Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

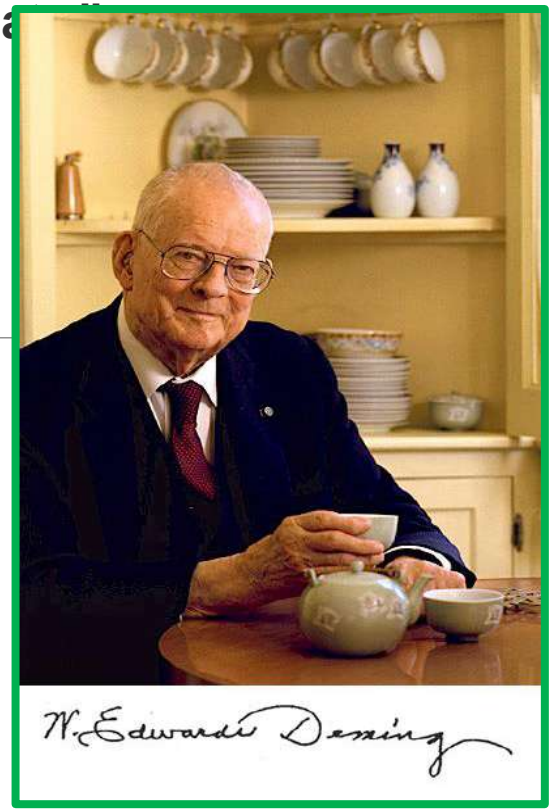
ТЕЛ. (484)3974658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

“In God we trust. All others must bring da

W. Edwards Deming



Цели и задачи дисциплины:

- Ознакомление специалистов с методами статистической обработки и анализа данных, в том числе с помощью методов, ранее изученных в курсе математической статистики, методов, изучаемых в спецкурсах,
- обучение представлению аналитической информации,
- обучение практическим приемам анализа данных на ЭВМ
- создание базиса для дальнейшего изучения разделов Науки о Данных (DATA SCIENCE), DATA MINING

Ранее не ориентировались на какой-то конкретный программный продукт – задача в том, чтобы понять методы, общие подходы, общие принципы реализации МАД (методов Анализа Данных) на ЭВМ

Проблемы:

сверхбыстрое развитие разделов DATA SCIENCE, DATA MINING

Аналитико-статистические методы – лишь часть анализа Больших Данных

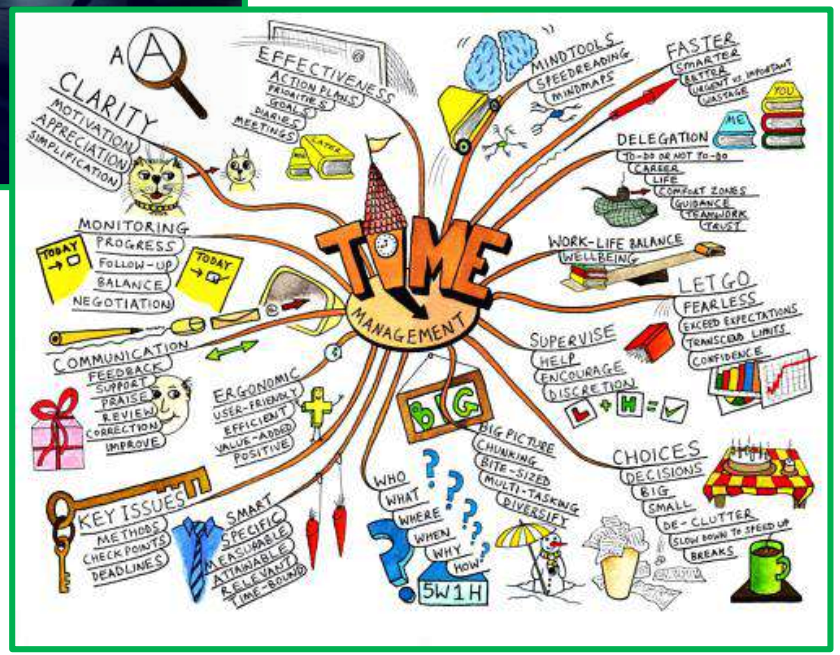
объективные сложности понимания основ методов (они сложные, особенно в эпоху спроса на быстрые решения),

Трудно визуализировать когда размерность больше трех

Легко посчитать, трудно найти нужный метод

Легко посчитать, трудно интерпретировать результаты

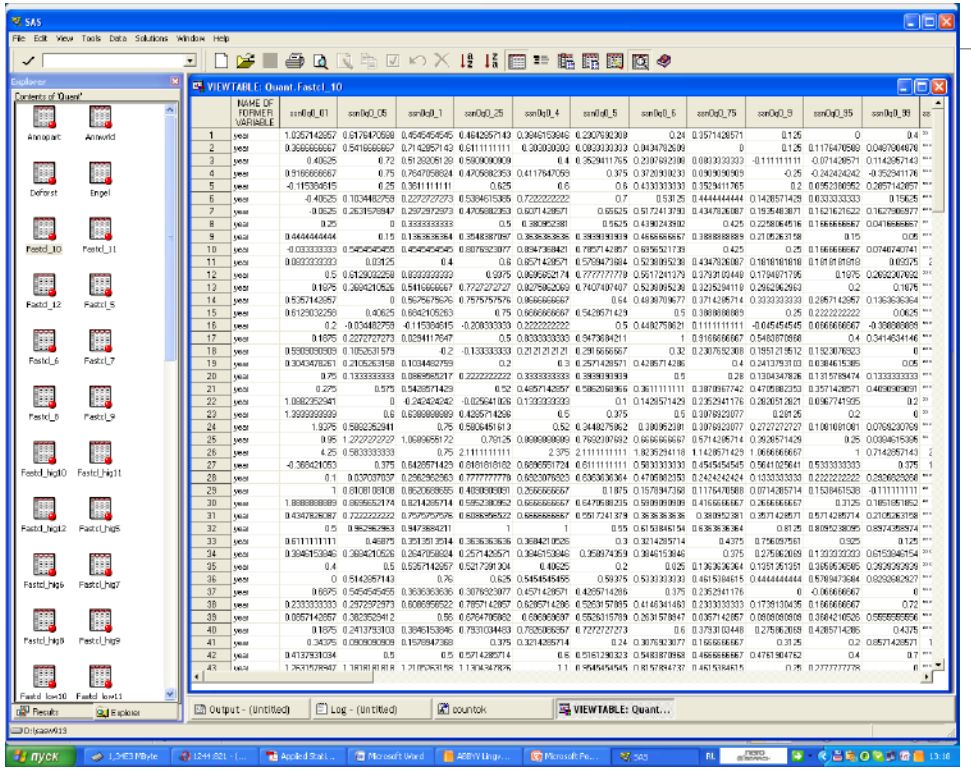
подавляющая часть аналитико-статистического ПО является коммерческим ПО, лишь немногое имеется в свободном доступе



Information is the oil of the 21st century, and analytics is the combustion engine.
— Peter Sondergaard

I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
—Sir Arthur Conan Doyle, author of Sherlock Holmes stories

Внутреннее представление данных – таблицы, или наборы данных (Data Sets)



Var1	Var2	…	VarN
Obs1			
Obs2			
…		.	
ObsL			

Столбцы, переменные, columns, variables
Строки, наблюдения, rows, observations

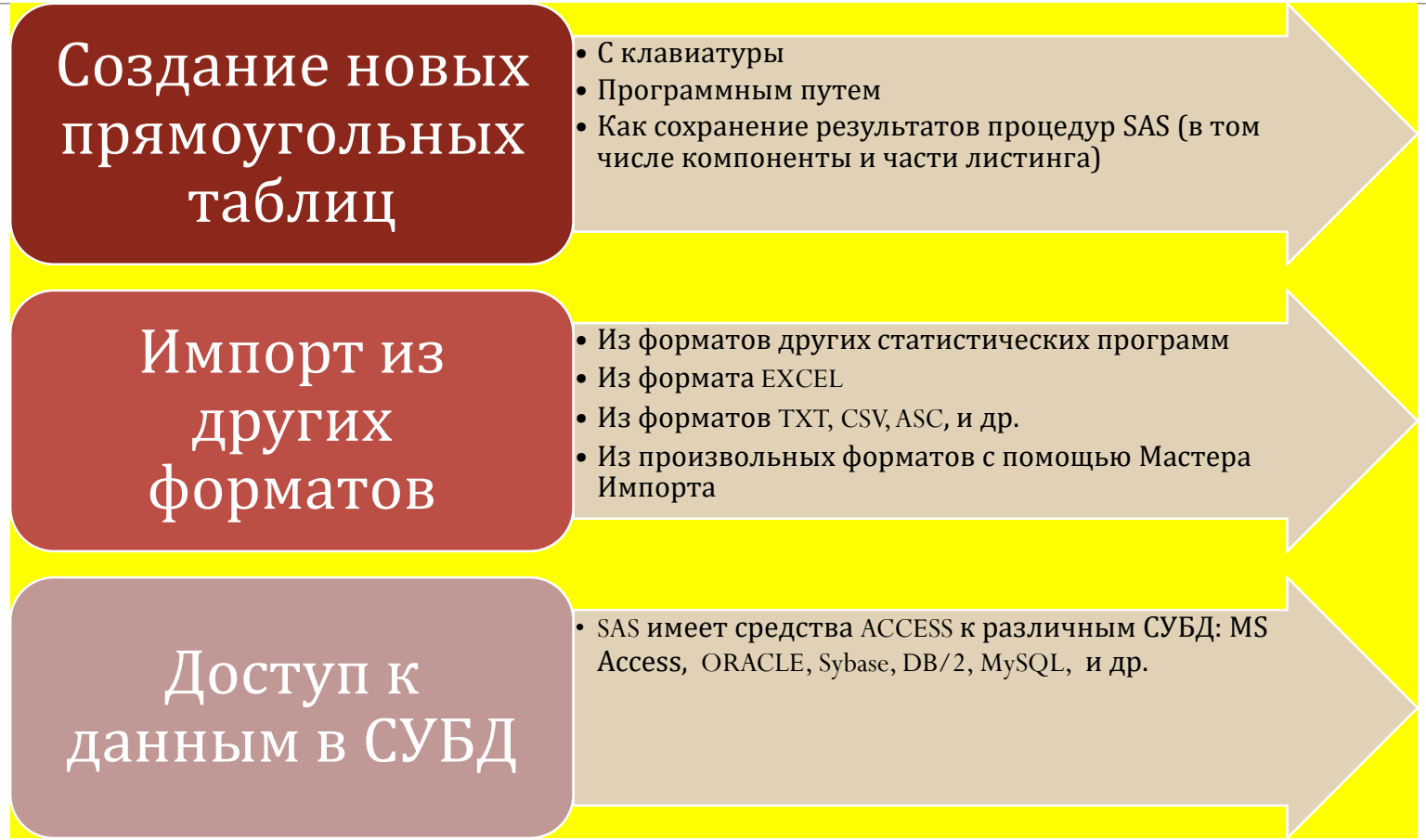
Отсутствующее значение, Требует специальной обработки

Разделы анализа данных (в традиционном понимании):

- Расчет и анализ элементарных статистик (одномерный анализ, univariate)
- Анализ связей между переменными (корреляционный анализ, частотный анализ)
- РЕГРЕССИИ:
  Линейный и нелинейный регрессионный анализ
- Классификация с обучением. Дискриминантный анализ
- Классификация без обучения. Кластерный анализ

- Анализ связей между группами переменных (канонические корреляции)
- Анализ главных компонент, понижение размерности и факторный анализ
- Анализ выживаемости
- Связи между нечисловыми переменными. Анализ соответствий
- Анализ и прогноз временнЫх рядов

Доступ к данным



Управление данными

SORTING - Сортировки – числовые и символьные переменные, возрастание и убывание, вложенные сортировки

INSERTING AND DELETING VARIABLES -Вставки и удаления переменных (столбцов)

INSERTING AND DELETING OBSERVATIONS - Вставки и удаления наблюдений (строк)

SUBSETTING OBSERVATIONS - Выделение подмножества наблюдений (в том числе случайным образом)

TRANSPOSING - Транспонирование

CONCATENATION - Конкатенация – слияния строк из нескольких таблиц

JOINING, MERGING - Объединение – объединения столбцов по значениям некоторых переменных или стыковка по номеру строки

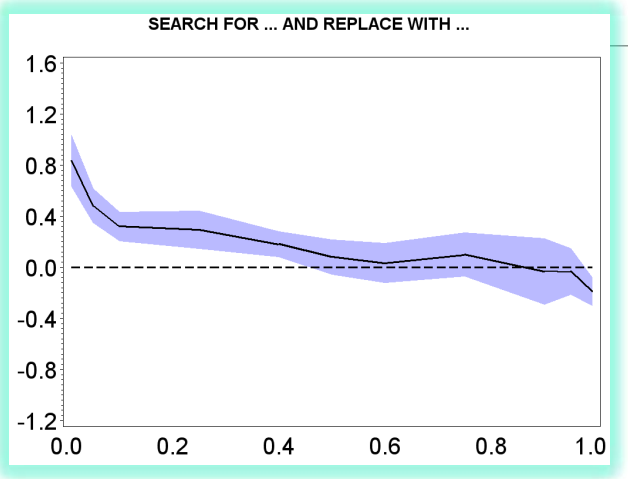
VARIABLE TRANSFORMATION (STANDARDIZING, etc.) – Преобразования переменных, в т.ч. стандартизация

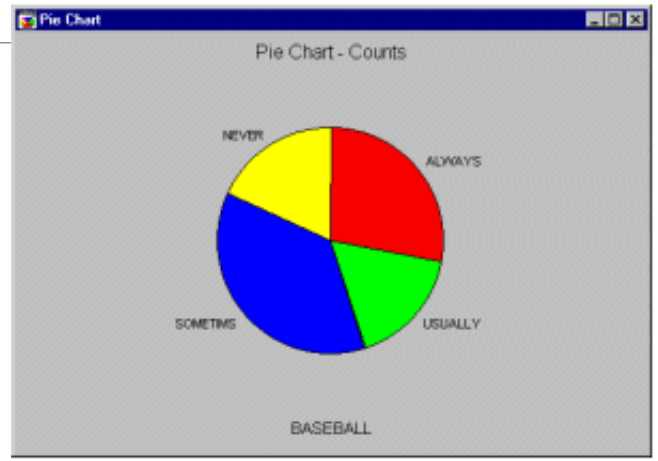
VARIABLE CALCULATOR - Калькулятор новых переменных

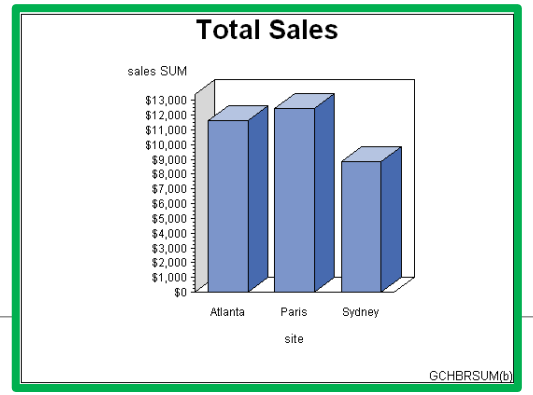
MISSING DATA PATTERN – Задание шаблона отсутствующих данных

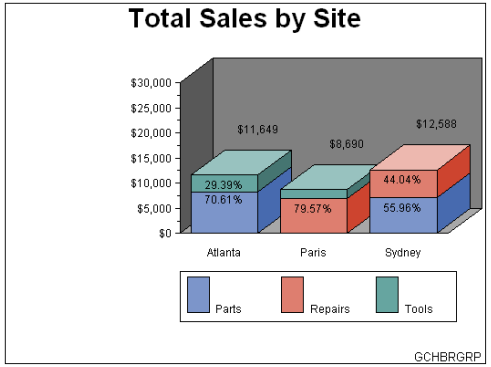
Визуализация данных - 1

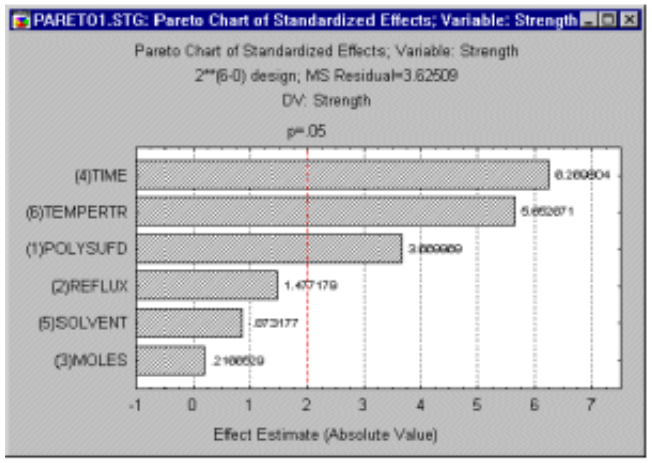
Визуализация – важнейший компонент анализа, без нее невозможно «заглянуть» в данные и увидеть *некоторые* закономерности. Однако она может помочь не во всем

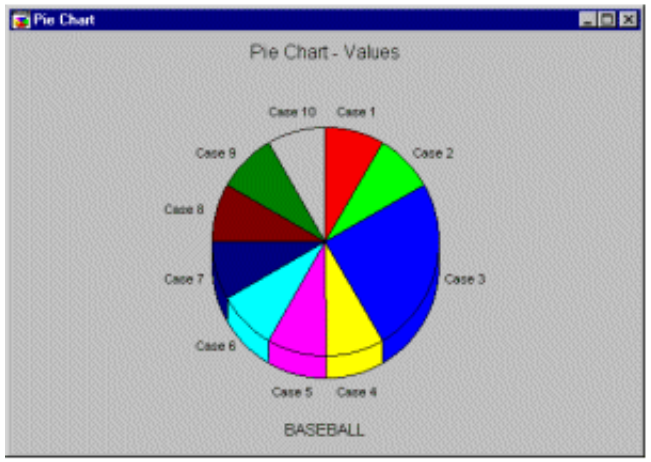


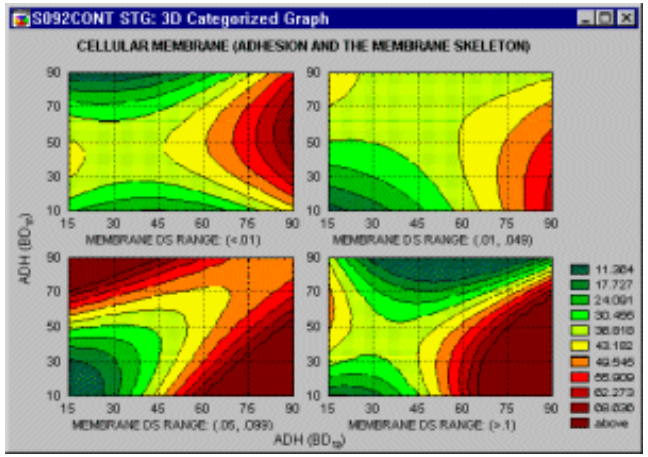


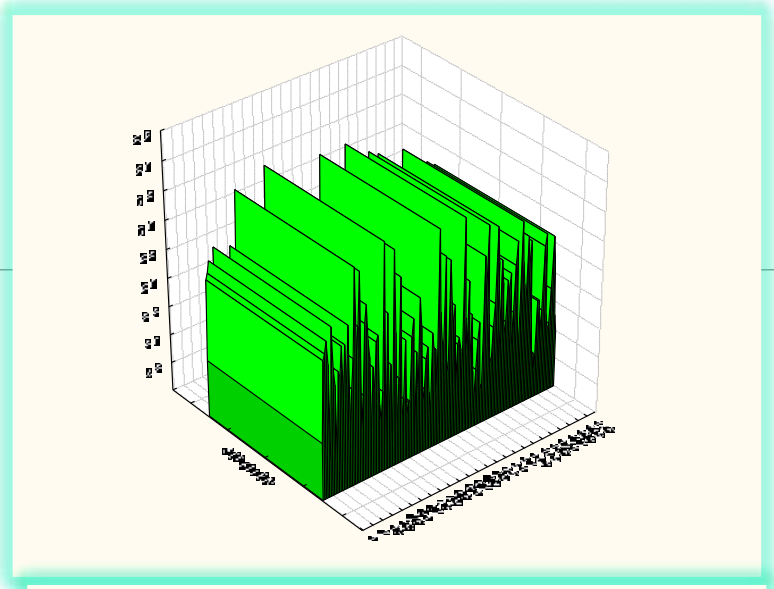


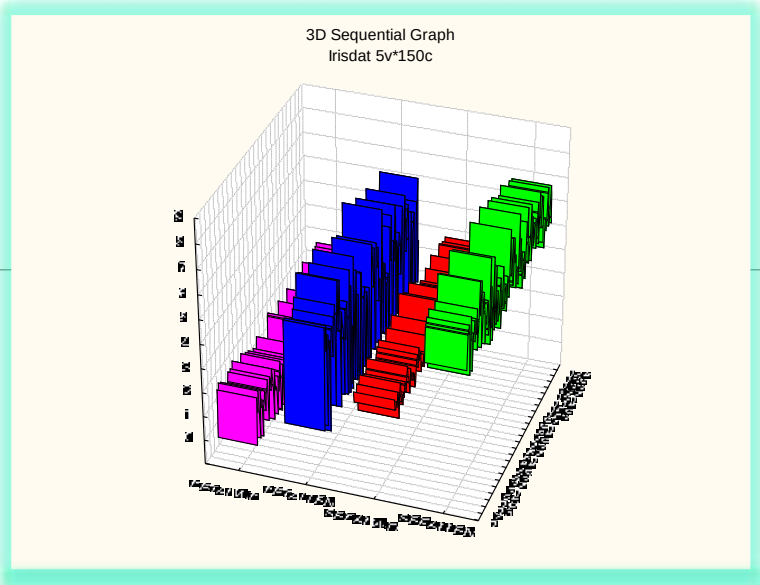


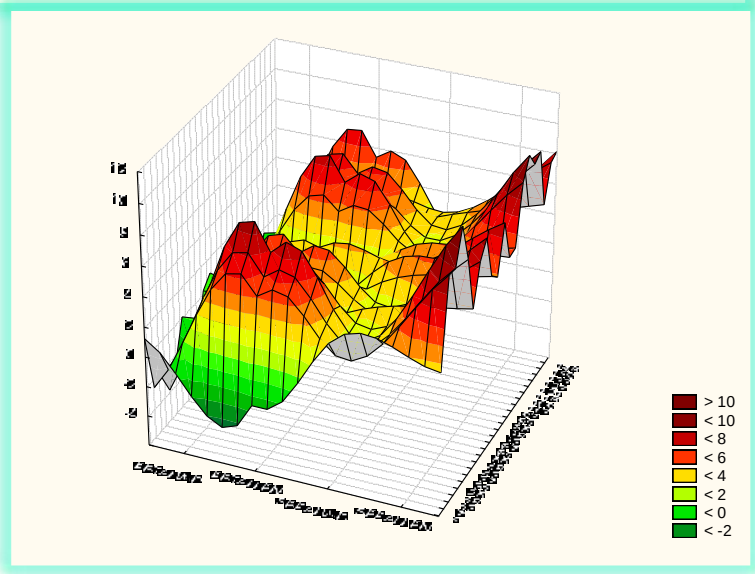


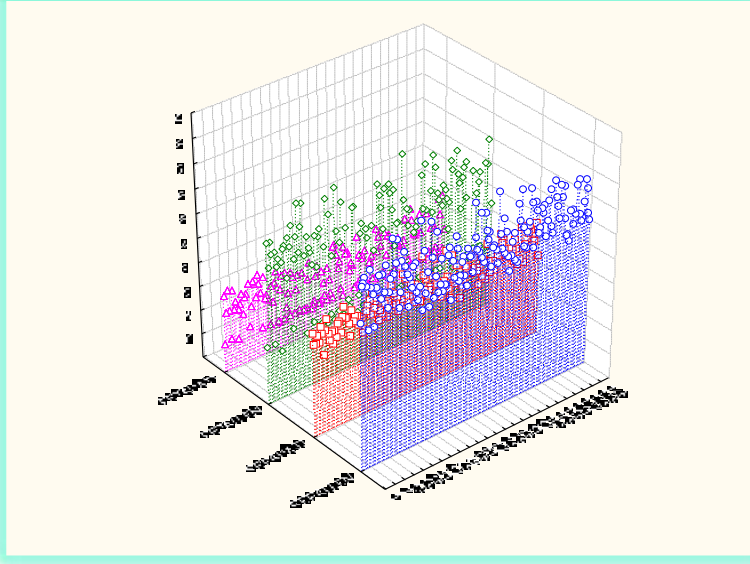


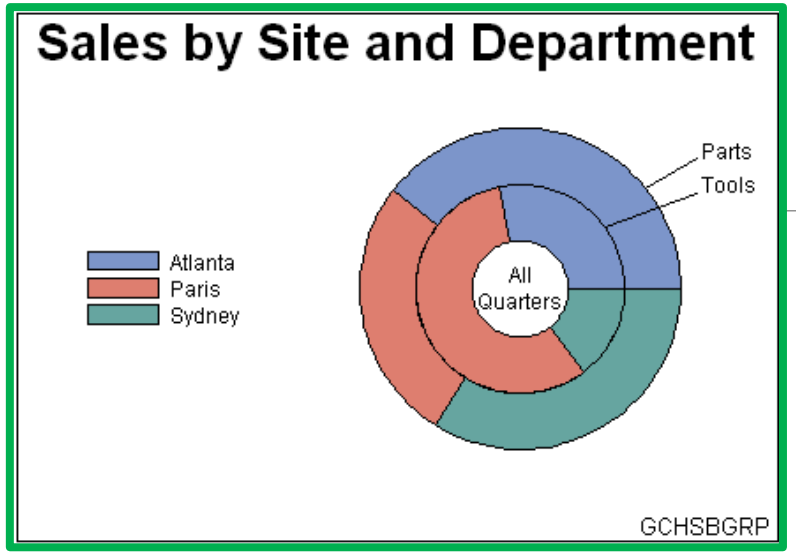


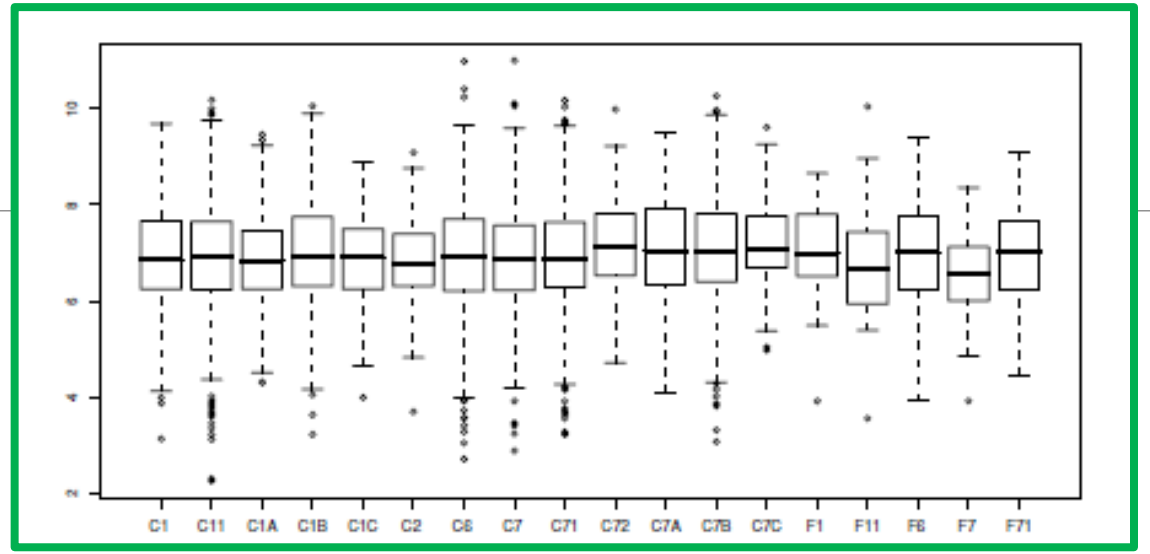


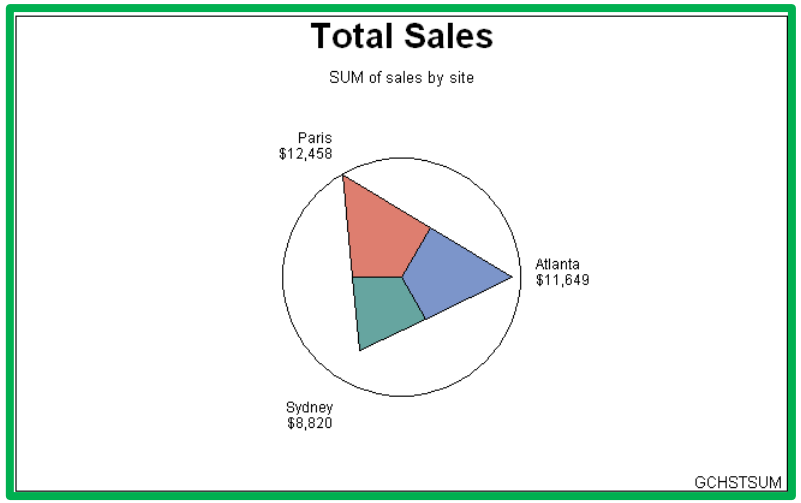

3D Sequential Graph
Irisdat 5v*150c

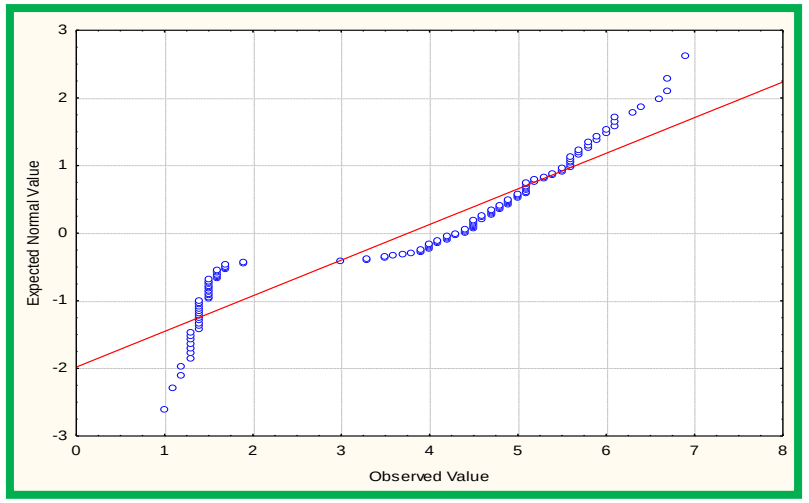

> 10
< 10
< 8
< 6
< 4
< 2
< 0
< -2









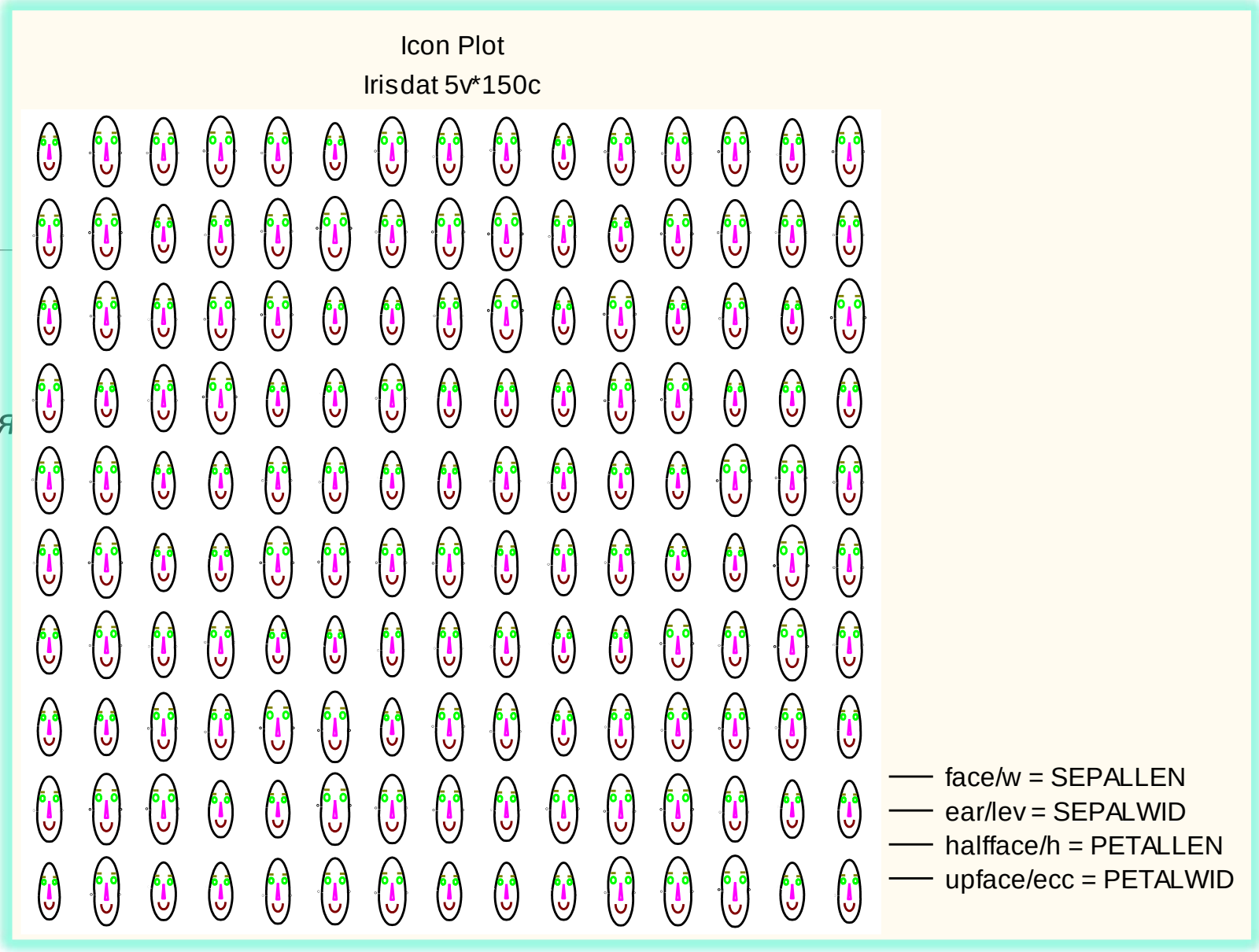


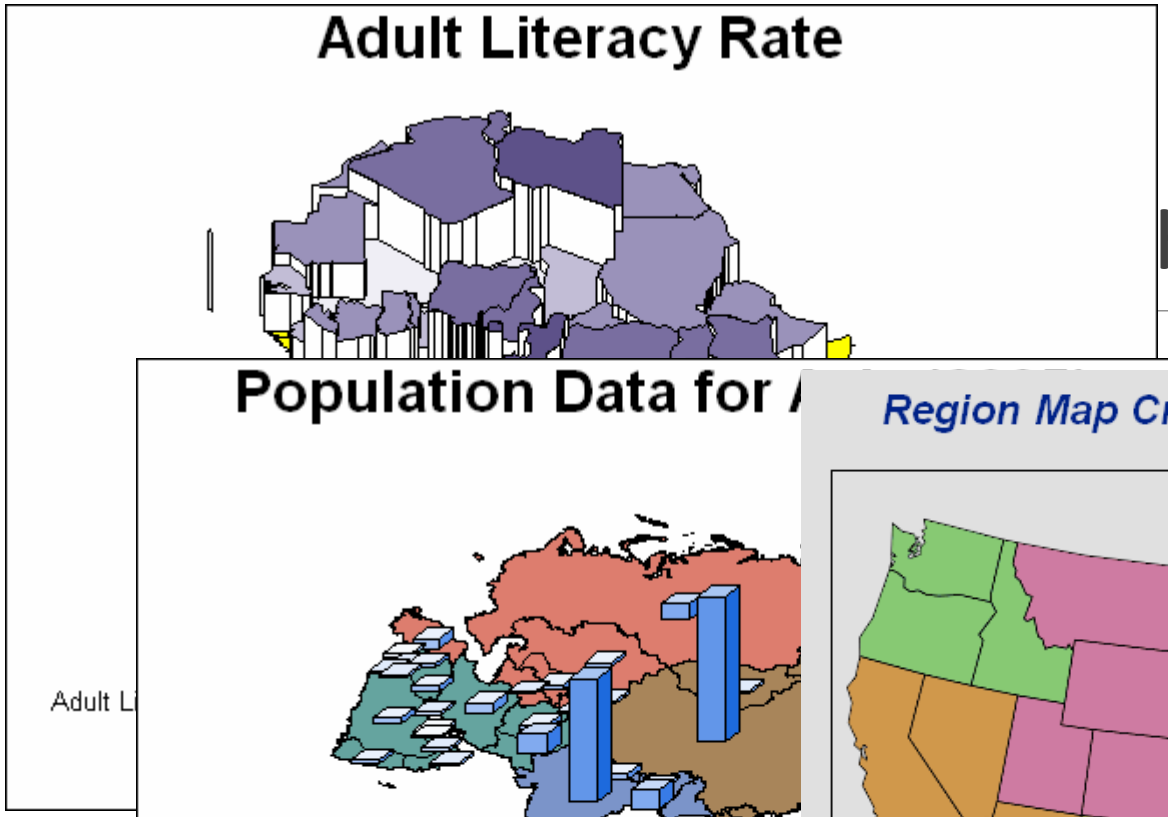
Как называются графики приведенных типов?

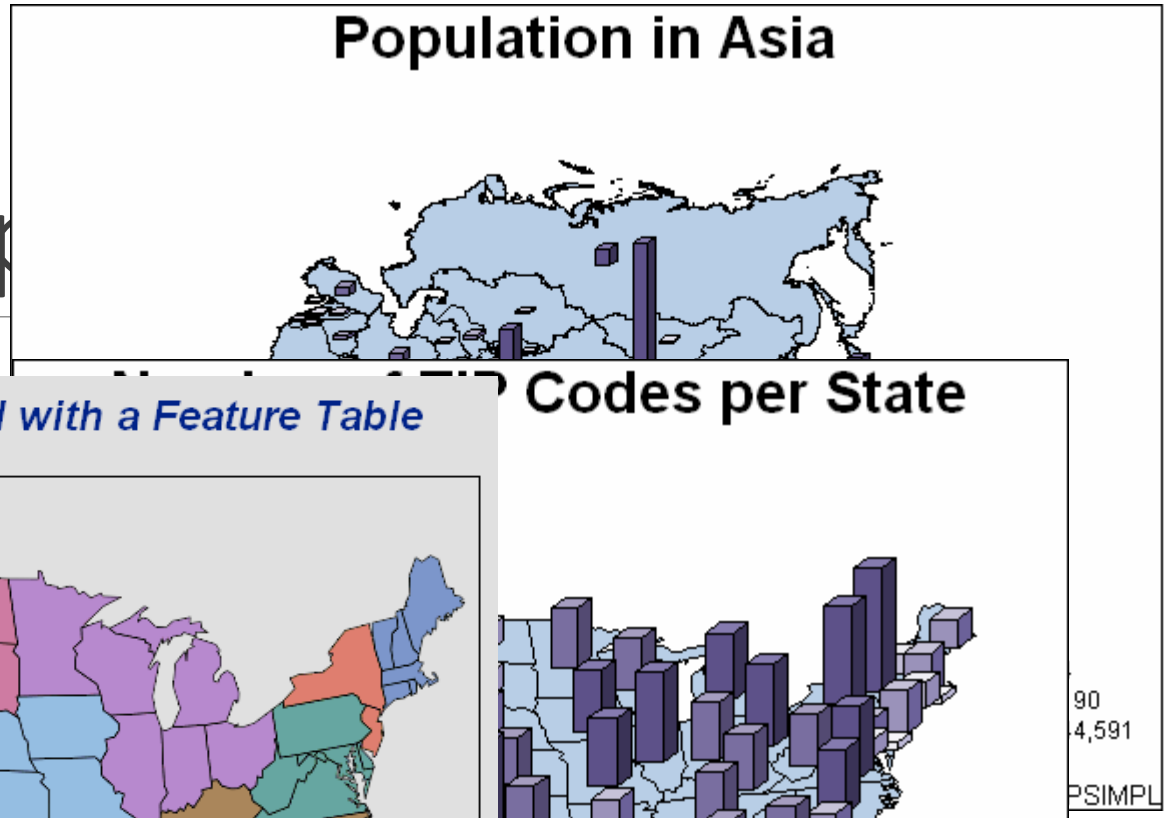
Лица Чернова

Chernoff Faces

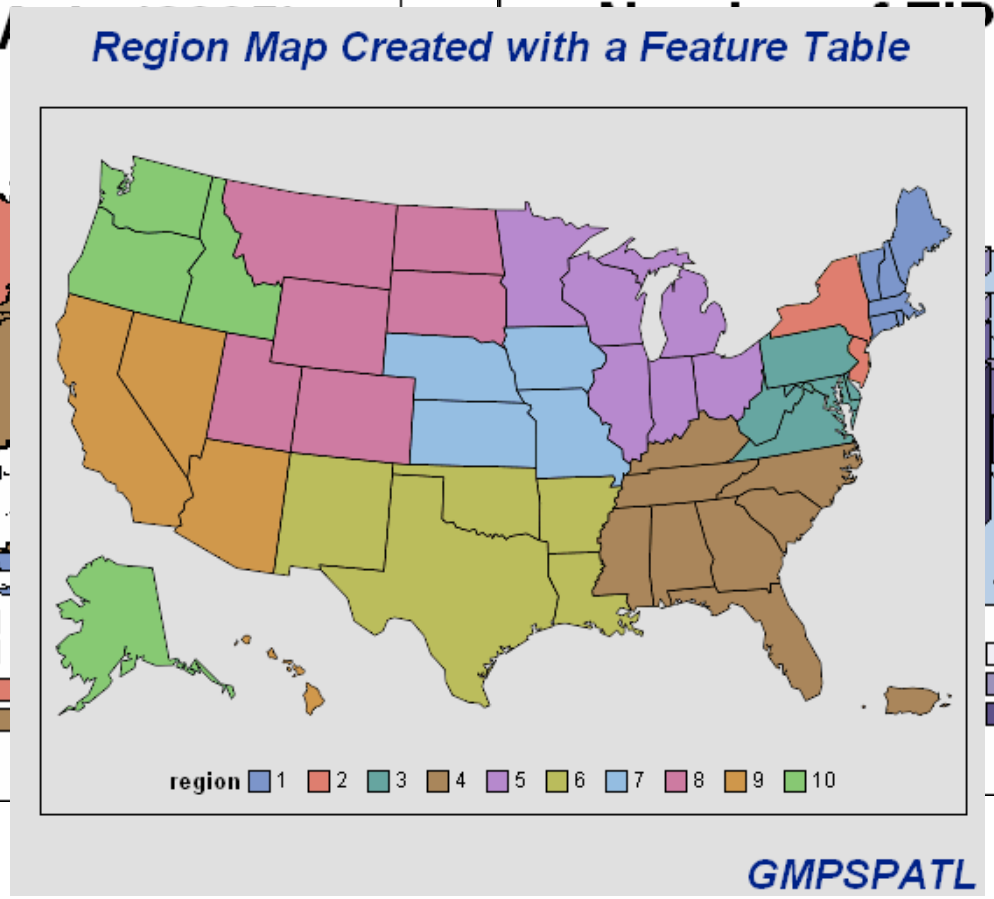
Так выглядит визуализация набора данных об Ирисе Фишера




Adult Literacy Rate


Population in Asia

Population Data for
Population (Millions)
0-1
100-500
Region
South-East Asia
Eastern Mediterranean

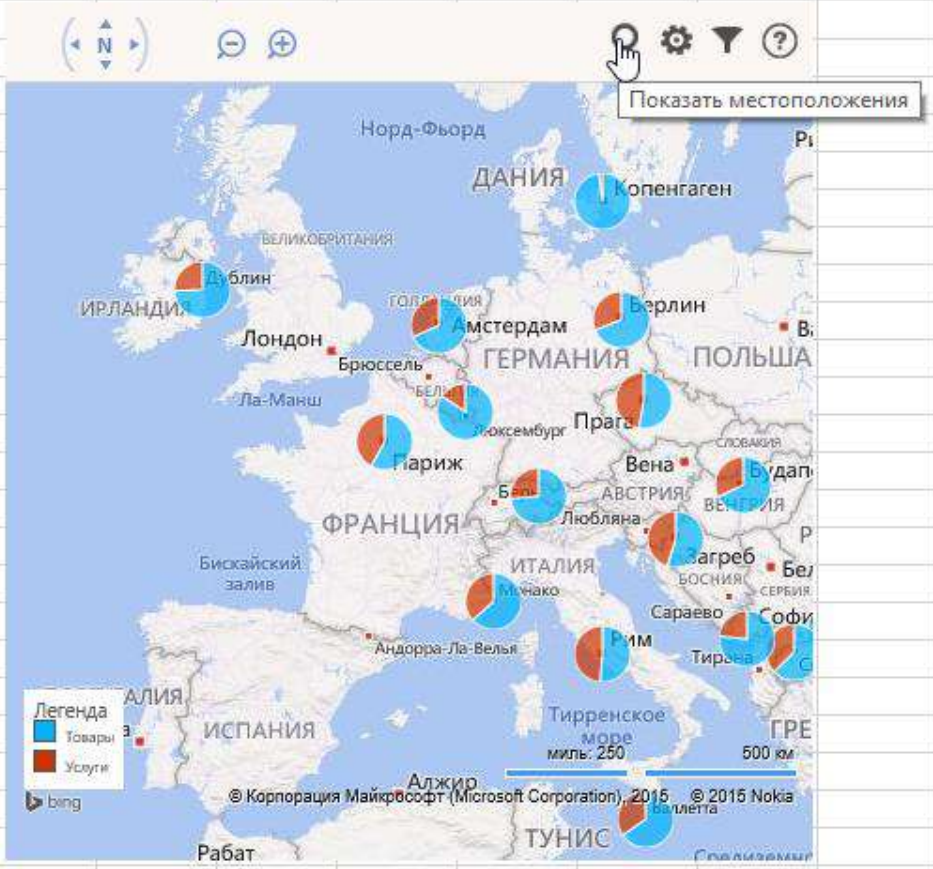

Region Map Created with a Feature Table
region 1 2 3 4 5 6 7 8 9 10
GMPSPATL

Codes per State
1 - 273
278 - 484
488 - 731
732 - 1031
1060 - 2650
GMPSTAT

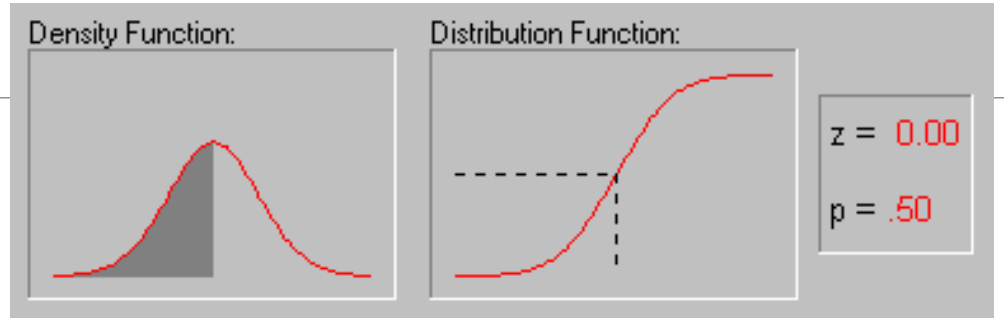
Графика, картография и 3D картография – теперь в EXCEL!!!!

Прибыль от продажи товаров и услуг в Европе

Город	Товары	Услуги
София	7 621 337	379 669
Загреб	4 637 460	3 811 353
Никосия	863 457	185 450
Прага	10 256 760	9 161 847
Копенгаген	5 568 854	119 320
Таллин	1 315 681	485 700
Хельсинки	5 357 537	1 439 115
Париж	6 165 983	4 458 140
Берлин	8 551 851	3 848 908
Афины	11 245 343	4 744 214
Будапешт	10 075 034	4 765 399
Дублин	4 434 925	1 532 589
Рим	6 051 711	5 717 670
Рига	2 366 515	426 267
Вадуц	35 322	12 626
Вильнюс	3 401 138	1 327 929
Люксембург	472 569	88 375
Скопье	2 054 800	1 230 016
Валлетта	408 009	216 736
Кишинёв	3 834 547	931 409
Монако	32 087	18 504
Подгорица	598 258	182 530
Амстердам	16 518 199	7 680 719



Одномерные статистики
Univariate statistics



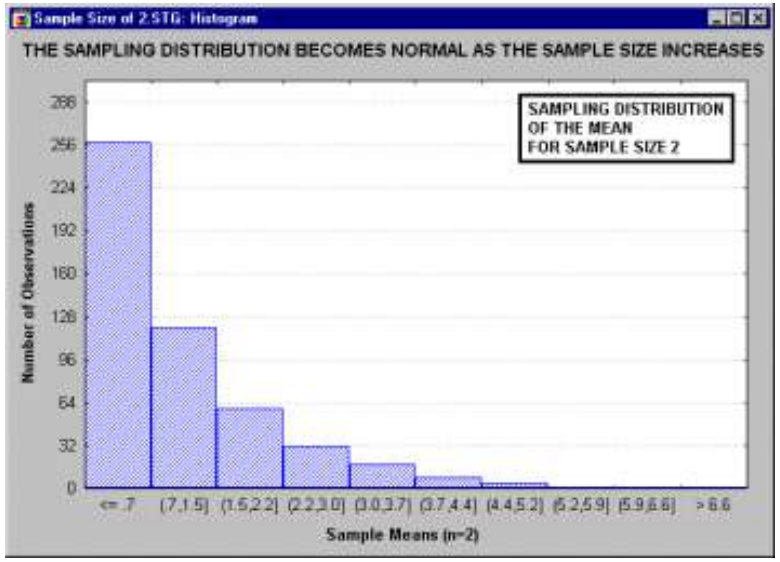
Функции распределения случайной величины ξ, заданной выборкой:

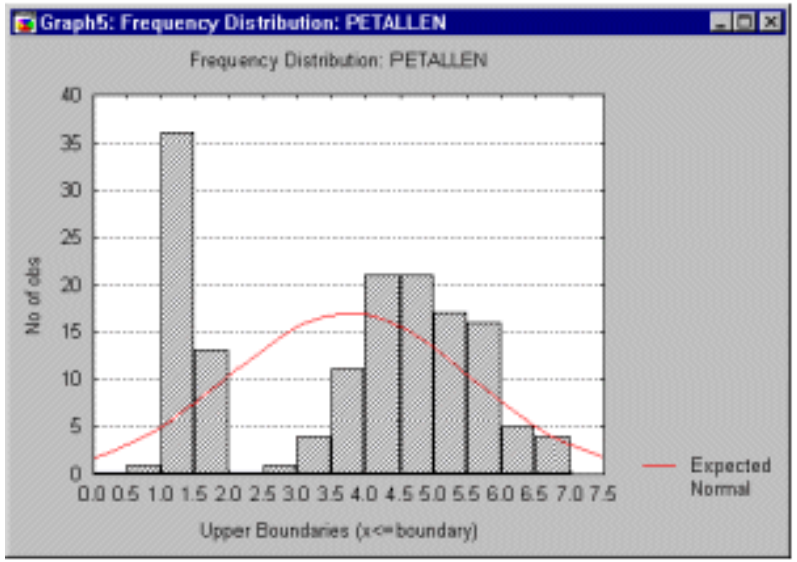
$$\xi_1, \xi_2, \ldots \xi_n$$

$$F_\xi(x) = P(\xi < x)$$

Плотность вероятности

$$f(x) = F_\xi\,'(x)$$





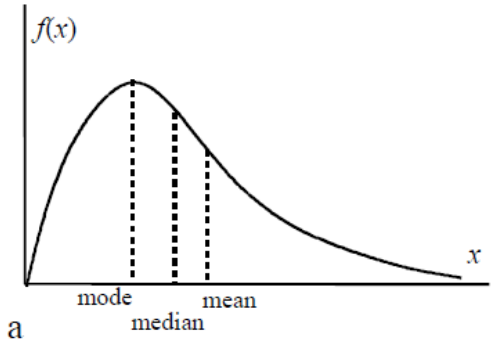
Гистограммы

Одномерные статистики
Univariate statistics

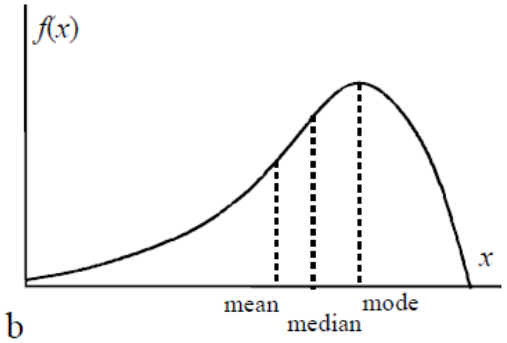
Выборочное среднее Mean	$E(\xi) = \bar{\xi} = (\sum_{i=1}^{l} \xi_i) / l$
Дисперсия (делитель – число наблюдений) Dispersion (for NOBS)	$D(\xi) = \sigma^2 = (\sum_{i=1}^{l} \xi_i - \bar{\xi})^2 / l$
Дисперсия (делитель – число степеней свободы) Dispersion (for DF)	$D(\xi) = \sigma^2 = (\sum_{i=1}^{l} \xi_i - \bar{\xi})^2 /(l-1)$
Коэффициент вариации Variance	$V(\xi) = \frac{\sigma_\xi}{E_\xi}$
Асимметрия Skewness	$\beta = \frac{E(\xi - \bar{\xi})^3}{(E(\xi - \bar{\xi})^2)^{3/2}}$
Эксцесс Kurtosis	$\gamma = \frac{E(\xi - \bar{\xi})^4}{(E(\xi - \bar{\xi})^2)^2} - 3$

Одномерные статистики
Univariate statistics

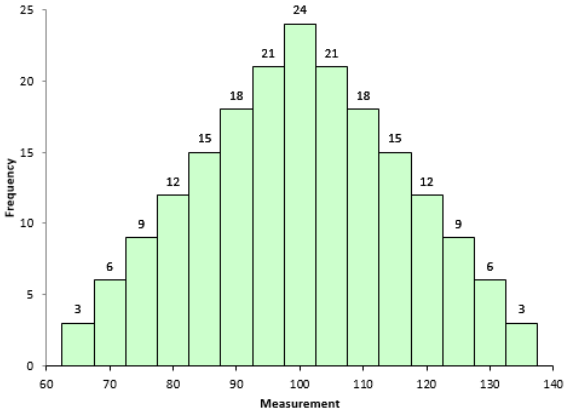
Элементарные статистики – асимметрия и эксцесс

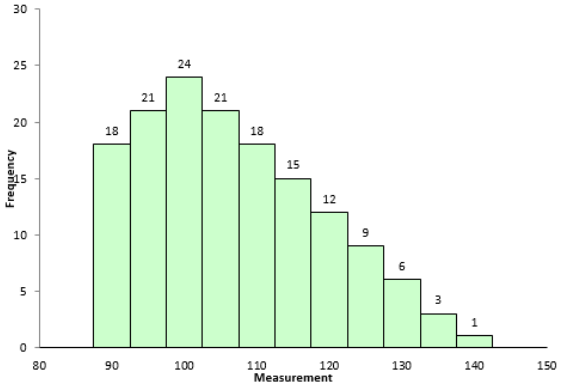


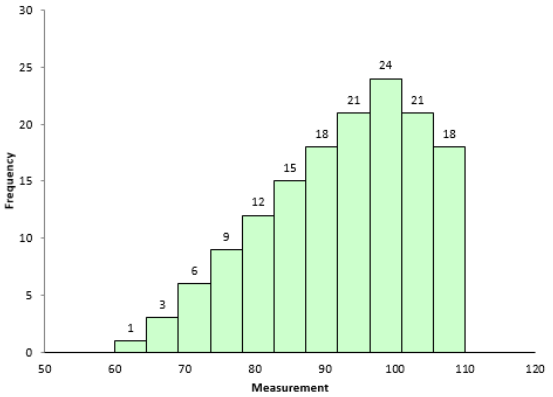
β>0



β<0







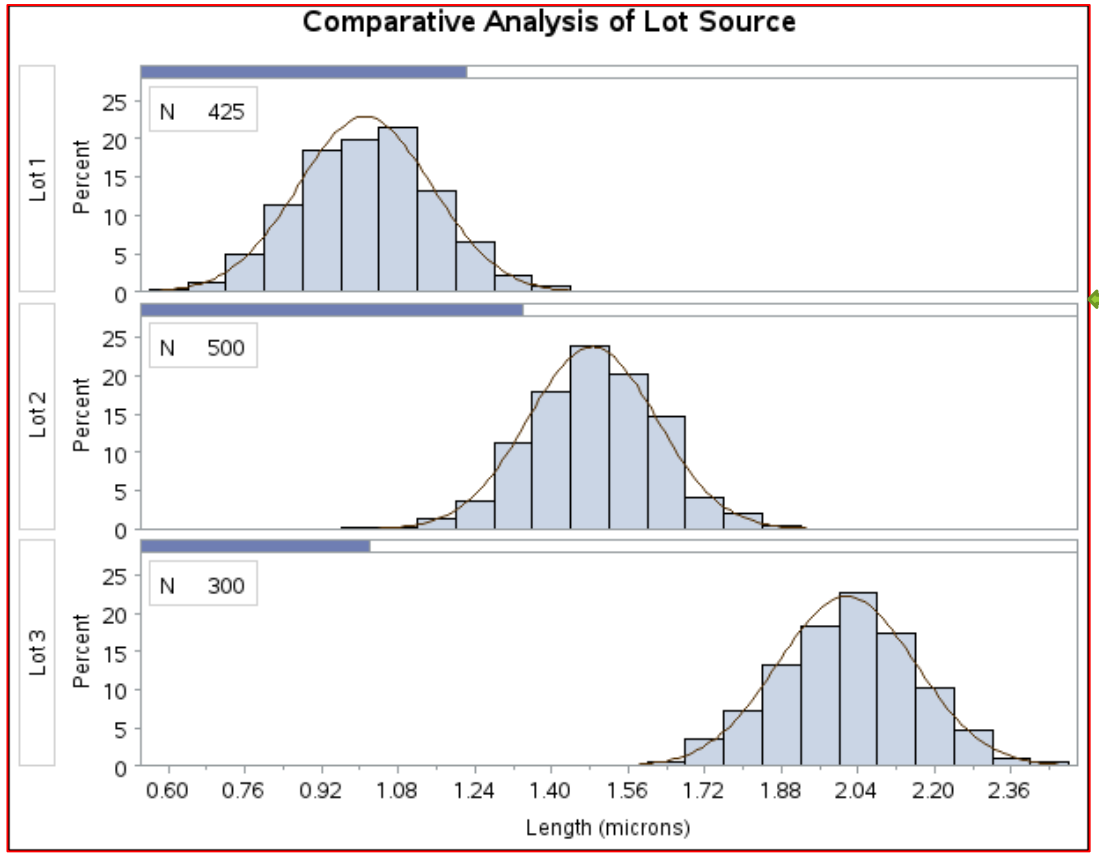
Визуализация 2- статистические графики

Гистограммы с наложенной плотностью нормального распределения для трех классов:

Добавление кривой плотности нормального распределения на сравнительную гистограмму

Histograms with Normal curve for three classes:

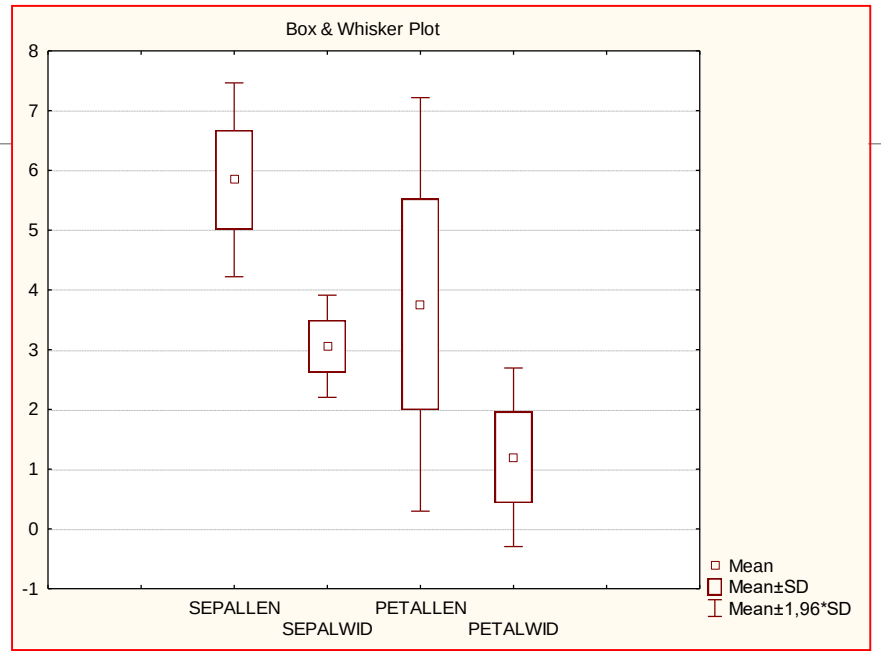
Adding Fitted Normal Curves to a Comparative Histogram

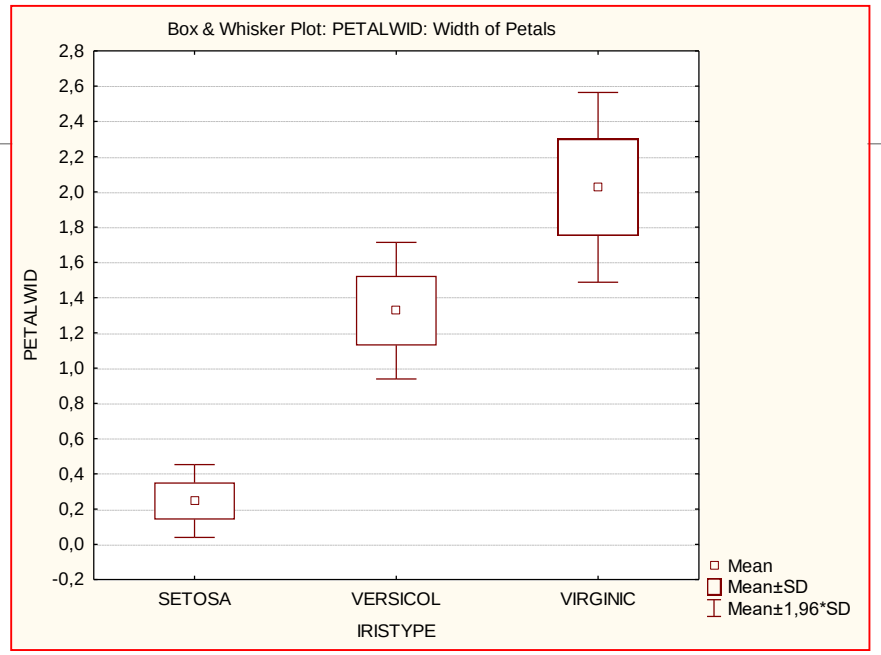


Классифицирующая (группирующая) переменная **Lot**

N – число наблюдений в группе

Ящики с усами для данных Ириса Фишера (box and whisker plots)

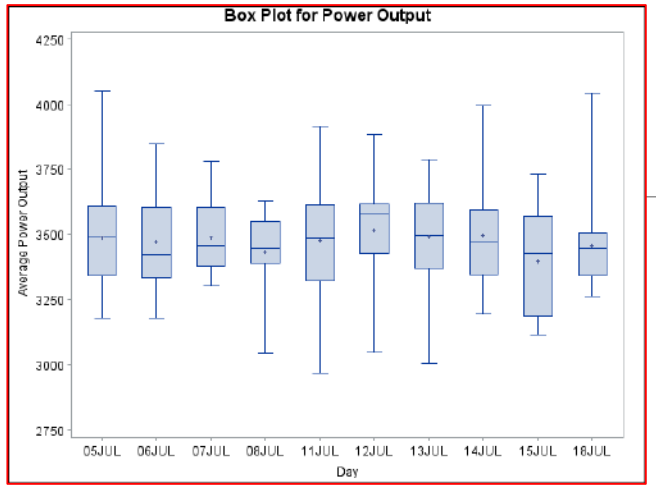


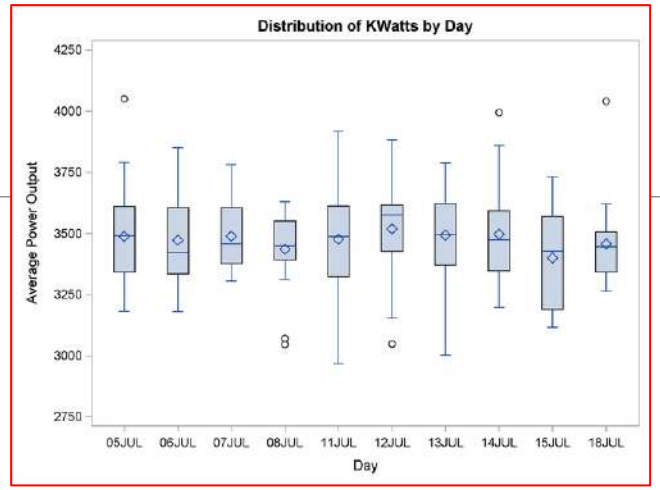


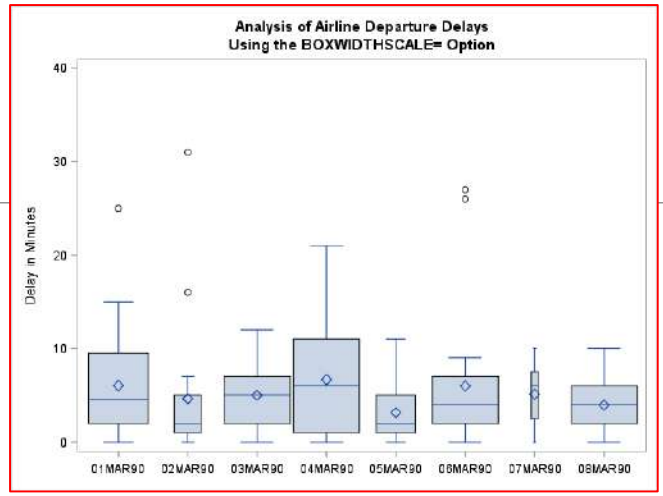
Для четырех переменных

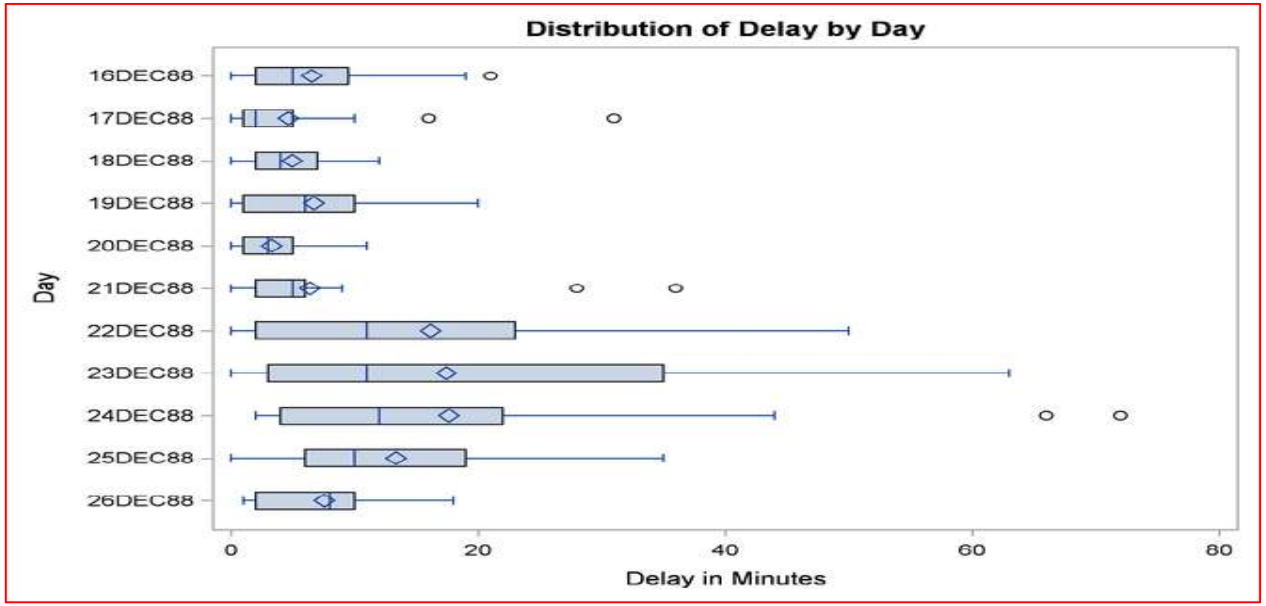
Для одной переменной, но по градациям категоризирующей переменной

Ящики с усами (box and whisker plots)

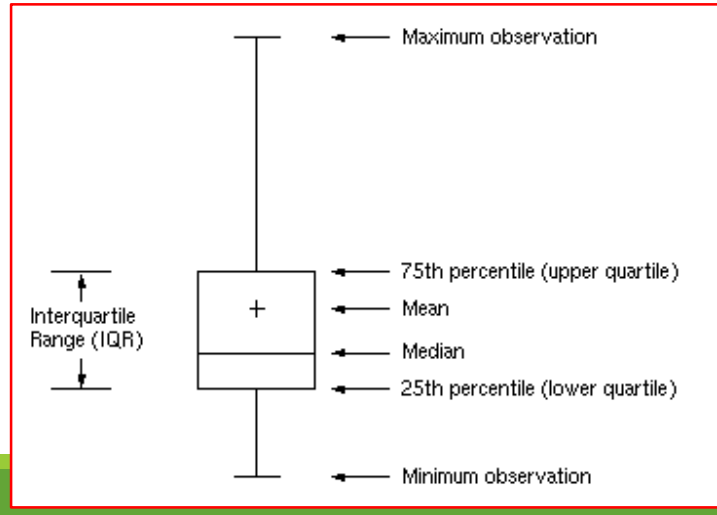




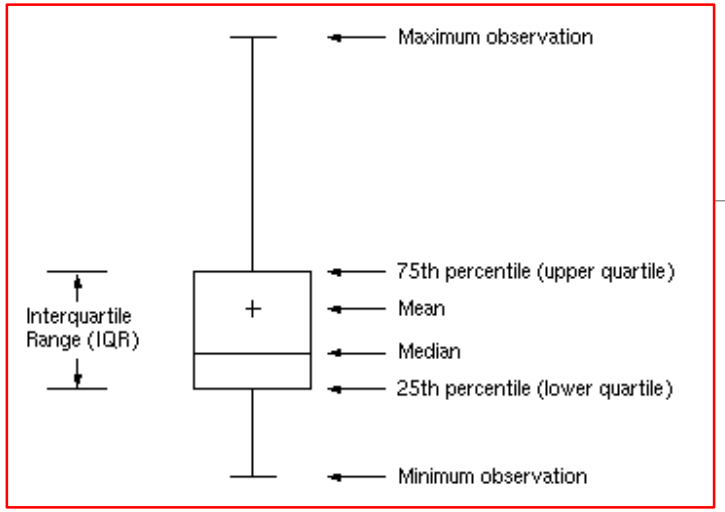




Ширина «ящика» пропорциональна числу наблюдений в группе



Квантили случайной величины(Quantiles)



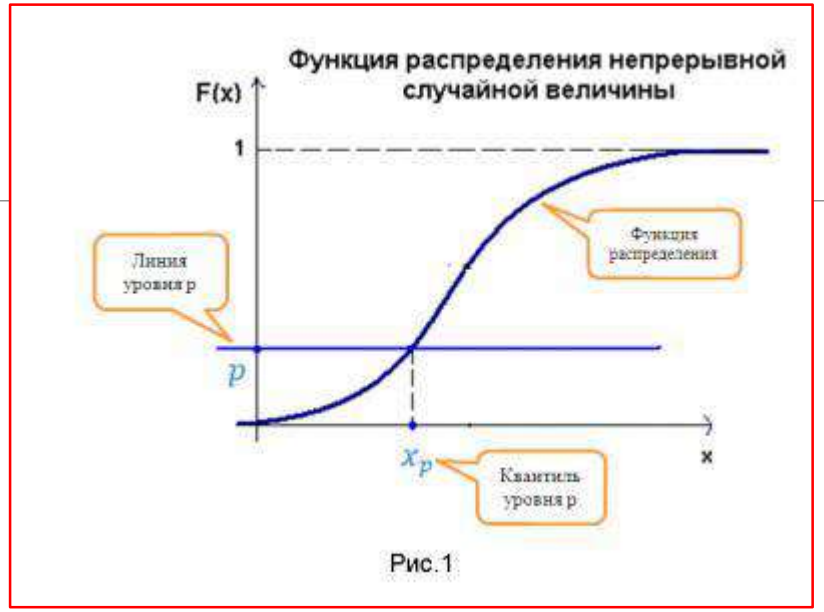


Рис.1

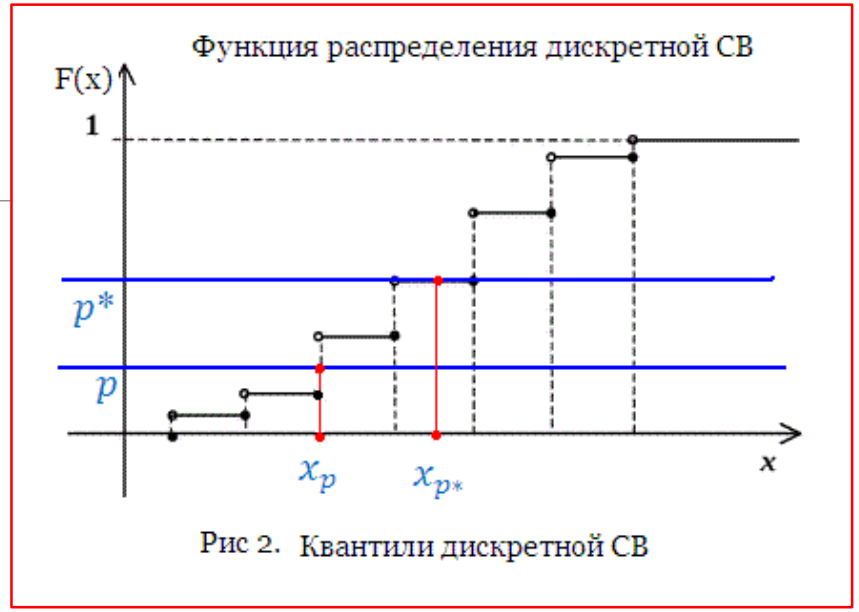


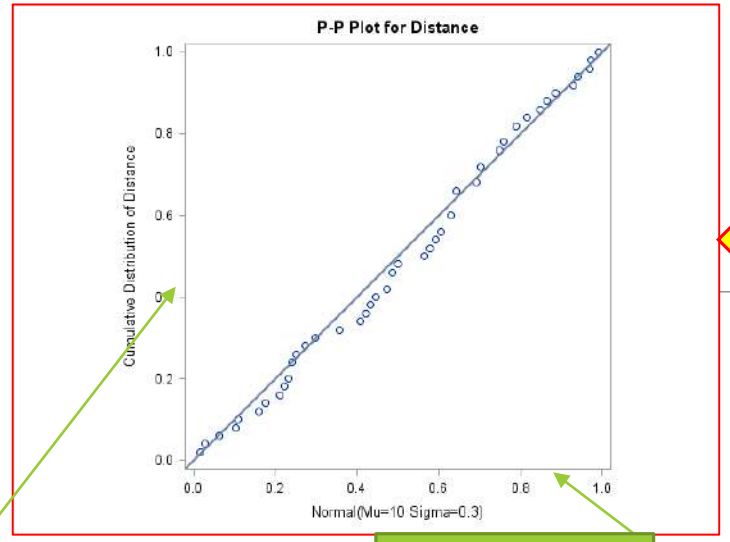
Рис 2. Квантили дискретной СВ

Квантилью x_p (*p*-квантилью, квантилью уровня *p*) случайной величины *X*, имеющей функцию распределения *F* (*x*), называют решение x_p уравнения *F* (*x*) = *p*

Для дискретной случайной величины X функция распределения F(x) имеет ступенчатый вид, функция не монотонна. Поэтому решение уравнения F(x) = p в общем случае не однозначно (в решение попадают интервалы). В таких случаях, для определенности квантилем назначают средину интервала

Квартилями называют квантили порядков 0,25, 0,5 и 0,75. Будем их обозначать соответственно как k1 , k2 , k3 . Квартили k1 и k3 называют обычно нижней и верхней квартилями. Вторая квартиль k2 совпадает с медианой распределения
Децилями называют квантили уровня 0,1, 0,2, 0,3, ...0,9, обозначают соответственно d1, d2, d3,...d9

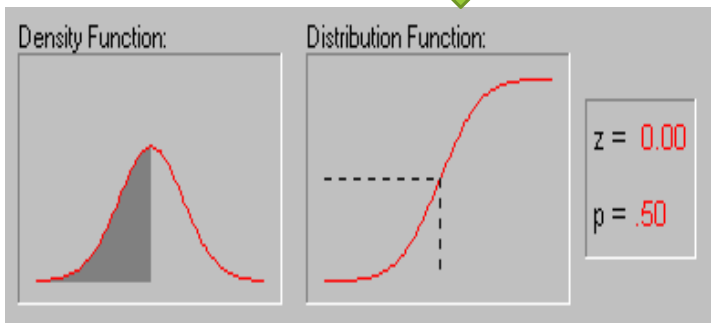
Графики P-P и Q-Q
P-P and Q-Q Plots
P-P графики



Квадрат, если масштабы осей одинаковы!

Эмпирические значения

Используется, когда исследуется Функция распределения



Это график связи **между двумя функциями распределения** (кумулятивными). В данном примере горизонтальная ось – значения функции распределения с определенными параметрами. Для нормального распределения достаточно среднего и сигмы.

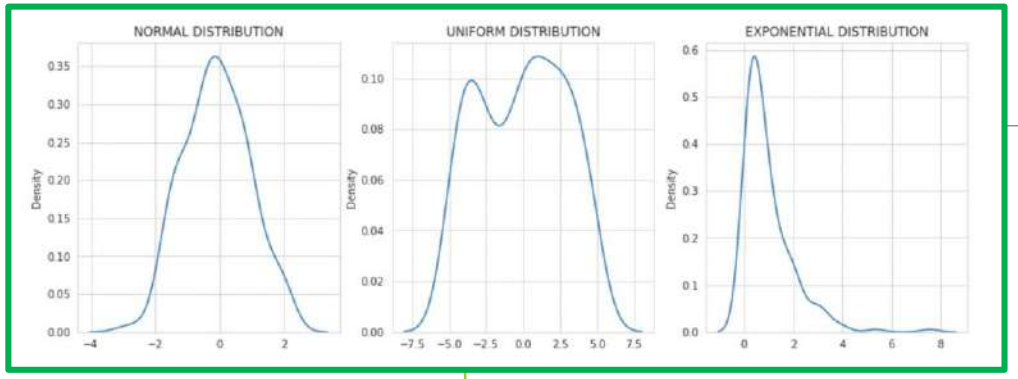
Вертикальная ось – значения кумулятивной функции распределения вычисленной на основе данных (эмпирическая).

Надо для **P-P графика** задать параметры теоретического распределения (в случае нормального теоретического распределения - среднее и сигму)!

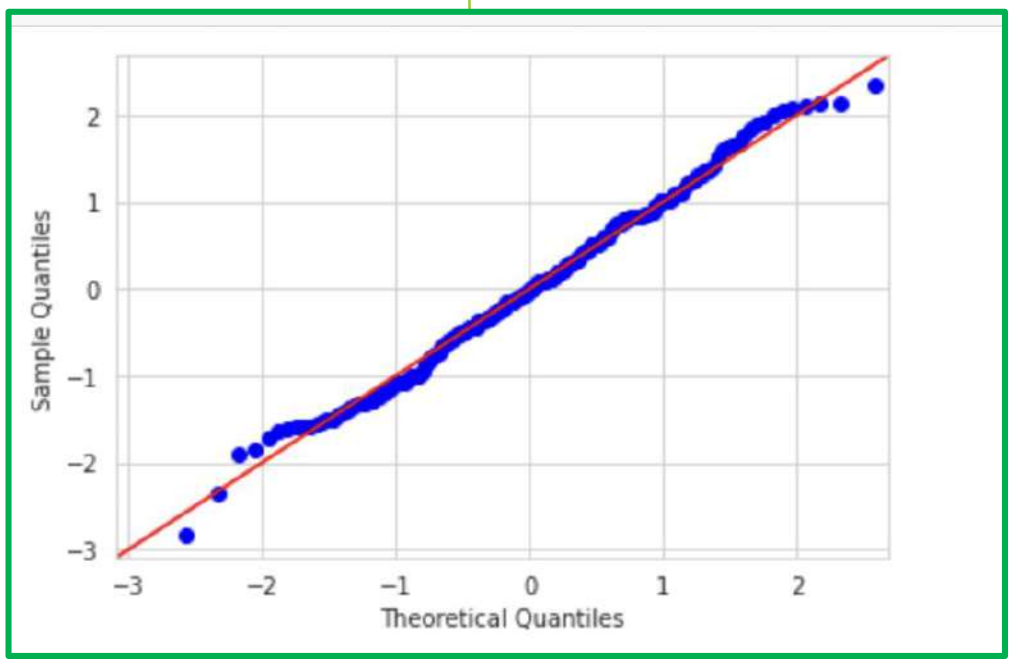
Различия между двумя кумулятивными функциями распределения на **P-P графике** лучше всего видны «ближе к середине», т.к. здесь скорость возрастания (т.е. плотность) – максимальна!

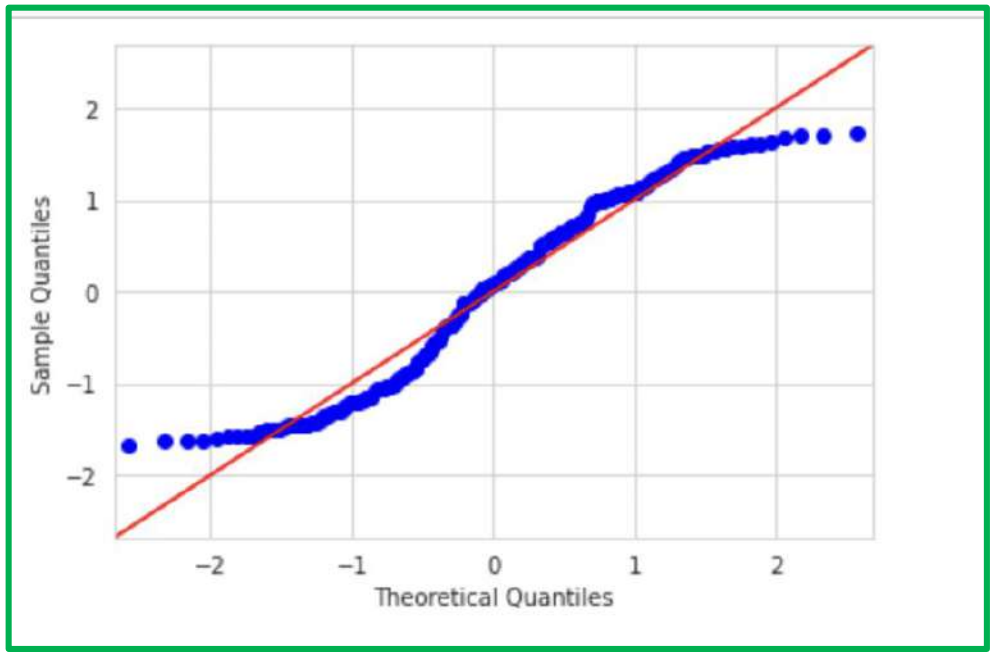
Оси могут меняться ролями в **P-P графиках (как и в Q-Q графиках!)**

Графики P-P и Q-Q
P-P and Q-Q Plots
Q-Q графики



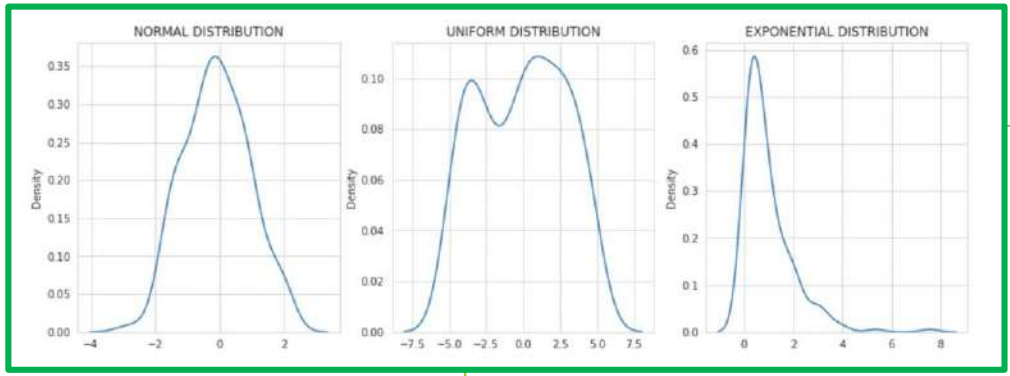
Надо явно указывать, с квантилями какого теоретического распределения ведется сравнение эмпирически вычисленных по выборке квантилей



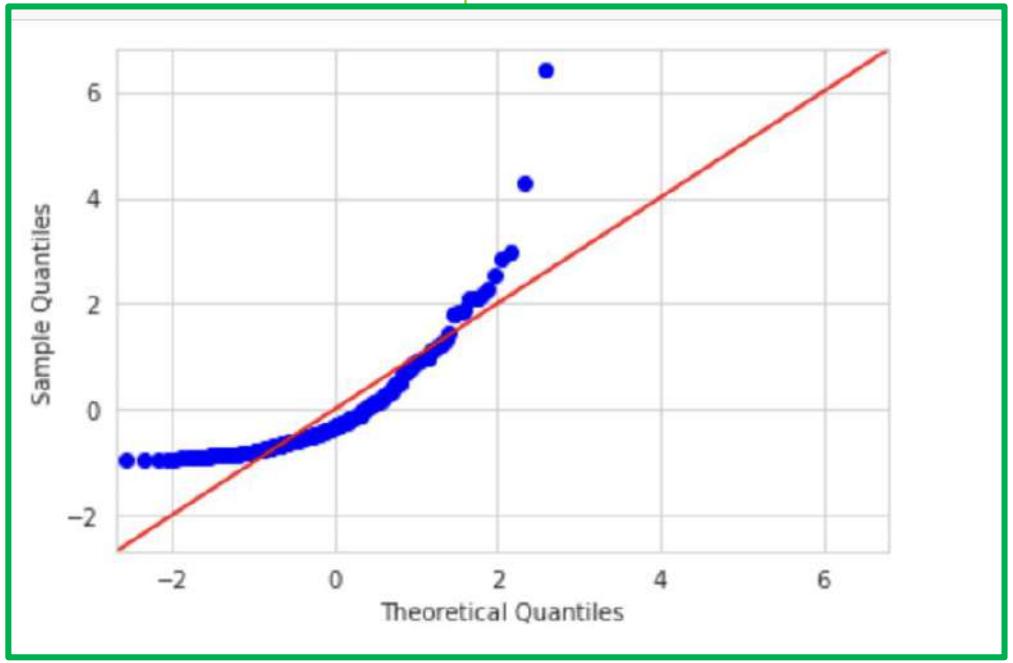


Данные имеют равномерное распределение, но его квантили сравниваются с теоретическими квантилями нормального распределения!!!!

Графики P-P и Q-Q
P-P and Q-Q Plots
Q-Q графики

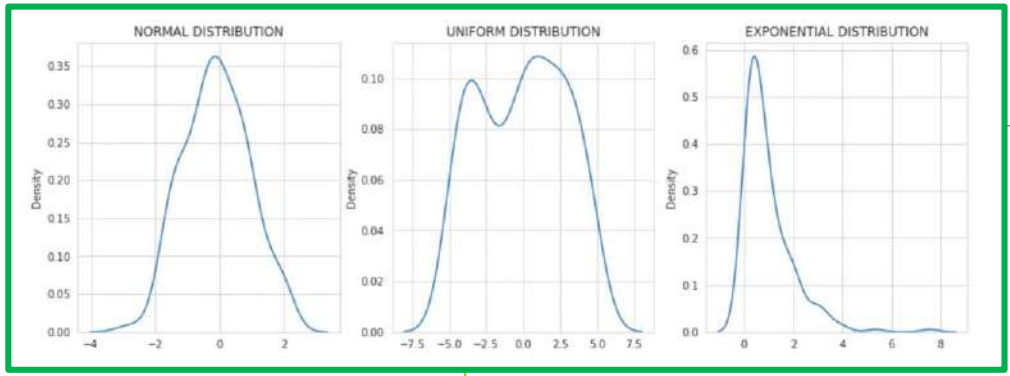


Надо явно указывать, с квантилями какого теоретического распределения ведется сравнение эмпирически вычисленных по выборке квантилей

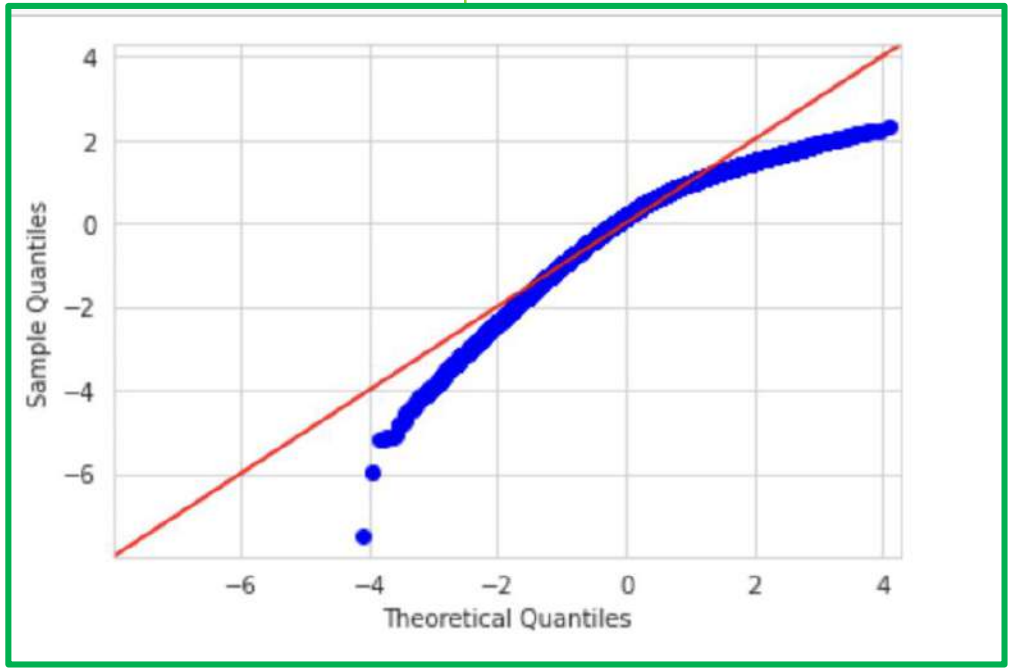


Данные имеют экспоненциальное распределение, но его квантили сравниваются с теоретическими квантилями нормального распределения!!!!

Графики P-P и Q-Q
P-P and Q-Q Plots
Q-Q графики



Надо явно указывать, с квантилями какого теоретического распределения ведется сравнение эмпирически вычисленных по выборке квантилей



*Данные имеют скошенное **вправо** распределение, но его квантили сравниваются с теоретическими квантилями нормального распределения!!!!*

Графики P-P и Q-Q
P-P and Q-Q Plots

Сравнение P-P и Q-Q Графиков

График P-P сравнивает эмпирическую кумулятивную функцию распределения числовой переменной, заданной в наборе данных, с определенной теоретической функцией распределения F(.).

График Q-Q служит для сопоставления квантилей числовой переменной, заданной в наборе данных, с квантилями стандартизованного (!) теоретического распределения из определенного семейства распределений.

Есть важные отличия в том, как строятся и как интерпретируются оба этих вида графиков.

Q-Q графики не зависят от параметров сдвига и масштаба распределения. Квантили теоретического распределения вычисляются по стандартизованным значениям определенного семейства распределений.

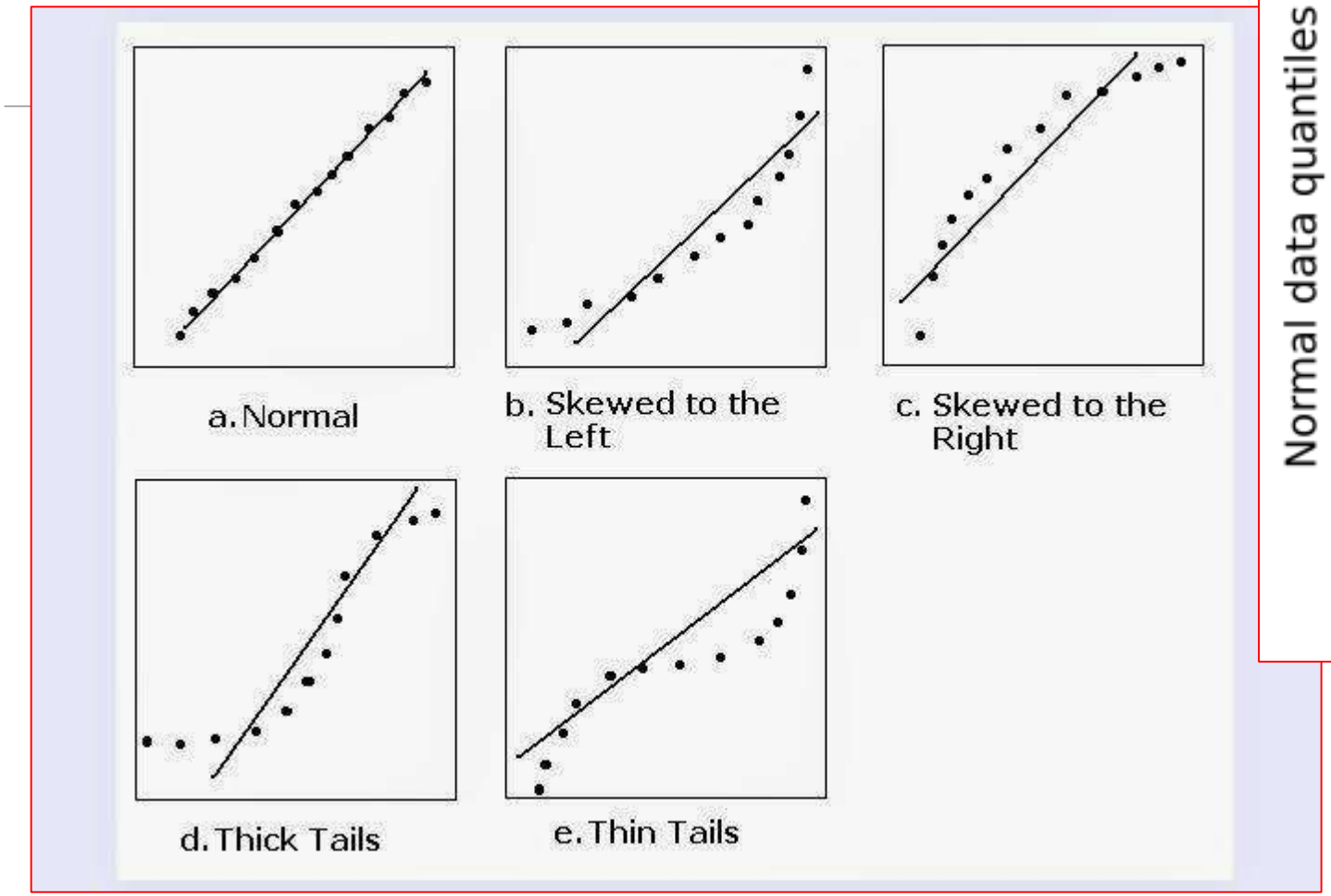
Линейность расположения точек на Q-Q графике не нарушается изменениями параметров сдвига и масштаба. На P-P графике изменения параметров сдвига и масштаба, напротив, могут нарушить линейность.

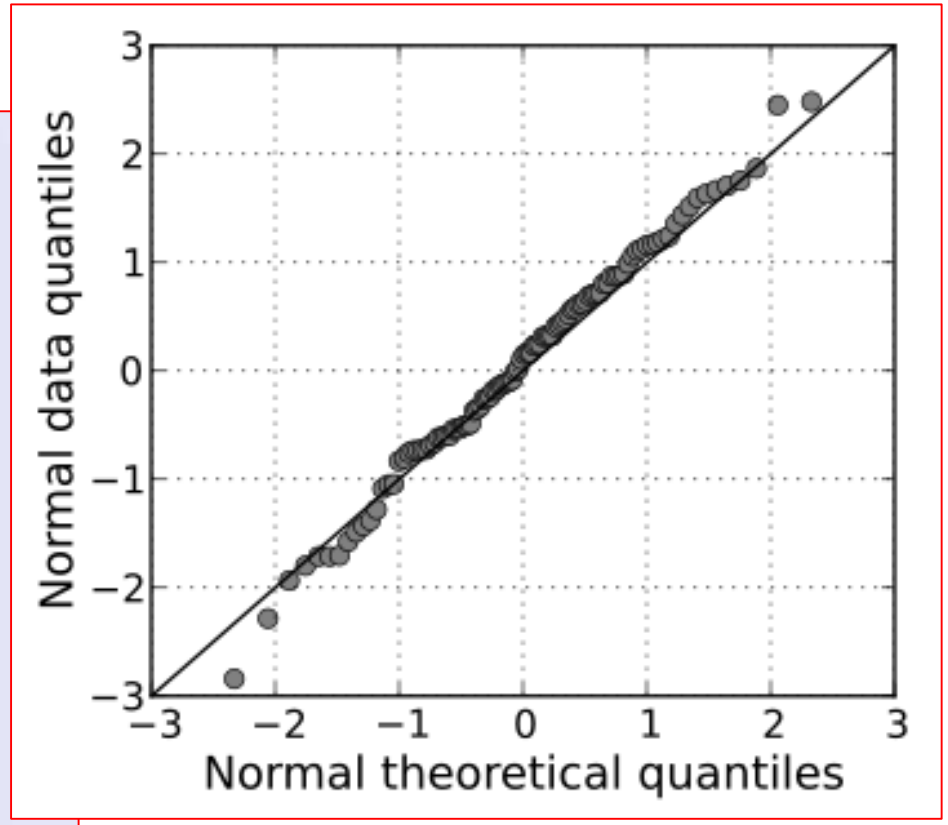
На Q-Q графике референтная линия зависит от параметров сдвига и масштаба теоретического распределения, эти параметры не надо задавать заранее, как в P-P графике, а можно, наоборот, определить из Q-Q графика! Остаток на нуле (intercept) и тангенс угла наклона определяют параметры сдвига и масштаба теоретического распределения. На P-P графике референтная линия – всегда диагональ прямоугольника или квадрата: y=x.

Достоинство P-P графиков – они лучше показывают различия распределений там, где плотность вероятности высока. Например, .для нормального распределения различия более наглядны на P-Pграфиках, чем на Q-Q графиках!

Достоинство Q-Q графиков – они лучше показывают различия распределений там, где плотность вероятности низка. Например, для распределений с «длинными хвостами» – различия лучше видны «на хвостах».

Графики P-P и Q-Q
P-P and Q-Q Plots





Роли вертикальной и горизонтальной осей в графиках **Q-Q** могут меняться!

Анализ связей между переменными (корреляционный анализ, частотный анализ)

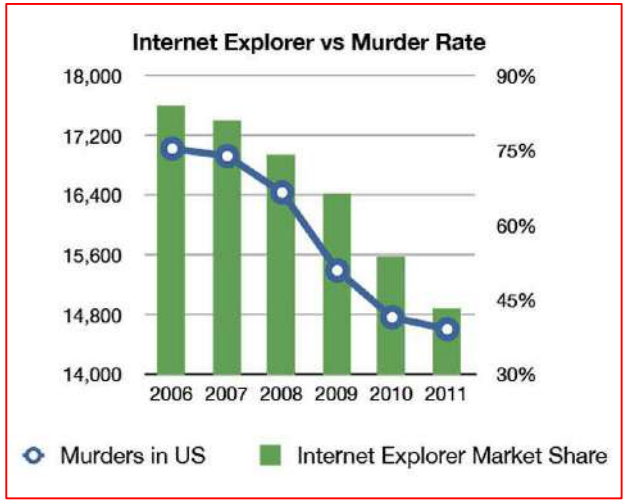
Дано: значения признаков X1,X2 измерены на объектах 1, . . . , n.

Насколько сильно признаки X1,X2 связаны между собой?

Набор данных, включает переменные (столбцы) X1,X2, число строк (наблюдений) равно n.

Рассматриваются числовые переменные – статистическая связь между ними, и нечисловые переменные – статистическая связь между ними

Статистическая взаимосвязь между числовыми случайными величинами - корреляция.



Корреляция - статистическая взаимосвязь между случайными величинами; **не является достаточным условием** причинно-следственной связи.

Корреляционный анализ: основные соотношения

Ковариации X и Y:

$$\mathrm{cov}(x,y)=M(xy)-M(x)M(y)$$

$$\mathrm{cov}(x,y)=\overline{xy}-\bar{x} * \bar{y}$$

Значения ковариаций зависят от единиц, в которых измерены величины X и Y

Коэффициент корреляции Пирсона (Pearson) – не зависит от единиц измерений

$$r(x,y)=\frac{\overline{xy}-\bar{x}*\bar{y}}{\sqrt{D(x)}\sqrt{D(y)}}=\frac{M\left(xy\right)-M(x)M(y)}{\sqrt{D(x)}\sqrt{D(y)}}$$

$$-1 \le r(x,y) \le 1$$

Другие обозначения:

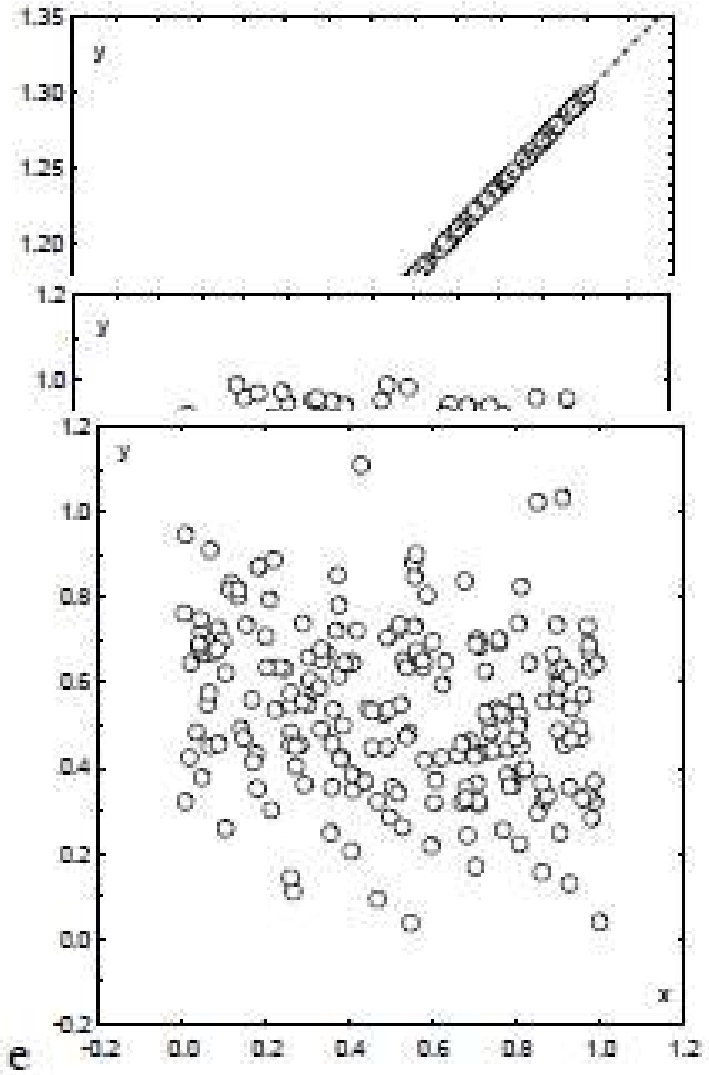
Коэффициент корреляции Пирсона случайных величин X_1 и X_2- мера силы линейной связи между ними:

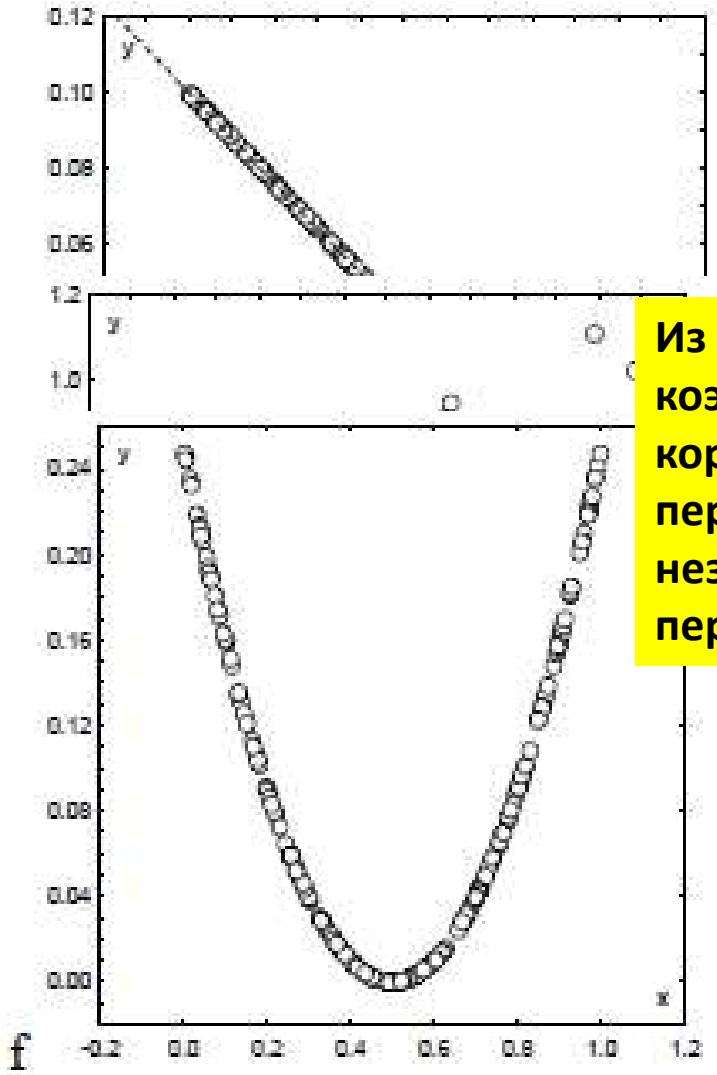
$$r_{X_1X_2} = \frac{\mathbb{E}\left((X_1 - \mathbb{E}X_1)\left(X_2 - \mathbb{E}X_2\right)\right)}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}}$$

$$r_{X_1X_2} \in [-1, 1].$$

Выборочный коэффициент корреляции Пирсона:

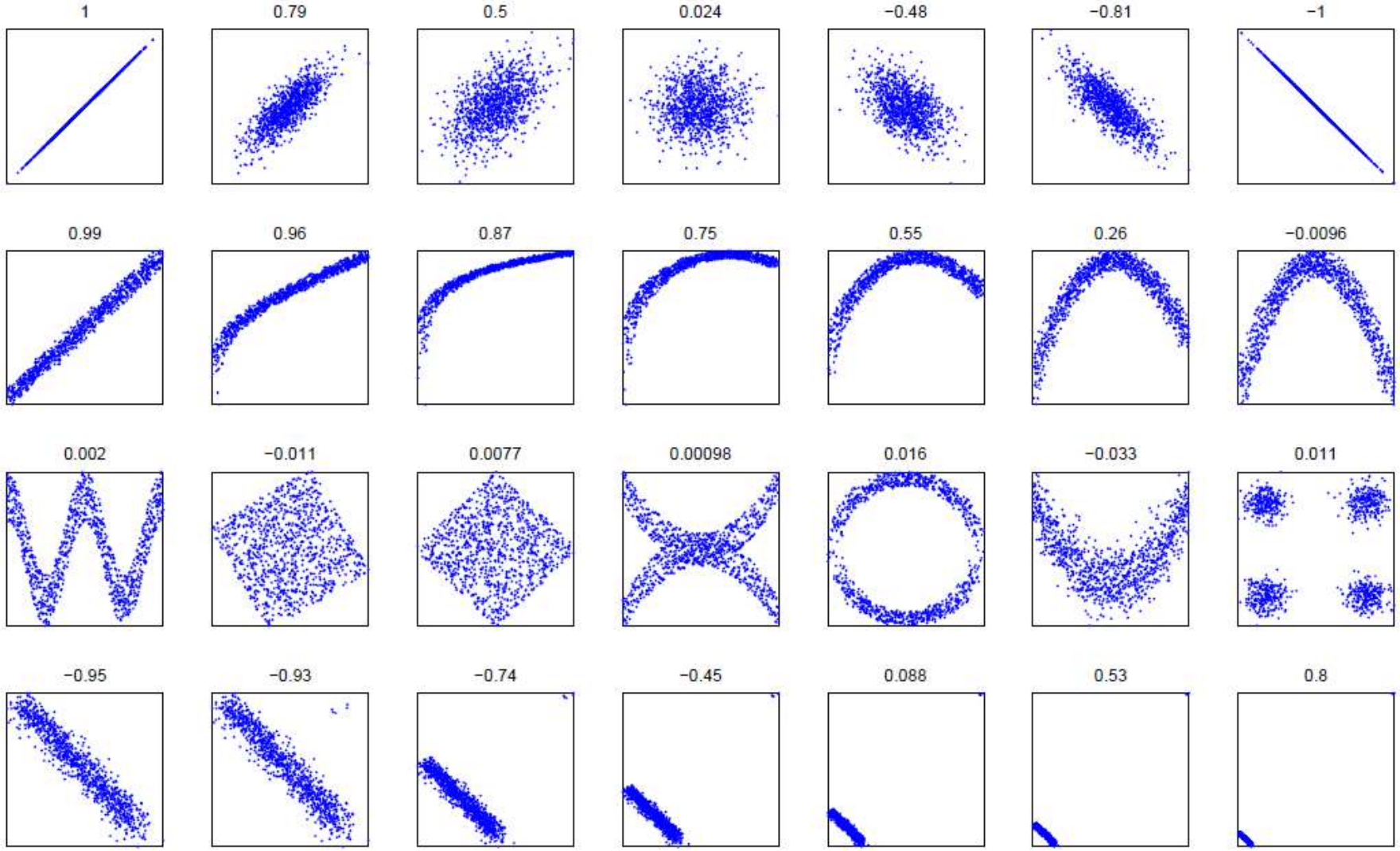
$$r_{X_1X_2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2 \sum_{i=1}^{n}\left(X_{2i} - \bar{X}_2\right)^2}}$$





Из равенства нулю коэффициента корреляции между переменными не следует независимость переменных!

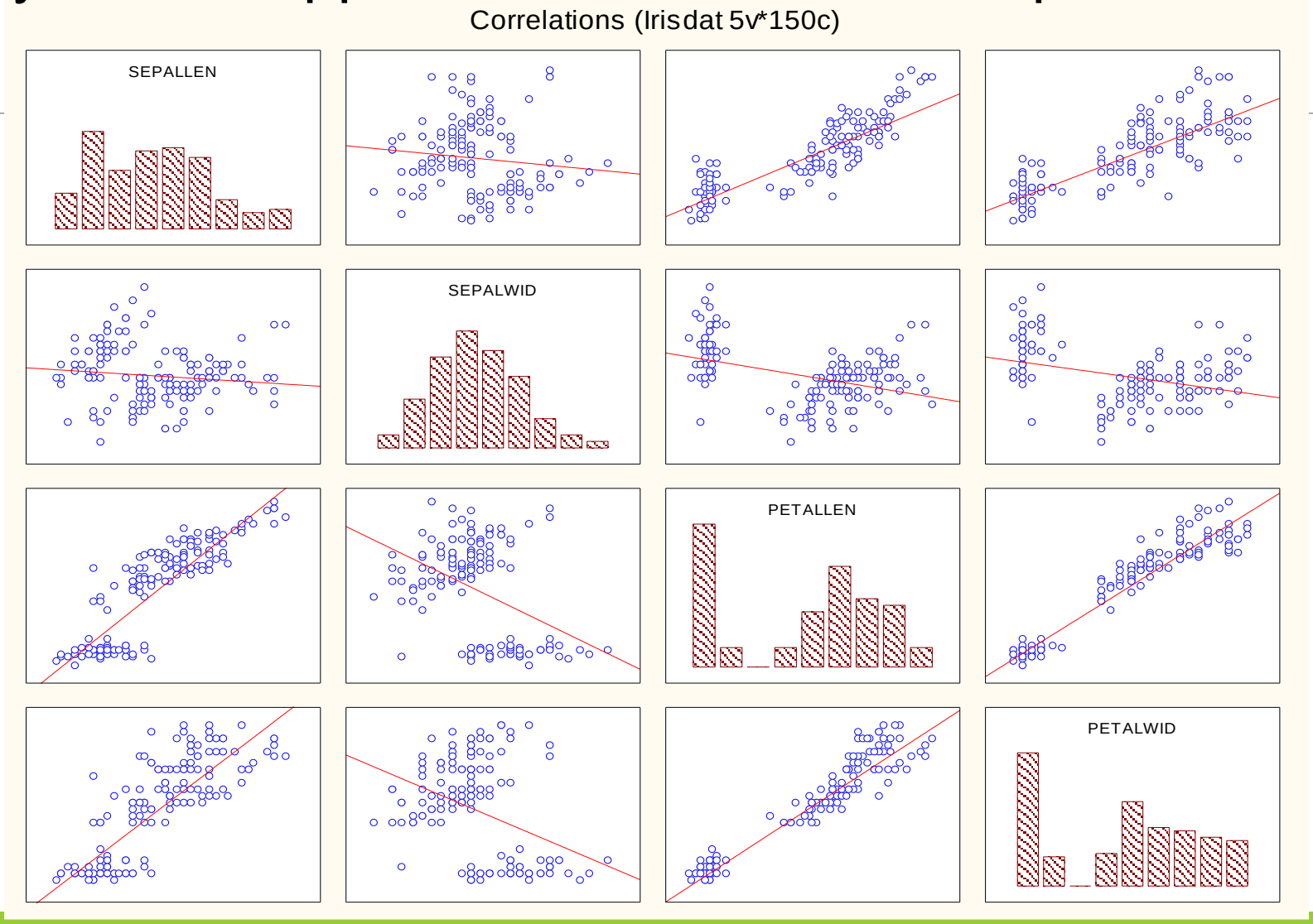
Корреляция Пирсона:



Из равенства нулю коэффициента корреляции между переменными не следует независимость переменных!

Добавление даже единичного выброса может существенно изменить коэффициент корреляции Пирсона!

Матричная диаграмма рассеяния (Matrix Scatterplot) – помогает судить о корреляциях нескольких переменных

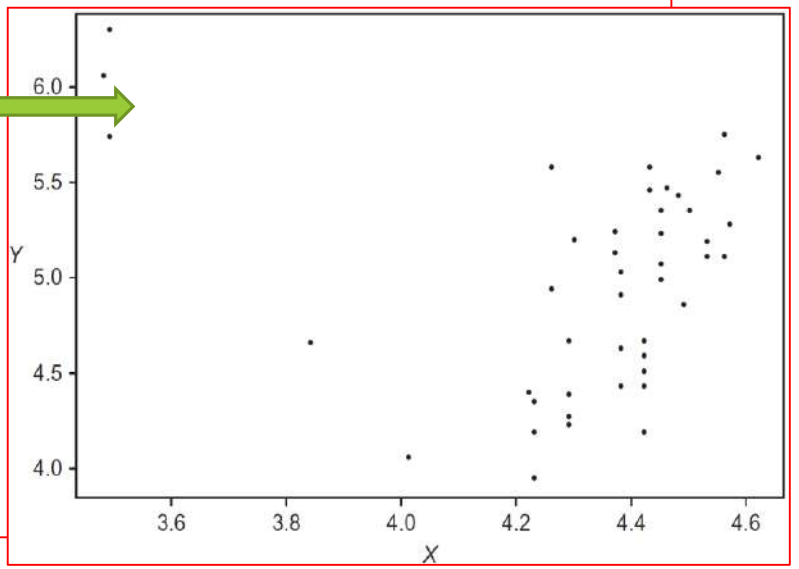


Корреляция Пирсона:

Недостатки выборочного коэффициента корреляции Пирсона:

- для распределений, отличных от нормального, перестаёт быть эффективной оценкой популяционного коэффициента корреляции;
- служит мерой только линейной взаимосвязи;
- неустойчив к выбросам.

Корреляция между логарифмами эффективной температуры на поверхности звезды (X) и интенсивности её света (Y) получается отрицательной ($r_{XY} = -0.21$) из-за наличия в выборке красных гигантов



Корреляционная матрица: пример пяти переменных

	N	ART	PRT	ARTG	PRTG
N	1.00	0.80	0.89	0.68	0.72
ART	0.80	1.00	0.98	0.96	0.97
PRT	0.89	0.98	1.00	0.91	0.93
ARTG	0.68	0.96	0.91	1.00	0.99
PRTG	0.72	0.97	0.93	0.99	1.00

	V	T
	r(x,v)	r(x,t)
	r(y,v)	r(y,t)
	r(z,v)	r(z,t)
	1	r(v,t)
		1

Ясно (из определения коэффициентов корреляции), что элементы корреляционной матрицы, расположенные симметрично по отношению к главной диагонали, равны. В связи с этим часто заполняется не вся корреляционная матрица, а лишь ее половина, считая от главной диагонали.

В случае, когда все случайные величины попарно
не коррелированы, все элементы корреляционной матрицы, кроме диагональных, равны нулю.

Корреляционная матрица: пример пяти переменных

	X	Y	Z	V	T
X	1	r(x,y)	r(x,z)	r(x,v)	r(x,t)
Y		1	(y,z)	r(y,v)	r(y,t)
Z			1	r(z,v)	r(z,t)
V				1	r(v,t)
T					1

Корреляционная матрица является симметричной, с единичной главной диагональю, положительно полуопределенной матрицей.

Квадратная матрица размерности n, не обязательно симметричная, называется положительно полуопределенной, если для любого вектора $Y=(y_1,y_2,\ldots,y_n)^T$ квадратичная форма $y^TRy>=0$ (не отрицательна). Квадратная матрица **R** положительно определена, если для любых $Y=(y_1,y_2,\ldots,y_n)^T$ квадратичная форма строго положительна.

Корреляционная матрица: пример пяти переменных

	N	ART	PRT	ARTG	PRTG
N	1.00	0.80	0.89	0.68	0.72
ART	0.80	1.00	0.98	0.96	0.97
PRT	0.89	0.98	1.00	0.91	0.93
ARTG	0.68	0.96	0.91	1.00	0.99
PRTG	0.72	0.97	0.93	0.99	1.00

	V	T
	r(x,v)	r(x,t)
	r(y,v)	r(y,t)
	r(z,v)	r(z,t)
	1	r(v,t)
		1

Ясно (из определения коэффициентов корреляции), что элементы корреляционной матрицы, расположенные симметрично по отношению к главной диагонали, равны. В связи с этим часто заполняется не вся корреляционная матрица, а лишь ее половина, считая от главной диагонали.

В случае, когда все случайные величины попарно не коррелированы, все элементы корреляционной матрицы, кроме диагональных, равны нулю.

Корреляционная матрица: пример пяти переменных

	X	Y	Z	V	T
X	1	r(x,y)	r(x,z)	r(x,v)	r(x,t)
Y		1	(y,z)	r(y,v)	r(y,t)
Z			1	r(z,v)	r(z,t)
V				1	r(v,t)
T					1

Корреляционная матрица является симметричной, с единичной главной диагональю, положительно полуопределенной матрицей.

Квадратная матрица размерности n, не обязательно симметричная, называется положительно полуопределенной, если для любого вектора $Y=(y_1,y_2,\ldots,y_n)^T$ квадратичная форма $y^TRy>=0$ (не отрицательна). Квадратная матрица **R** положительно определена, если для любых $Y=(y_1,y_2,\ldots,y_n)^T$ квадратичная форма строго положительна.

Проблема пропусков в данных и корреляционная матрица

Возникает, когда число переменных, для которых ведется расчет корреляций, больше двух

При пропуске значения по одной переменной:

- Исключать из расчетов ВСЕ наблюдение?
- Исключать из расчетов суммирование только для тех пар, в которых присутствует переменная с отсутствующим значением?

Во втором варианте корреляционная матрица теряет свойство положительной полуопределенности, и ее использование для решения многих задач **многомерного** анализа становится невозможным

Программы всегда предусматривают выбор одного из вариантов учета пропущенных значений при расчете корреляций и в других

Коэффициент корреляции Спирмена (Spearman):

Коэффициент корреляции Спирмена $\rho(X_1X_2)$ случайных величин X_1 и X_2 - мера **силы монотонной** корреляции между ними; **равен коэффициенту корреляции Пирсона между рангами наблюдений.**

Выборочный коэффициент корреляции Спирмена:

Ранг (RANK) -
Номер наблюдения, присвоенный ему при процедуре ранжировки. Наблюдения ранжируют (им присваиваются ранги), упорядочивая их по величине и назначая им номера (называемые рангами), соответствующие их месту (номеру в упорядоченном ряду) в упорядочении. Обычно наблюдения ранжируются от меньшего к большему.

$$\rho_{X_1X_2} = \frac{\sum_{i=1}^{n} \left(\operatorname{rank}(X_{1i}) - \frac{n+1}{2}\right)\left(\operatorname{rank}(X_{2i}) - \frac{n+1}{2}\right)}{\frac{1}{12}(n^3 - n)} =$$

$$= 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(\operatorname{rank}(X_{1i}) - \operatorname{rank}(X_{2i})\right)^2,$$

$\operatorname{rank}(X_{1i}), \operatorname{rank}(X_{2i})$ -ранги i-х наблюдений в соответствующих выборках

Корреляция Спирмена равна 1 только когда ранги абсолютно совпадают
Корреляция Спирмена равна -1 только когда ранги абсолютно не совпадают (характер их изменений противоположен у переменных)

$$S=\sum_{i=1}^{n}(QX-QY)^2$$

$$\rho_{Spearman}=1-\frac{6S}{n^3-n}=1-\frac{6(\sum_{i=1}^{n}(QX-QY)^2}{n^3-n}$$

Непараметрические меры связи между переменными. Ранговые корреляции

- Тау Кендалла (Kendall):
  - Для X вводим псевдопеременную QX, значения которой равны рангу соответствующего значения переменной X, для Y вводим соответственно QY, значения которой равны рангу соответствующего значения переменной Y
  - Вычисляем число перестановок наблюдений , необходимых для QY, чтобы последовательность рангов QY совпала с последовательностью рангов QX
  - К –число таких перестановок,
  - К может быть от 0 до n(n-1)/2:

$$\tau_{\text{Kendall}} = 1 - \frac{4K}{n(n-1)}$$

Коэффициент корреляции Кендалла (Kendall TAU):

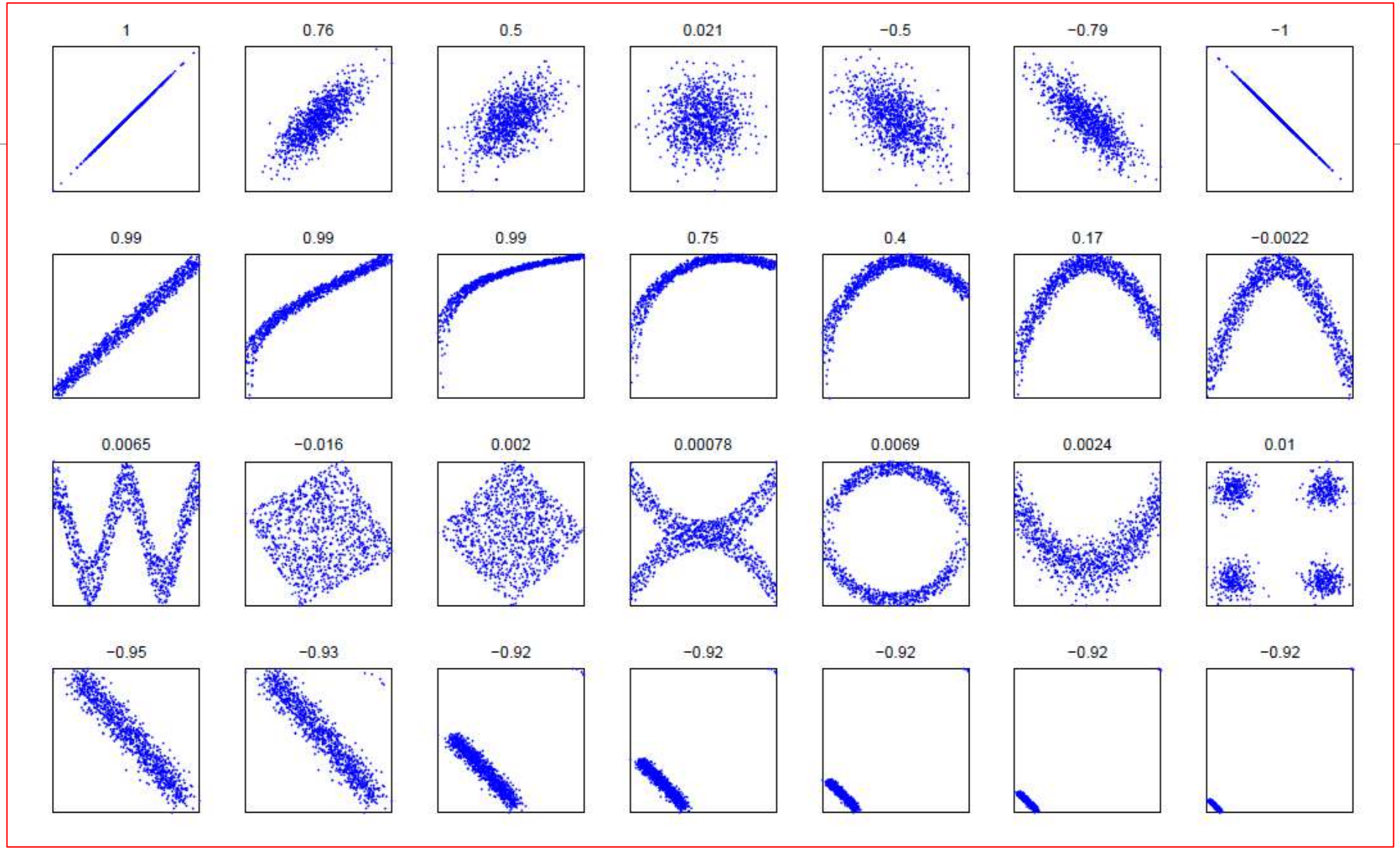
Коэффициент корреляции Кендалла $\tau X_1 X_2$ случайных величин X_1 и X_2 - мера их взаимной неупорядоченности; также оценивает силу монотонной корреляции между величинами.

Выборочный коэффициент корреляции Кендалла:

$$\tau_{X_1 X_2} = 1 - \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^{n} [[X_{1i} < X_{1j}] \neq [X_{2i} < X_{2j}]] = \frac{C - D}{C + D}$$

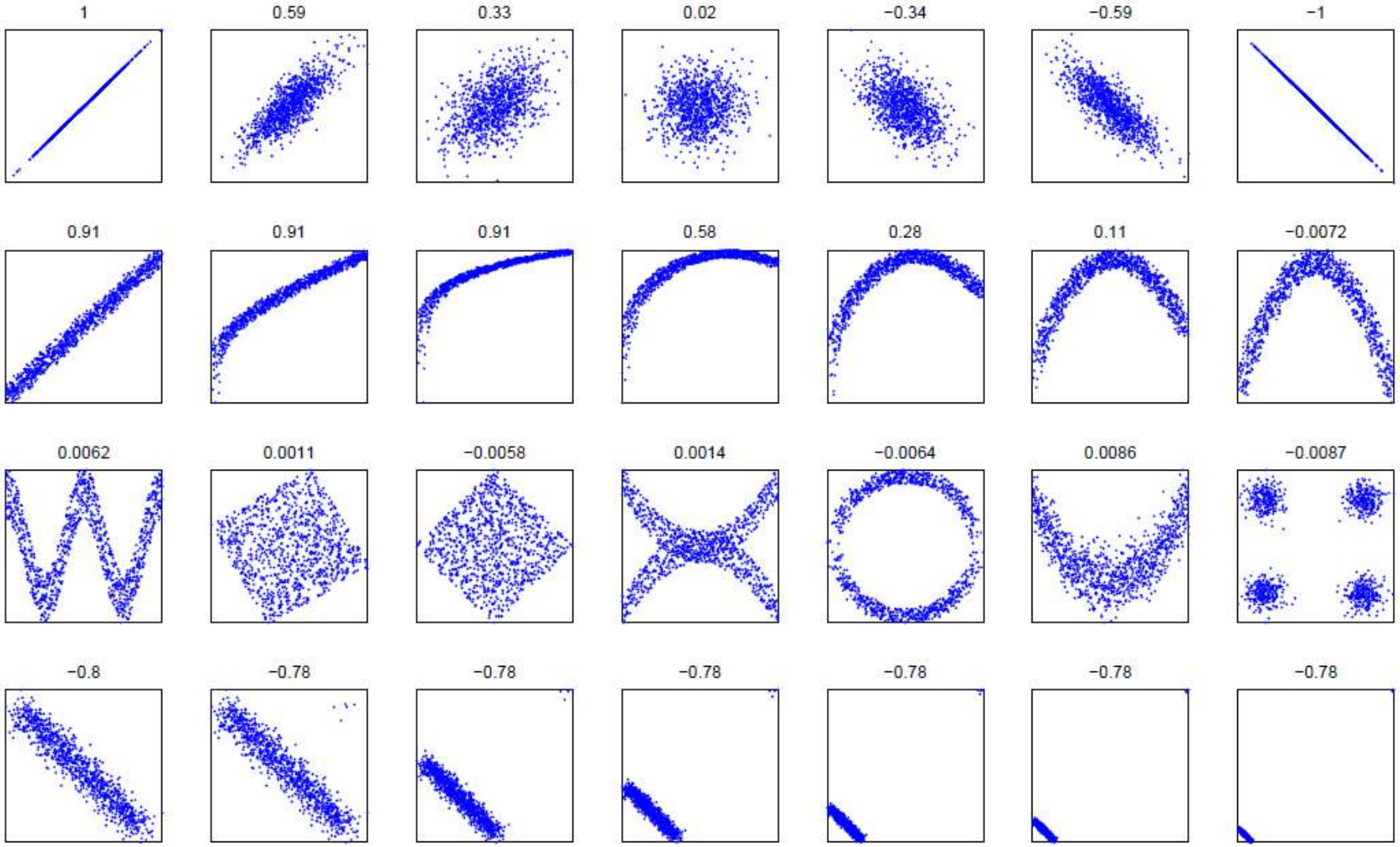
где C - число согласованных пар, D - число несогласованных пар

Коэффициент корреляции Спирмена (Spearman):



Коэффициенты корреляции Спирмена и тау Кендалла – практически нечувствительны к появлению грубых ошибок (выбросов) в данных

Коэффициент корреляции Кендалла (Kendall):



Коэффициенты корреляции Спирмена и тау Кендалла – практически нечувствительны к появлению грубых ошибок (выбросов) в данных

Статистическое ПО
Statistical Software

Наименование	Адрес	Последняя версия (февр. 2018)	Примечание
SAS	**www.sas.com**	**v.9.4 и SAS Viya**	
SPSS	www.spss.com	25 (8 августа 2017)	(аббревиатура англ. «Statistical Package for the Social Sciences» — «статистический пакет для социальных наук») В 2010 поглощен IBM и стал одним из компонентов IBM Software Group Business Analytics Portfolio
Statistica	www.statsoft.com, www.statsoft.ru	v.13.3	Statsoft с марта 2014 г стала частью DELL Software, сейчас перекуплена и поглощена интеграционной компанией TIBCO
SYSTAT	www.systat.com	v.12 (?)	MYSTAT – студенческий вариант, SigmaPlot, SigmaStat, AutoSignal,
JMP	www.sas.com	JMP 13.2.1 и JMP Pro 13.2.1	
MINITAB	www.minitab.com	v.17	

Аналитико-статистическое и другое ПО

S-Plus
StatGraphics
MATHCAD
MATLAB
R Software (www.r-project.org) – свободно распространяемый продукт
EXCEL (?) имеется вкладка «Анализ данных», обладает рядом неудобств, и, строго говоря, к аналитико-статистическому ПО отнесен быть не может

Статистическое ПО – и не только статистика в традиционном понимании!
Statistical Software

Имеют все необходимые компоненты для решения широкого круга проблем, связанных с Data Science

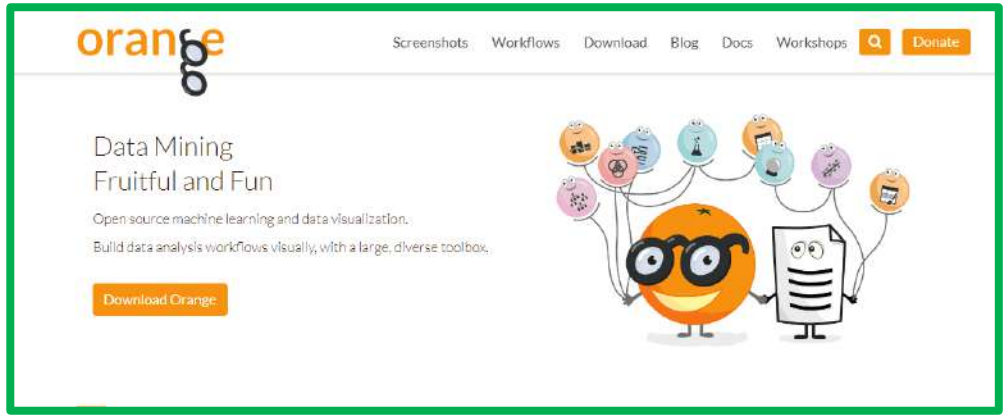
Каждый участник этой «троицы» – для миллионов своих приверженцев – это «наше все»

Владение хотя бы одной компонентой – требование к Data Scientist, Data Analyst

Решения SAS для нас с апреля 2022 г. недоступны….

Будем пробовать работать с ORANGE и RAPID MINER

https://orangedatamining.com



Спасибо за внимание!

Лекция (состоящая из трех частей плюс практические вопросы !) - окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

ТЕЛ. (484) 5830658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

Регрессионный анализ – позиционирование *где его место?*

Имеется множество объектов. Каждый объект описывается вектором его наблюдаемых характеристик (признаков) $x \in X$ и скрытых характеристик $y \in Y$ (целевая переменная).

Существует (на всем множестве – на генеральной совокупности) некоторая функция $f : X \to Y$

Задача: имея **ограниченный** набор объектов (обучающая выборка), построить функцию $a: X \to Y$, приближающую f на всем множестве объектов (на генеральной совокупности).

Какие бывают признаки (переменные в анализе, характеристики объектов, и т.д.)?

- Вещественный признак – принимает вещественные значения
- Бинарный признак - может принимать 2 значения
- Категориальный признак - может принимать конечное число значений
- Порядковый признак – упорядоченный категориальный признак

Регрессионный анализ – позиционирование *где его место?*

- Пусть обучающая выборка объектов выборка объема (размера) N .
  Обозначим:

$$x_1, \ldots, x_N = X_{train}, \{y_1, \ldots, y_N\} = Y_{train}$$

Тогда выделим следующие возможные типы задач:

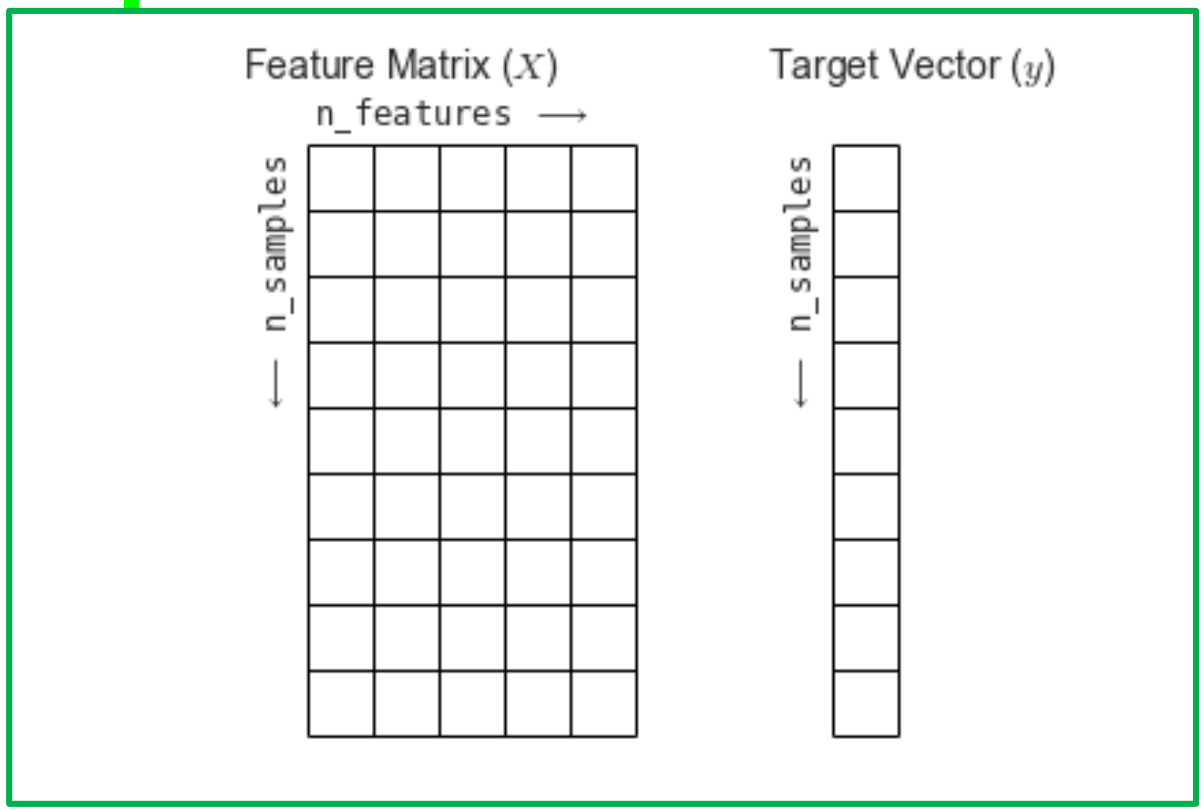
- Обучение с учителем (supervised learning): Известны X_{train}, Y_{train}
- Обучение без учителя (unsupervised learning): Известно только X_{train}
- Частичное обучение (semi-supervised learning): Известно для всех ***N*** объектов X_{train} и для некоторых (!!!!) ***l*** объектов (***l<N)*** объектов из X_{train} - известна целевая переменная ***y***

Регрессионный анализ – позиционирование *где его место?*

- По типу целевой переменной задача *обучение с учителем* разбивается на несколько классов.
- Будем различать 2 постановки:
- Если ***Y*** может принимать значения из конечного множества чисел – ***Y={1,2, …,M}*** - тогда это ***задача классификации***
- Если ***Y*** может принимать значения из конечного множества нечисловых значений - тогда это тоже ***задача классификации – нечисловые значения – их конечное множество - можно кодировать числами из конечного множества***
- Если ***Y*** может принимать произвольные значения из множества вещественных чисел– ***Y=$\mathbb{R}$*** - тогда это ***задача регрессии***

Регрессионный анализ – позиционирование
где его место?

- ***В любом варианте задача подразумевает наличие матрицы данных, содержащей признаки X – ее строки – объекты, ее столбцы- признаки***

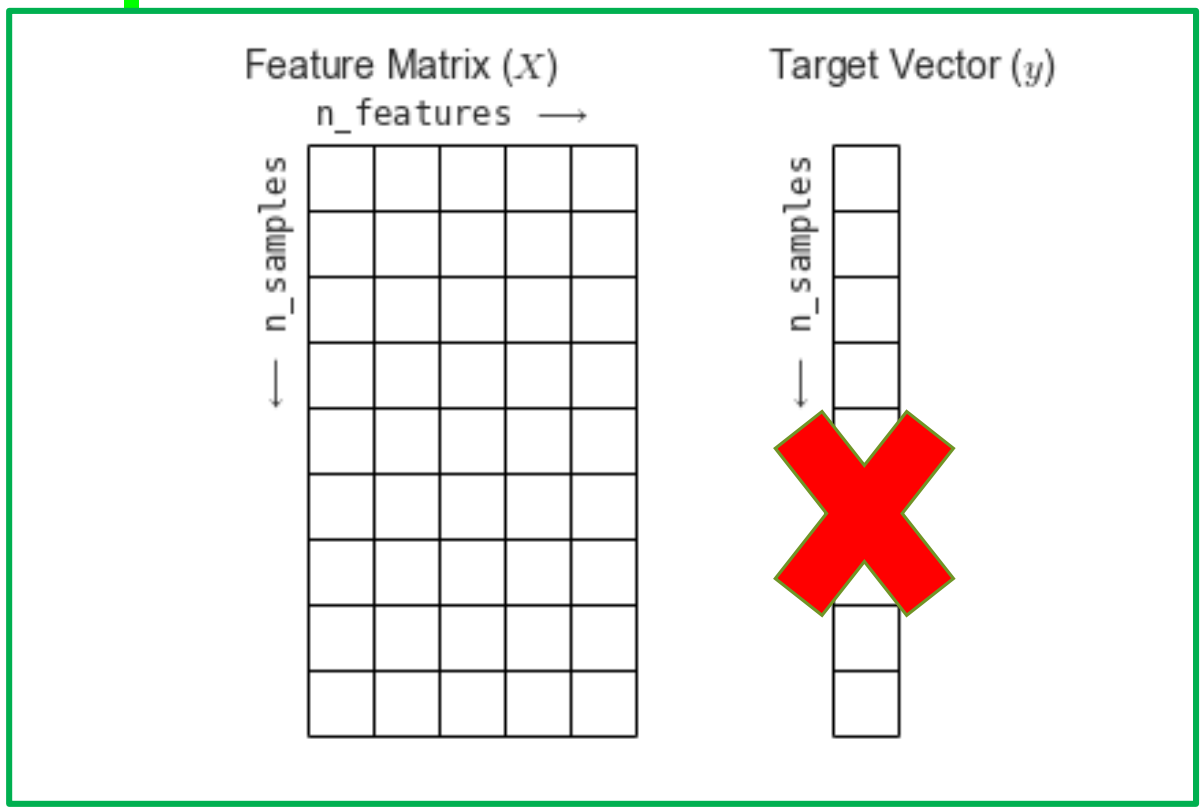


Если вариант задачи – обучение с учителем, то подразумевается наличие вектора значений целевой переменной Y

Y – значения из конечного множества чисел - классификация
Y - произвольные значения из множества вещественных чисел - регрессия

Регрессионный анализ – позиционирование
где его место?

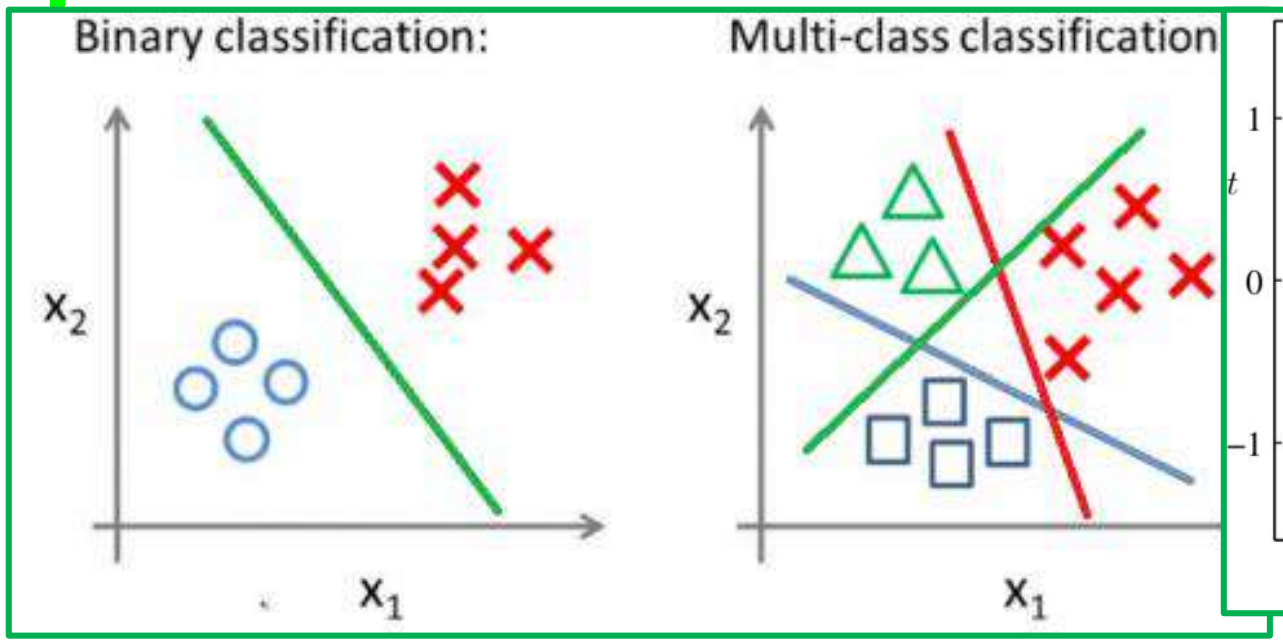
- ***В любом варианте задача подразумевает наличие матрицы данных, содержащей признаки X – ее строки – объекты, ее столбцы- признаки***

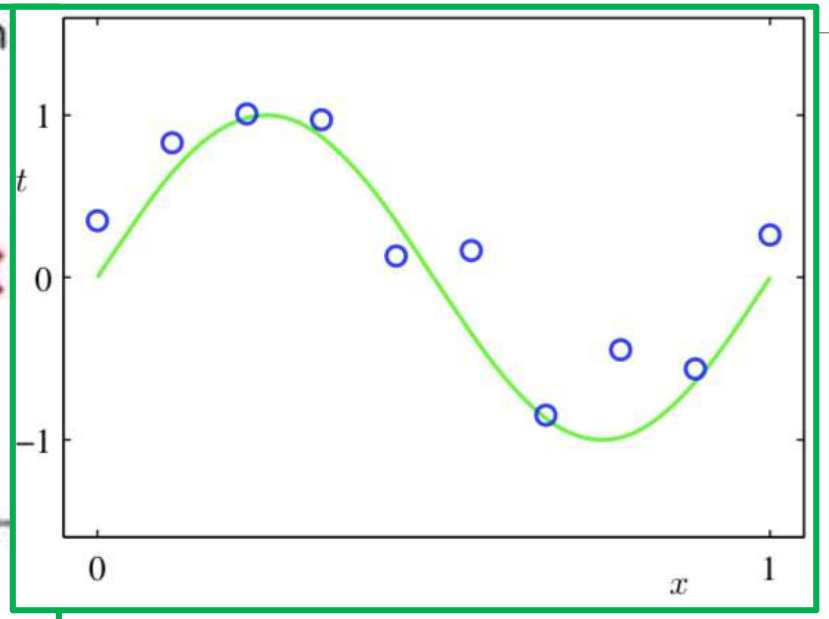


Если вариант задачи – обучение без учителя, или самообучение, то НЕ подразумевается наличие вектора значений целевой переменной Y

Регрессионный анализ – позиционирование
где его место?

- ***При обучении с учителем:***





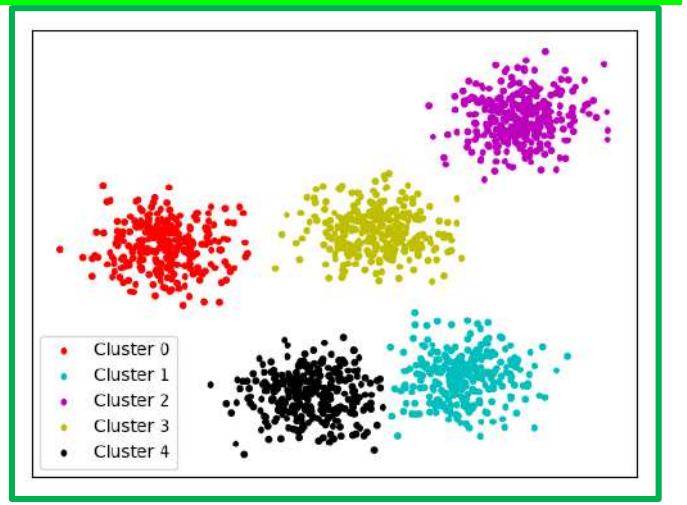
Классификация

Регрессия

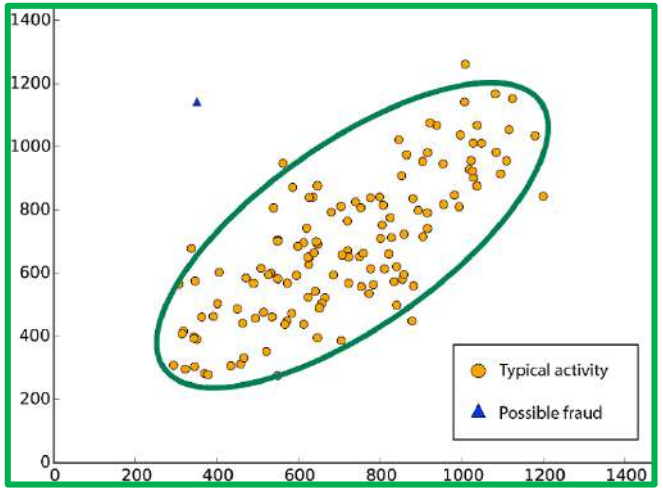
Регрессионный анализ – позиционирование
где его место?

- ***При обучении без учителя:***

- ***Кластеризация –*** разбиение объектов на такие группы, что объекты в одних группах похожи, а в разных - отличаются
- • ***Поиск аномалий -*** поиск объектов, отличающихся от всех остальных
- • ***Снижение размерности*** -- уменьшение числа признаков



Кластеризация



Поиск аномалий

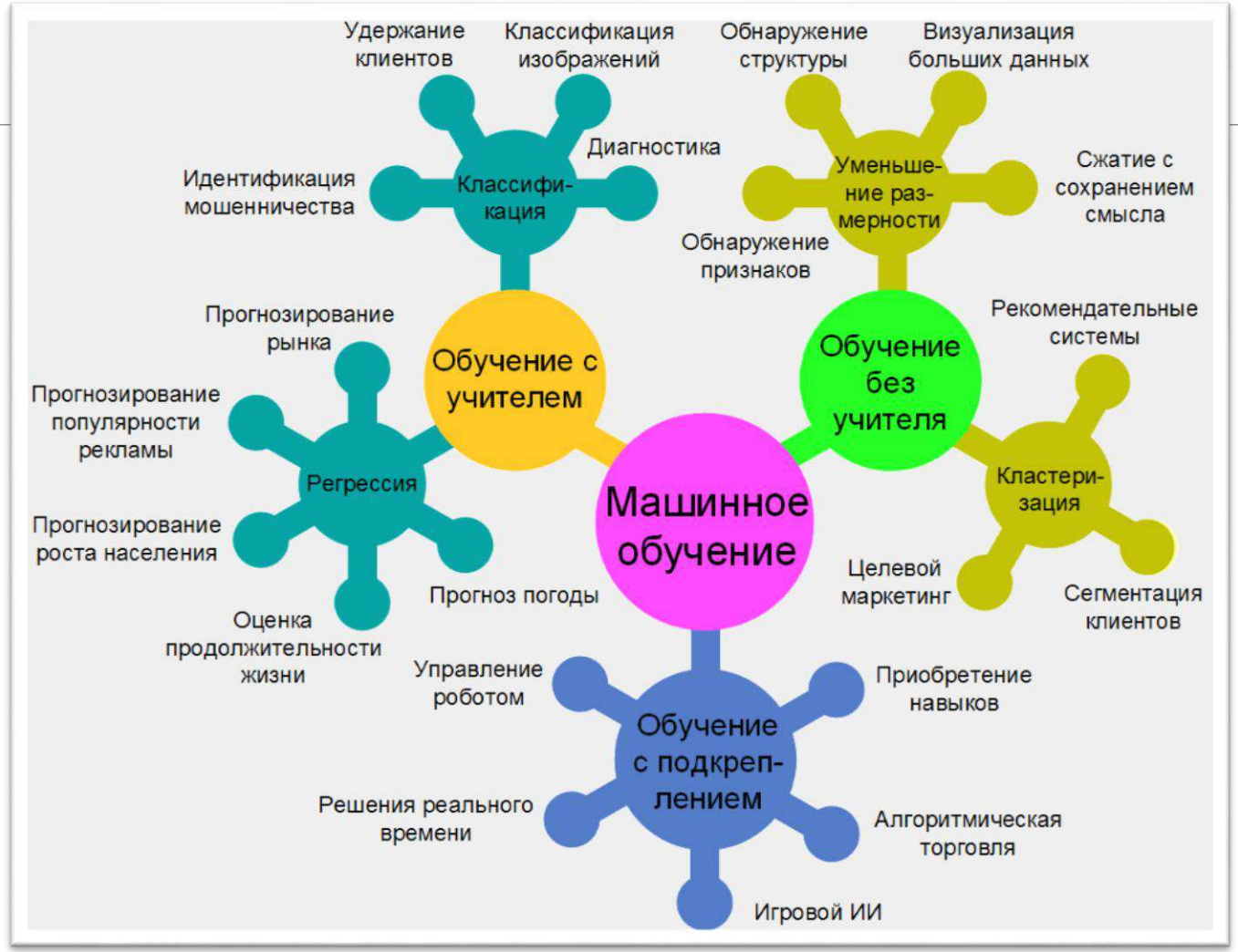
На будущее: как решать прикладные задачи:

Прежде чем делать прикладную задачу, нужно разобрать ее постановку!

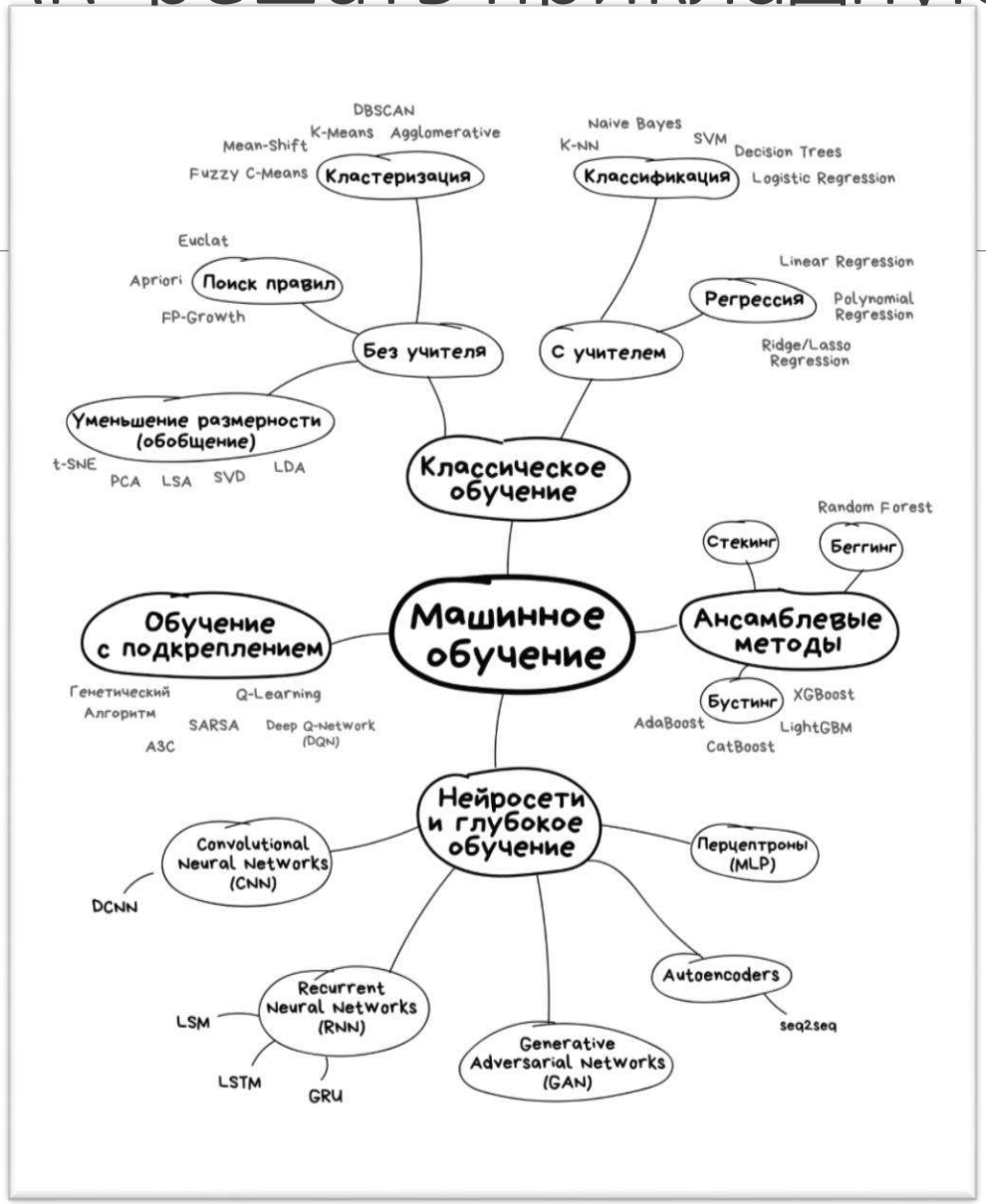
Делаем по принципу:

1. Что является объектом в задаче?
2. Что является целевой переменной?
3. С учителем или без?
4. Регрессия или классификация? Кластеризация или поиск аномалий?
5. Какие данные нам нужны?
6. Какие признаки нужно извлечь?

На будущее: КАК решать прикладную задачу? :



На будущее: КАК решать прикладную задачу? :



На будущее: как понять, хорошо ли решили прикладную задачу, хорош ли алгоритм a:

Функция потерь (loss function) $L(a, x, y)$ - неотрицательная функция, показывающая величину ошибки алгоритма a на объекте x с целевой переменной y.

Функционал качества $Q(a, X, Y) = \frac{1}{N} \sum_{i=1}^{N} L(a, x_i, y_i), x_i \in X, y_i \in Y$

Принцип минимизации эмпирического риска:

$a^* = \underset{A}{\operatorname{argmin}}\, Q(a, X_{train}, Y_{train})$, где A - семейство алгоритмов.

Примеры функций потерь:

- Классификация - $L(a, x, y) = [a(x) = y]$
- Регрессия - $L(a, x, y) = |a(x) - y|$

Формула обучения:

Learning = Representation + Evaluation + Optimization

На будущее: как понять, хорошо ли решили прикладную задачу, хорош ли алгоритм a:

Самый важный вопрос: открыли ли мы закон природы или просто подогнали наш алгоритм $a(x)$ под обучающую выборку?

Не обязательно, что $\underset{A}{\operatorname{argmin}} Q(a, X_{train}, Y_{train})$ - полезный алгоритм.

? Можете придумать пример алгоритма, у которого ошибка на обучении 0, но он совершенно бесполезен?

Это несложно!!

Финальный алгоритм проверяем на контрольной выборке X_{test}, Y_{test}, которую он раньше не видел.

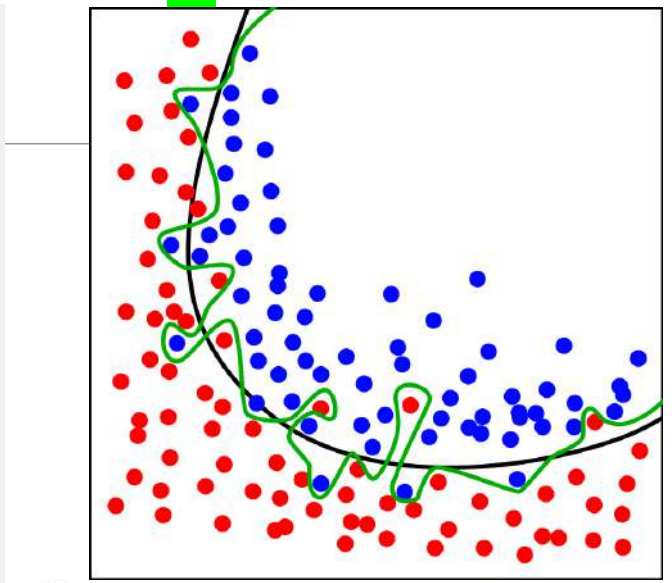
На будущее: как понять, хорошо ли решили прикладную задачу, хорош ли алгоритм a:

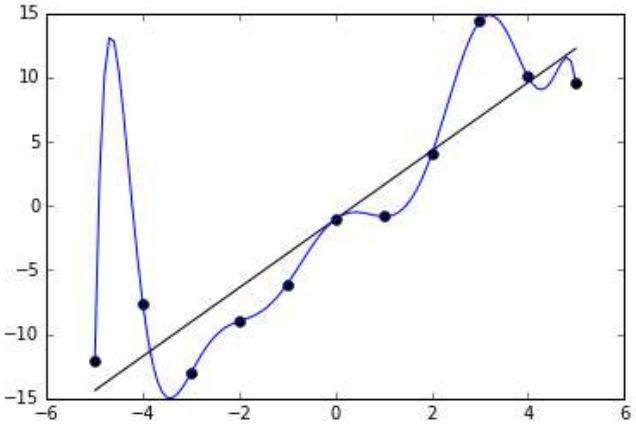
Проблема **переобучения** - значения $Q(a, X_{train}, Y_{train})$ значительно меньше, чем значение $Q(a, X_{test}, Y_{test})$ на контрольной выборке.

Если $Q(a, X_{test}, Y_{test})$ примерно равна $Q(a, X_{train}, Y_{train})$, то говорят, что алгоритм обладает **обобщающей способностью**.

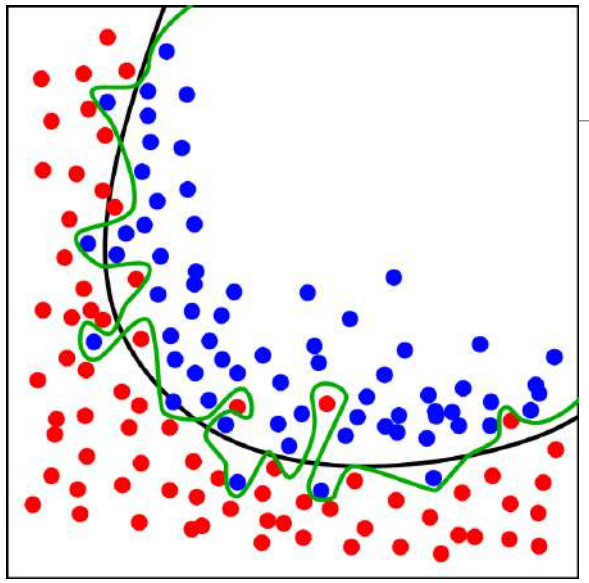
Переобучение есть всегда из-за индуктивной постановки задачи - нахождение закона природы по неполной выборке!

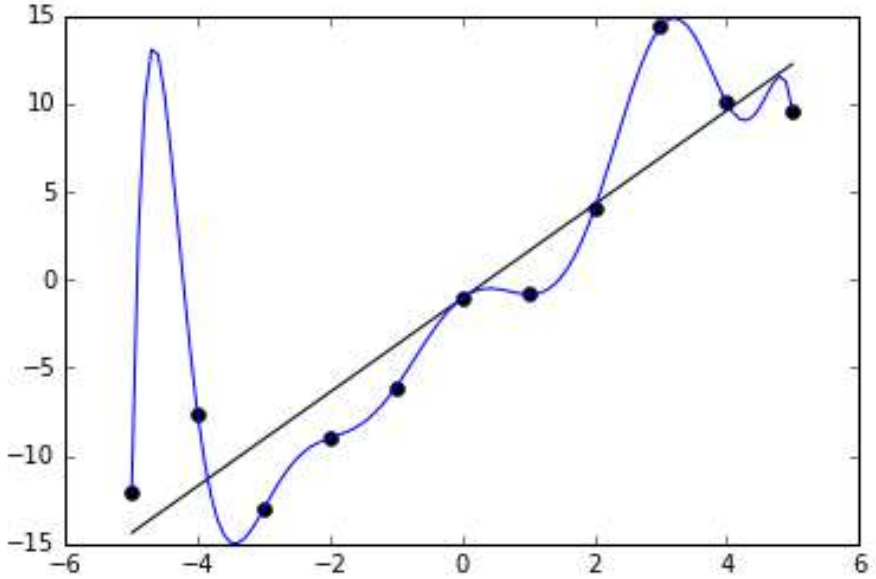
Но еще она может быть из-за излишней **сложности** модели.





На будущее: как бороться с переобучением?





- • **Искать** больше данных – хороший совет , но не всегда реальный!!!
- • **Упрощать** семейство решений , используя экспертные знания о структуре решения.
- **Не бояться** ошибок на обучающей выборке, если алгоритм существенно упрощен
- **Помнить:** Без знания предметной области невозможно решать прикладную задачу!
- **Помнить:** Нет идеального алгоритма, решающего ***все задачи лучше других!***. (теоремы Вольперта и МакРеди - The No Free Lunch Theorem)

Регрессионный анализ – *где его место?*
Применяем ранее сформулированную методологию

Решаем задачу обучения с учителем

Функционал качества

$$Q(a, X, Y) = \frac{1}{N}\sum_{i=1}^{N} L(a, x_i, y_i), x_i \in X, y_i \in Y$$

Принцип минимизации эмпирического риска:

$$a^* = \underset{A}{\operatorname{argmin}}\, Q(a, X_{train}, Y_{train}), A$$ — семейство алгоритмов.

Формула обучения:

Learning = Representation + Evaluation + Optimization

В нашей выборке есть Y!

Регрессионный анализ – *где его место?*

Функций $a(x)$ которые идеально описывают обучающую выборку бесконечно много. Нужно сузить функциональное пространство перебора. **Параметризуем** искомую функцию

модель $a(x, w)$ описывается вектором весов w.

Тогда задача превращается в поиск весов:

$$w^* = \arg\min_w Q(w, X_{train}, y_{train})$$

Примеры:

$$a(x, w) = x_1 w_1 + x_2 w_2$$

$$a(x, w) = x_1 x_2 x_3 w_1$$

$$a(x, w) = 1[x_1 < w_1]$$

В том числе, алгоритм, который просто запоминает обучающую выборку, воспроизведет ее ИДЕАЛЬНО, без ошибок! Но это не является обучением!

Надо сразу оговорить, в каком классе функций ищем решение
А затем в этом оговоренном классе искать конкретную функцию – конкретный $w=w^*$, обеспечивающий минимум

Когда получаем ЛИНЕЙНУЮ модель?

Пусть объект описывается D признаками $f_1, f_2, \ldots f_D$.
Тогда модель:

$a(x, w) = w_0 + \sum_{j=1}^{D} f_j w_j$ называется **линейной моделью**,

где w — D-мерный вектор признаков $w_1, w_2, \ldots w_D$

Далее будем считать, что в векторе признаков есть тождественно равный единице признак f_0, тогда формула упростится до:

Representation: $a(x, w) = \sum_{j=0}^{D} f_j w_j = \boldsymbol{x} \cdot \boldsymbol{w}$

Параметры модели интерпретируемы. w_i — значение, на которое изменится предсказание, если признак f_i увеличить на единицу.

Какая гипотеза лежит в основе линейной модели?

Введение в вектор признаков признака (столбца), тождественно равного единице, позволяет ввести в вид искомого решения свободный член - INTERCEPT

Когда получаем ЛИНЕЙНУЮ РЕГРЕССИОННУЮ модель?

Если целевая переменная Y - вещественное число, то такую модель называют линейной регрессией. Для нее функция потерь может иметь самые различные конкретные выражения:

$$Q(a, X, Y) = \frac{1}{N}\sum_{i=1}^{N} L(a, x_i, y_i), x_i \in X, y_i \in Y$$

Функции потерь:

1. **Квадратичная**(Q — MSE)

$$L(a, x, y) = (a(x) - y)^2$$

2. **Абсолютная**(Q — MAE)

$$L(a, x, y) = |a(x) - y|$$

3. **Логарифмическая** (Q — MSLE)

$$L(a, x, y) = (\log(a(x) + 1) - \log(y + 1))^2$$

4. **Абсолютная-процентная** (Q —MAPE) $L(a, x, y) = \frac{|a(x) - y|}{y}$

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} L(a(x_i, w), y_i) = \underset{w}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{x_i} \cdot \boldsymbol{w}, y_i), x_i \in X, y_i \in Y$$

Иначе говоря, если имеем целевую функцию-вещественное число (левая часть равенства) ПЛЮС линейную модель (правая часть равенства), то имеем ЛИНЕЙНУЮ РЕГРЕССИОННУЮ модель!!!

Может рассматриваться огромное разнообразие функций потерь – здесь приведены лишь простейшие!

Мы должны в классе a найти оптимальное решение w^*

Получаем оптимизационную задачу – как ее решить?

Для квадратичной функции потерь – идем общим путем:

$$Q(X,\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x_i}\cdot\boldsymbol{w} - y_i)^2 = \frac{1}{N}||X\cdot\boldsymbol{w} - \boldsymbol{y}||^2$$

X — матрица (N,D), w — вектор весов $(D,1)$, y — вектор ответов $(N,1)$

$X\cdot\boldsymbol{w}$ — вектор предсказаний $(N,1)$

Необходимое условие минимума:

$\nabla_{\boldsymbol{w}}Q(\boldsymbol{w})$ — градиент, вектор частных производных. Необходимое условие минимума — градиент равен нулю.

$$Q(\boldsymbol{w}) = \frac{1}{N}||X\cdot\boldsymbol{w} - \boldsymbol{y}||^2 = \frac{1}{N}(X\cdot\boldsymbol{w} - \boldsymbol{y})^T(X\cdot\boldsymbol{w} - \boldsymbol{y})$$

Линейная регрессия - невероятно популярный алгоритм машинного обучения.

Плюсы алгоритма:

- Быстро учится
- Быстро предсказывает
- Легко интерпретируется
- Легко хранить в памяти
- Легко применять с дифференцируемой функцией потерь

Весомый минус –

- Не способен учитывать нелинейные зависимости в данных.

Как можно улучшить модель?

$$a(x, w) = \sum_{j=0}^{D} f_j w_j = \boldsymbol{x} \cdot \boldsymbol{w}$$

ПОМНИМ:
Невозможно сделать правильное измененное признаковое пространство без понимания самой задачи!

Параметры модели интерпретируемы. Это большой плюс линейной модели. **W_i - значение, на которое изменится предсказание, если признак f_i увеличить на единицу.**

Категориальные признаки кодируем:

- One-hot кодирование - категориальный признак с k значениями превращаем в k бинарных признаков!
- Кодирование через целевую переменную (нельзя включать переменную самого объекта)
- Кодирование через вещественные признаки

Для вещественных признаков применяем нелинейные функции - возводим в степень, берем синус и т.д.

Учитываем взаимодействия:

- Пару вещественных перемножаем, делим и т.д.
- Для пары бинарных используем логические операции

Изменяем масштаб признаков (вспомним ковариацию и корреляцию!)

Для исходных числовых признаков применяем стандартизацию и нормализацию:

- Стандартизация $f_j = \frac{f_j - mean(f_j)}{std(f_j)}$
- Min-max нормализация $f_j = \frac{f_j - min(f_j)}{max(f_j) - min(f_j)}$

Как ещё можно улучшить модель?

$$a(x, w) = \sum_{j=0}^{D} f_j w_j = \boldsymbol{x} \cdot \boldsymbol{w}$$

Изменяем вид целевой функции – вводим дополнительные ограничения!
Например, проводим регуляризацию:

Хотим еще сузить функциональное пространство $a(x)$, чтобы улучшить обобщающую способность.

Накладываем дополнительные штрафы за превышение вводимых нами ограничений, если решение удаляется от нашего представления о правильном решении

Например, можем дополнительно штрафовать, если у нас большие по абсолютной величине коэффициенты w_i

Это тема, требующая отдельного рассмотрения!

Спасибо за внимание!

Лекция -окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

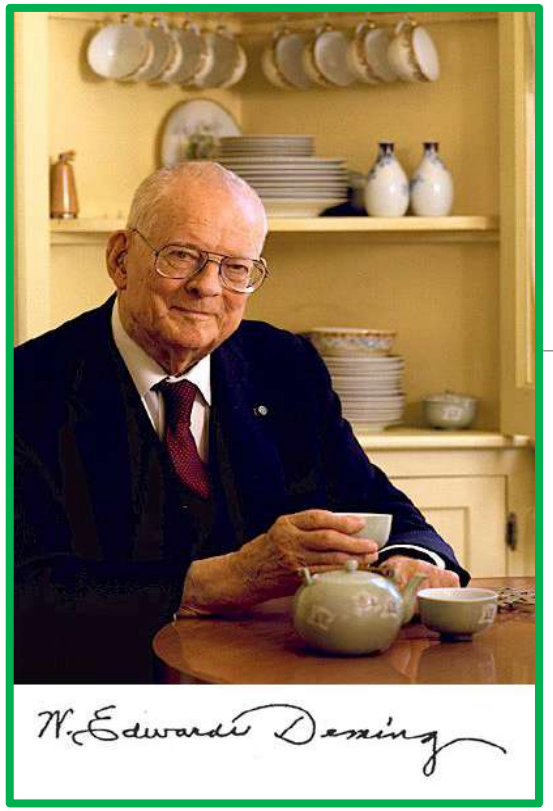
СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

ТЕЛ. (484)3974658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU



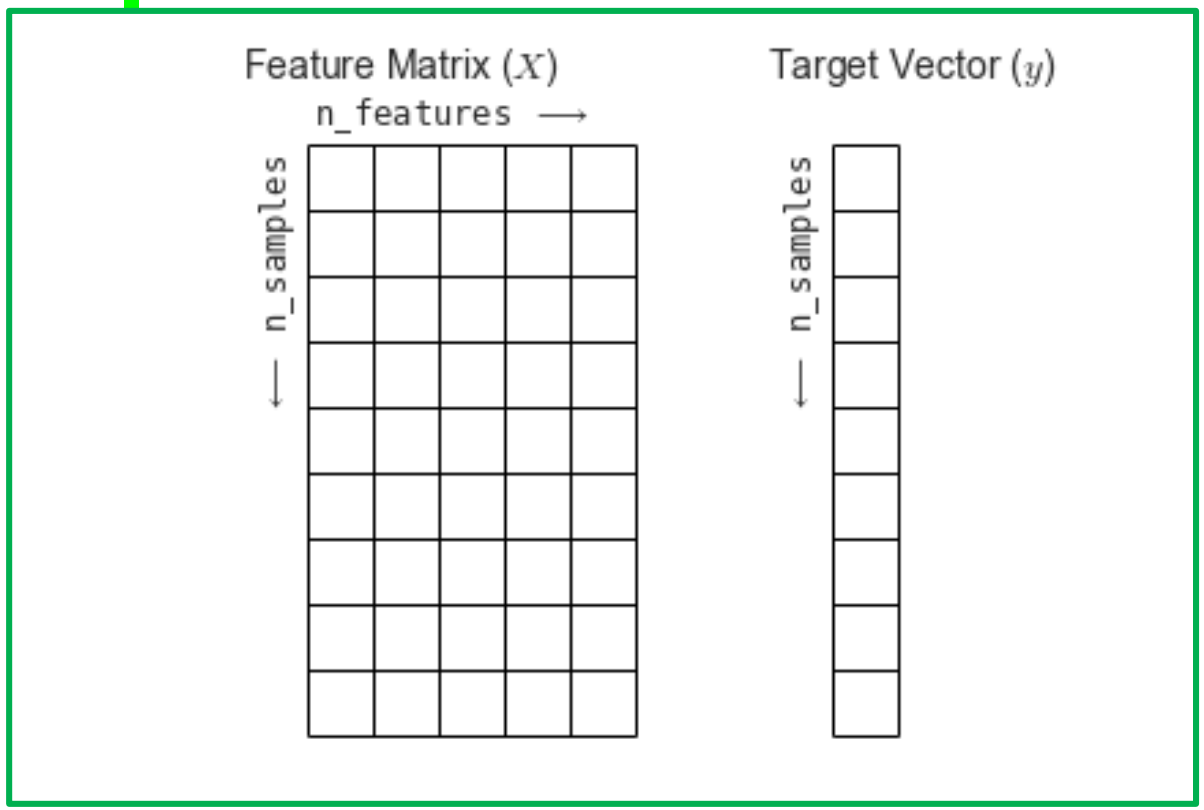
“In God we trust. All others must bring data”
W. Edwards Deming
“We are drowning in information and starving for knowledge”
–Rutherford D. Roger
“The purpose of computing is insight, not numbers”
-Richard W. Hamming

Information is the oil of the 21st century, and analytics is the combustion engine.
— Peter Sondergaard

I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
—Sir Arthur Conan Doyle, author of Sherlock Holmes stories

Регрессионный анализ – позиционирование
где его место?

- ***В любом варианте задача подразумевает наличие матрицы данных, содержащей признаки X – ее строки – объекты, ее столбцы- признаки***



Если вариант задачи – обучение с учителем, то подразумевается наличие вектора значений целевой переменной Y

Y – значения из конечного множества чисел - классификация
Y - произвольные значения из множества вещественных чисел - регрессия

Регрессионный анализ – *где его место?*

Решаем задачу обучения с учителем

Функционал качества

$$Q(a, X, Y) = \frac{1}{N} \sum_{i=1}^{N} L(a, x_i, y_i), x_i \in X, y_i \in Y$$

Принцип минимизации эмпирического риска:

$$a^* = \underset{A}{\operatorname{argmin}}\, Q(a, X_{train}, Y_{train}), A$$ – семейство алгоритмов.

Формула обучения:

Learning = Representation + Evaluation + Optimization

В нашей выборке есть Y!

Регрессионный анализ – *где его место?*

Функций $a(x)$ которые идеально описывают обучающую выборку бесконечно много. Нужно сузить функциональное пространство перебора. **Параметризуем** искомую функцию

модель $a(x, w)$ описывается вектором весов w.

Тогда задача превращается в поиск весов:

$$w^* = \arg\min_w Q(w, X_{train}, y_{train})$$

Примеры:

$$a(x, w) = x_1 w_1 + x_2 w_2$$

$$a(x, w) = x_1 x_2 x_3 w_1$$

$$a(x, w) = 1[x_1 < w_1]$$

В том числе, алгоритм, который просто запоминает обучающую выборку, воспроизведет ее ИДЕАЛЬНО, без ошибок! Но это не является обучением!

Надо сразу оговорить, в каком классе функций ищем решение
А затем в этом оговоренном классе искать конкретную функцию – конкретный $w=w^*$, обеспечивающий минимум

Когда получаем ЛИНЕЙНУЮ модель?

Пусть объект описывается D признаками $f_1, f_2, \ldots f_D$.
Тогда модель:

$a(x, w) = w_0 + \sum_{j=1}^{D} f_j w_j$ называется **линейной моделью**,

где w — D-мерный вектор признаков $w_1, w_2, \ldots w_D$

Далее будем считать, что в векторе признаков есть тождественно равный единице признак f_0, тогда формула упростится до:

Representation: $a(x, w) = \sum_{j=0}^{D} f_j w_j = \boldsymbol{x} \cdot \boldsymbol{w}$

Параметры модели интерпретируемы. w_i — значение, на которое изменится предсказание, если признак f_i увеличить на единицу.

Какая гипотеза лежит в основе линейной модели?

Введение в вектор признаков признака (столбца), тождественно равного единице, позволяет ввести в вид искомого решения свободный член - INTERCEPT

Когда получаем ЛИНЕЙНУЮ РЕГРЕССИОННУЮ модель?

Если целевая переменная Y - вещественное число, то такую модель называют линейной регрессией. Для нее функция потерь может иметь самые различные конкретные выражения:

$$Q(a, X, Y) = \frac{1}{N}\sum_{i=1}^{N} L(a, x_i, y_i), x_i \in X, y_i \in Y$$

Функции потерь:

1. **Квадратичная**$(Q - \text{MSE})$

$$L(a, x, y) = (a(x) - y)^2$$

2. **Абсолютная**$(Q - \text{MAE})$

$$L(a, x, y) = |a(x) - y|$$

3. **Логарифмическая** $(Q - \text{MSLE})$

$$L(a, x, y) = (\log(a(x) + 1) - \log(y + 1))^2$$

4. **Абсолютная-процентная** $(Q - \text{MAPE})$ $L(a, x, y) = \frac{|a(x) - y|}{y}$

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} L(a(x_i, w), y_i) = \underset{w}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{x_i} \cdot \boldsymbol{w}, y_i), x_i \in X, y_i \in Y$$

Иначе говоря, если имеем целевую функцию-вещественное число (левая часть равенства) ПЛЮС линейную модель (правая часть равенства), то имеем ЛИНЕЙНУЮ РЕГРЕССИОННУЮ модель!!!

Может рассматриваться огромное разнообразие функций потерь – здесь приведены лишь простейшие!

Мы должны в классе a найти оптимальное решение w^*

Получаем оптимизационную задачу – как ее решить?

Для квадратичной функции потерь – идем общим путем:

$$Q(X, \boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x_i}\cdot\boldsymbol{w} - y_i)^2 = \frac{1}{N}||X\cdot\boldsymbol{w} - \boldsymbol{y}||^2$$

X — матрица (N, D), w — вектор весов $(D,1)$, y — вектор ответов $(N,1)$

$X\cdot\boldsymbol{w}$ — вектор предсказаний $(N,1)$

Необходимое условие минимума:

$\nabla_{\boldsymbol{w}}Q(\boldsymbol{w})$ — градиент, вектор частных производных. Необходимое условие минимума — градиент равен нулю.

$$Q(\boldsymbol{w}) = \frac{1}{N}||X\cdot\boldsymbol{w} - \boldsymbol{y}||^2 = \frac{1}{N}(X\cdot\boldsymbol{w} - \boldsymbol{y})^T(X\cdot\boldsymbol{w} - \boldsymbol{y})$$

Линейная регрессия - невероятно популярный алгоритм машинного обучения.

Плюсы алгоритма:

- Быстро учится
- Быстро предсказывает
- Легко интерпретируется
- Легко хранить в памяти
- Легко применять с дифференцируемой функцией потерь

Весомый минус –

- Не способен учитывать нелинейные зависимости в данных.

Как можно улучшить модель?

$$a(x, w) = \sum_{j=0}^{D} f_j w_j = \boldsymbol{x} \cdot \boldsymbol{w}$$

ПОМНИМ:
Невозможно сделать правильное измененное признаковое пространство без понимания самой задачи!

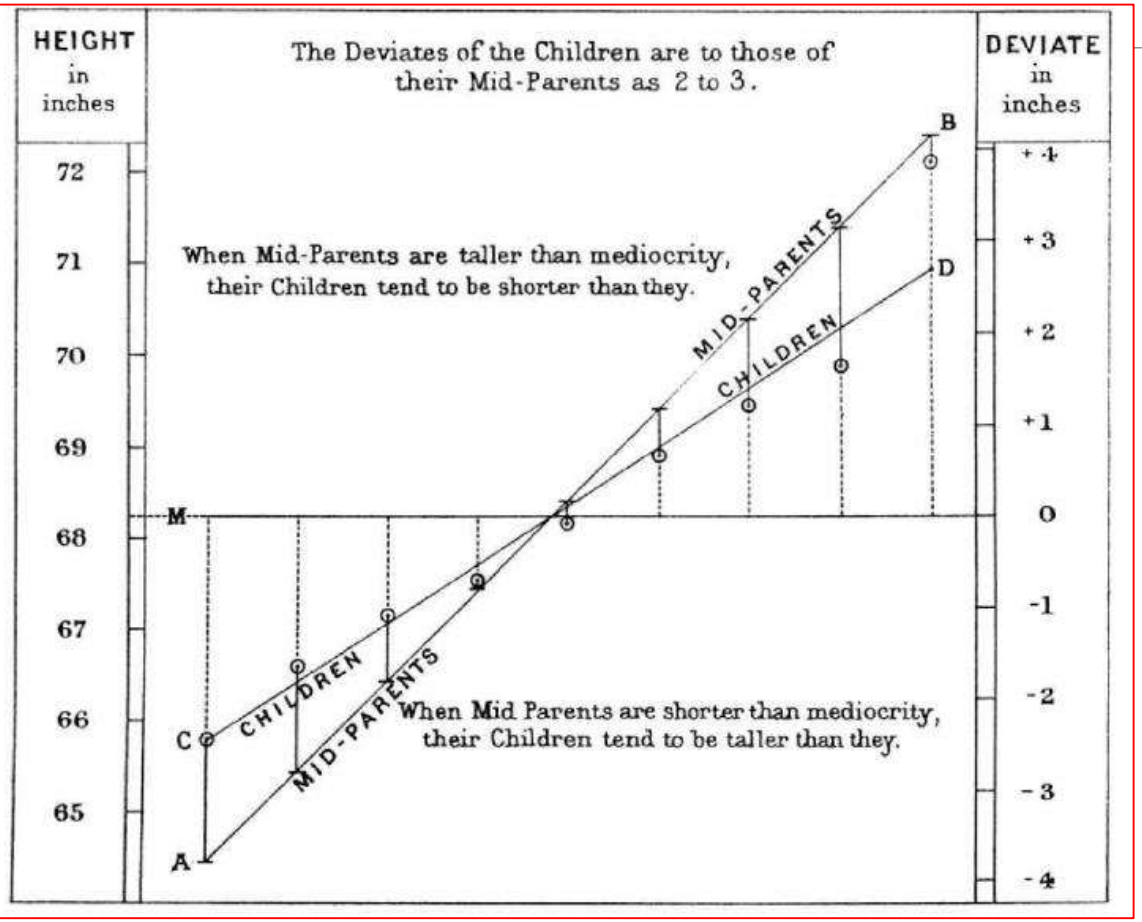
Параметры модели интерпретируемы. Это большой плюс линейной модели. W_i - значение, на которое изменится предсказание, если признак f_i увеличить на единицу.
Категориальные признаки кодируем:
• One-hot кодирование - категориальный признак с k значениями превращаем в k бинарных
признаков!
• Кодирование через целевую переменную (нельзя включать переменную самого объекта)
• Кодирование через вещественные признаки

Для вещественных признаков применяем нелинейные
функции - возводим в степень, берем синус и т.д.
Учитываем взаимодействия:
• Пару вещественных перемножаем, делим и т.д.
• Для пары бинарных используем логические операции
Изменяем масштаб признаков (вспомним ковариацию и корреляцию!)
Для исходных числовых признаков применяем стандартизацию и нормализацию:

- Стандартизация $f_j = \frac{f_j - mean(f_j)}{std(f_j)}$
- Min-max нормализация $f_j = \frac{f_j - min(f_j)}{max(f_j) - min(f_j)}$

Регрессионный анализ: откуда термин «регрессия»?

Впервые такая постановка - термин «регрессия» -появляется в работе Гальтона 1885 г. «Регрессия к середине в наследственности роста».



$$y - \bar{y} \approx \frac{2}{3}(x - \bar{x})$$

Отклонения для родителей

Отклонения для детей

Регрессионный анализ – традиционная постановка – решение МНК

Матричные обозначения:

Всего k+1 столбцов. Столбец для свободного члена (intercept)

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}; \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Метод наименьших квадратов (**не всегда применяется при постановке и формулировке решения задачи!**):

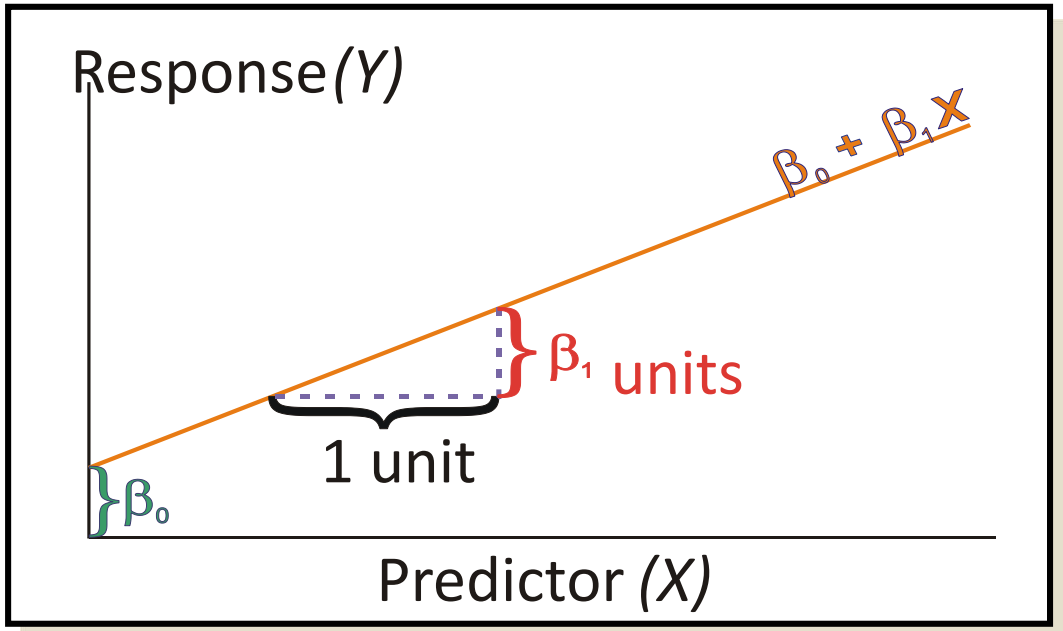
$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \beta_j x_{ij} \right)^2 \to \min_{\beta};$$

$$\|y - X\beta\|_2^2 \to \min_{\beta};$$

$$2X^T (y - X\beta) = 0,$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y,$$

$$\hat{y} = X \left(X^T X\right)^{-1} X^T y.$$



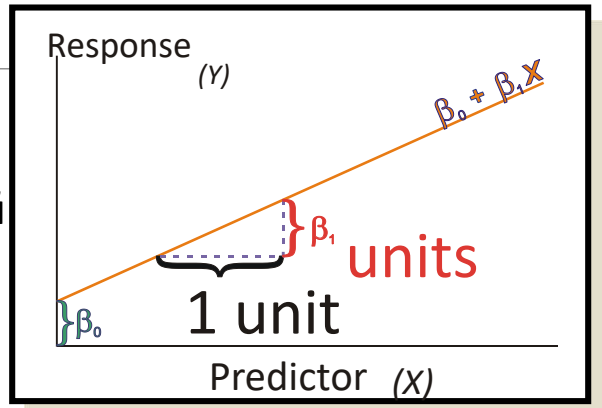
Регрессионный анализ – простейший случай: линейная модель, единственный предиктор

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

- Линейная регрессия – аппроксимируем зависимость линейной функцией

β_0 - Свободный член линейного выражения, INTERCEPT

β_1 - Коэффициент линейного выражения, тангенс угла наклона прямой



$$\mathrm{E}[\varepsilon_i] = 0; \quad \mathrm{V}[\varepsilon_i] = \sigma^2; \quad \mathrm{V}[\varepsilon_i \varepsilon_j] = 0, \quad \forall i \neq j.$$

- Случайная ошибка

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 .$$

Дифференцируем и приравниваем Производные =0

MIN

$$\begin{cases} \sum y_i = n b_0 + b_1 \sum x_i \\ \sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2 \end{cases}$$

Коэффициенты надо рассчитать так, чтобы они были в выбранном классе функций «наилучшими» в некотором смысле.
Наиболее распространены оценки Метода Наименьших Квадратов (МНК)
OLS –Ordinary Least Squares

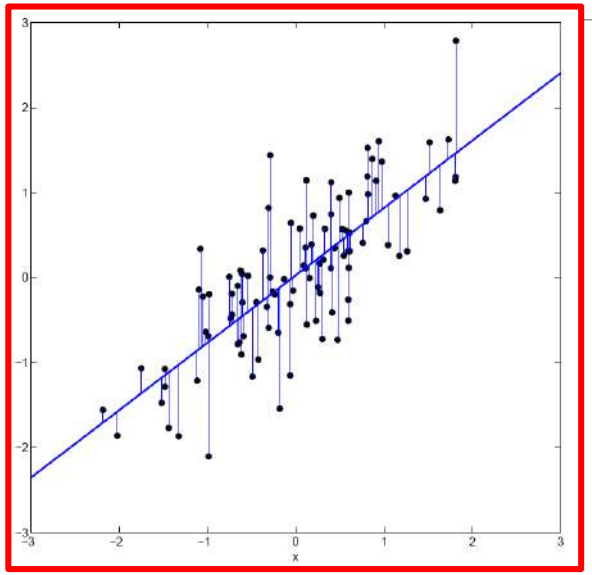
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x} .$$

Формулы расчета коэффициентов в простейшем случае одного предиктора

Регрессионный анализ – качество подгонки модели

Качество подгонки модели: Общая (total), объясненная моделью (explained) и остаточная (residual) суммы квадратов (sum of squares):



$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total Sum of Squares);}$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(Explained Sum of Squares);}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(Residual Sum of Squares);}$$

$$TSS = ESS + RSS.$$

Коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

$R^2 = r^2_{y\hat{}y}$ - **квадрат коэффициента множественной корреляции наблюденных и восстановленных (предсказанных моделью) значений y!!!!!!!!**

Если идеально восстановлены значения, R^2 равен 1

Регрессионный анализ: отбор предикторов, пошаговая регрессия (stepwise regression)

Стандартный коэффициент детерминации всегда увеличивается при добавлении регрессоров в модель, поэтому для отбора признаков и сравнения моделей с разным числом предикторов его использовать нельзя. Для сравнения моделей, содержащих разное число признаков, можно использовать приведённый коэффициент детерминации (adjusted RSQUARE):

$$R_a^2 = \frac{ESS/(n-k-1)}{TSS/(n-1)} = 1 - (1-R^2)\,\frac{n-1}{n-k-1}$$

Приведенный коэффициент детерминации даёт штраф за дополнительно включённые факторы, где *n* — количество наблюдений, а k — количество параметров.
Данный показатель всегда меньше единицы, но теоретически может быть и меньше нуля (только при очень маленьком значении обычного коэффициента детерминации и большом количестве факторов). Поэтому теряется интерпретация показателя как «доли» общей дисперсии, объясненной регрессионной моделью. Тем не менее, применение этого показателя в сравнении моделей, в том числе при переборе, вполне обоснованно.
Могут быть разные стратегии перебора в пошаговой регрессии: с последовательным включением, с последовательным исключением, и т.д.

Регрессионный анализ – простейший случай: линейная модель (реализация в среде SAS)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	688.66931	688.66931	254.100	0.0005
Error	3	8.13069	2.71023		
C Total	4	696.80000			

Root MSE	1.64628	R-square	0.9883
Dep Mean	58.20000	Adj R-sq	0.9844
C.V.	2.82866		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.700990	3.49727376	1.058	0.3676
HUMERUS	1	0.825743	0.05180152	15.941	0.0005

Регрессионный анализ – простейший случай: линейная модель

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	688.66931	688.66931	254.100	0.0005
Error	3	8.13069	2.71023		
C Total	4	696.80000			

Root MSE	1.64628	R-square	0.9883
Dep Mean	58.20000	Adj R-sq	0.9844
C.V.	2.82866		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.700990	3.49727376	1.058	0.3676
HUMERUS	1	0.825743	0.05180152	15.941	0.0005

Регрессионный анализ – предположения линейной модели

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \ldots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке $(\operatorname{rank} X = k + 1)$.
4. Случайность ошибок: $\mathbb{E}(\varepsilon | X) = 0$.
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}(\varepsilon | X) = \sigma^2$.

В предположениях (1-4) МНК-оценки коэффициентов β являются несмещёнными:

$$\mathbb{E}\hat{\beta}_j = \beta_j,\ j = 0, \ldots, k,$$

и состоятельными:

$$\forall \gamma > 0 \ \lim_{n \to \infty} P\left(\left|\beta_j - \hat{\beta}_j\right| < \gamma\right) = 1,\ j = 0, \ldots, k.$$

Регрессионный анализ – предположения линейной модели

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \ldots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке $(\operatorname{rank} X = k + 1)$.
4. Случайность ошибок: $\mathbb{E}(\varepsilon | X) = 0$.
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}(\varepsilon | X) = \sigma^2$.

(предположения Гаусса-Маркова).

Теорема Гаусса-Маркова: в предположениях (1-5) МНК-оценки имеют наименьшую дисперсию в классе оценок β, линейных по y.

Регрессионный анализ: дисперсии МНК оценок

В предположениях (1-5) дисперсии МНК-оценок коэффициентов β задаются следующим образом:

$$\mathbb{D}\left(\hat{\beta}_j \middle| X\right) = \frac{\sigma^2}{TSS_j\left(1 - R_j^2\right)},$$

где $TSS_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$, R_j^2 — коэффициент детерминации при регрессии x_j на все остальные признаки из X.

- Чем больше дисперсия ошибки σ^2, тем больше дисперсия оценки $\hat{\beta}_j$.
- Чем больше вариация значений признака x_j в выборке, тем меньше дисперсия оценки $\hat{\beta}_j$.
- Чем лучше признак x_j объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия оценки $\hat{\beta}_j$.

Регрессионный анализ: дисперсии МНК оценок

В предположениях (1-5) дисперсии МНК-оценок коэффициентов β задаются следующим образом:

$$\mathbb{D}\left(\hat{\beta}_j \middle| X\right) = \frac{\sigma^2}{TSS_j\left(1 - R_j^2\right)},$$

где $TSS_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$, R_j^2 — коэффициент детерминации при регрессии x_j на все остальные признаки из X.

- Чем больше дисперсия ошибки σ^2, тем больше дисперсия оценки $\hat{\beta}_j$.
- Чем больше вариация значений признака x_j в выборке, тем меньше дисперсия оценки $\hat{\beta}_j$.
- Чем лучше признак x_j объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия оценки $\hat{\beta}_j$.

Регрессионный анализ: дисперсии МНК оценок

В матричном виде:

$$\mathbb{D}\left(\hat{\beta}\middle| X\right) = \sigma^2 \left(X^T X\right)^{-1}.$$

Если столбцы X почти линейно зависимы, то матрица $X^T X$ плохо обусловлена, и дисперсия оценок $\hat{\beta}_j$ велика.

Близкая к линейной зависимость между двумя или более признаками x_j называется мультиколлинеарностью.
Мультиколлинеарность – одна из основных «бед» линейного регрессионного анализа на основе МНК (но не единственная)!
Проблема мультиколлинеарности решается с помощью отбора признаков или использования регуляризаторов.
Отбор признаков – задача, решаемая средствами пошаговой регрессии

Регрессионный анализ: проверка предположений Гаусса-Маркова

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \ldots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке $(\operatorname{rank} X = k + 1)$.
4. Случайность ошибок: $\mathbb{E}(\varepsilon | X) = 0$.
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}(\varepsilon | X) = \sigma^2$.

Предположения (1-2) проверить нельзя.
Предположение (3) легко проверяется, без его выполнения построить модель вообще невозможно.
Если оно не выполняется, программные средства не строят регрессионные модели.
Предположения (4-6) об ошибке ε необходимо проверять.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i,\ i = 1, \ldots, n.$$

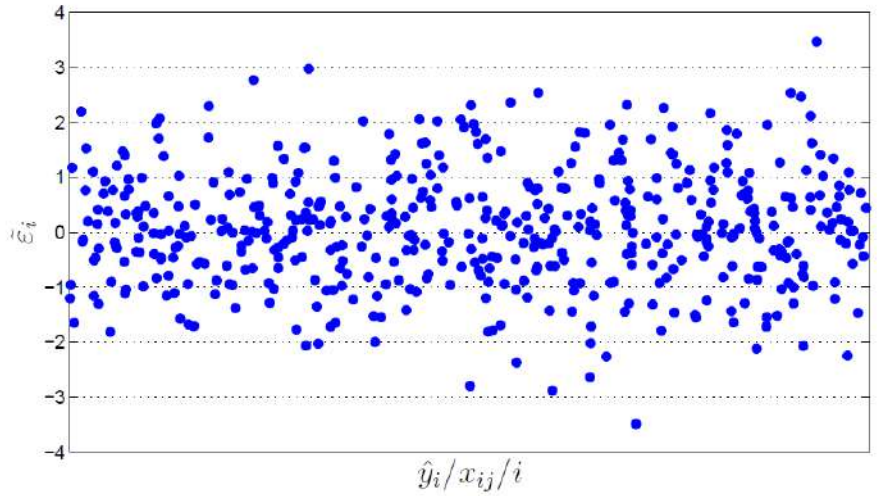
Остатки (residuals)

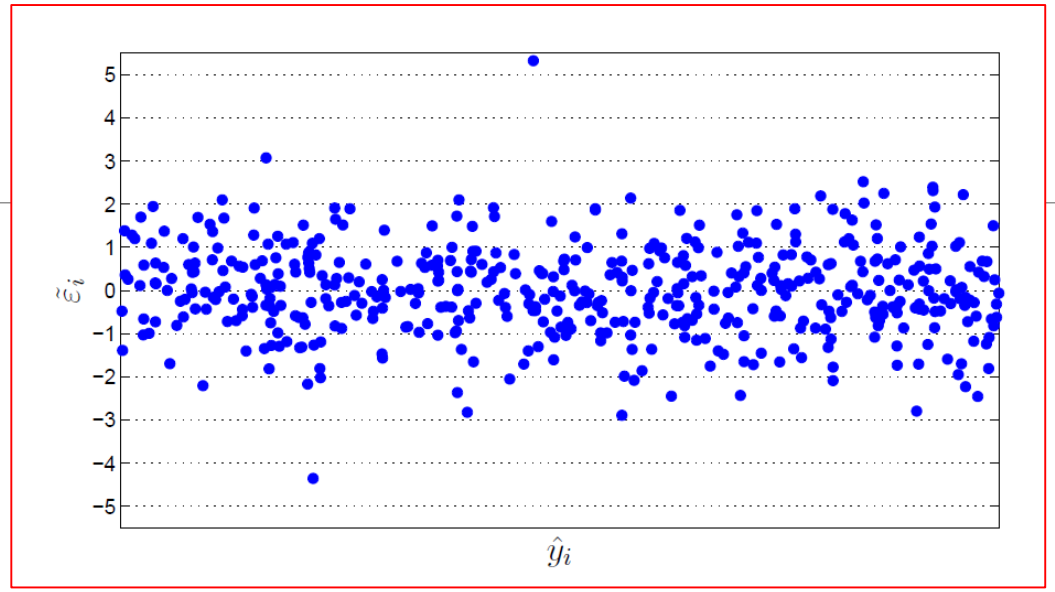
$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}},\ i = 1, \ldots, n.$$

Стандартизованные остатки

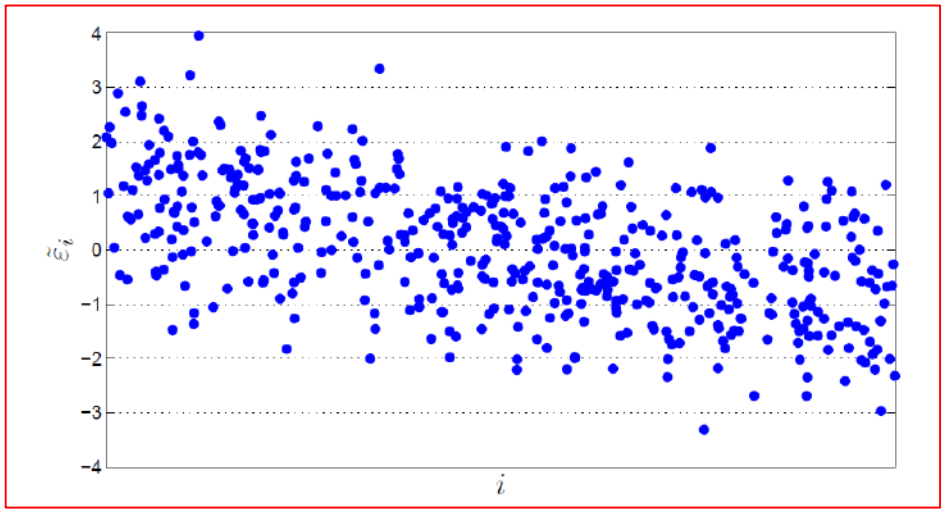
Регрессионный анализ: анализ остатков

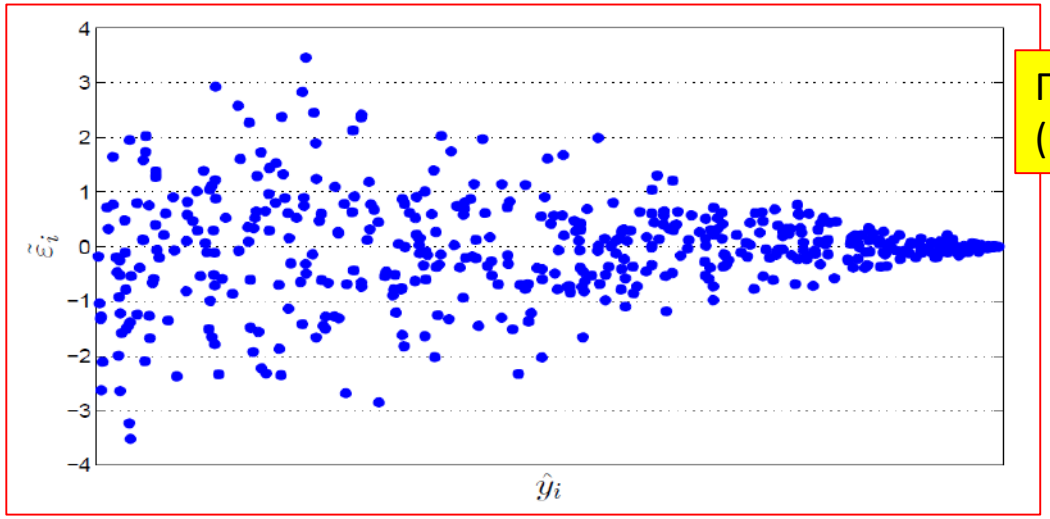
Строятся графики зависимости $\tilde{\varepsilon}_i$ от $\hat{y}_i, \quad x_{ij}, j = 1, \ldots, k, \quad i.$





Есть выбросы (outliers)!

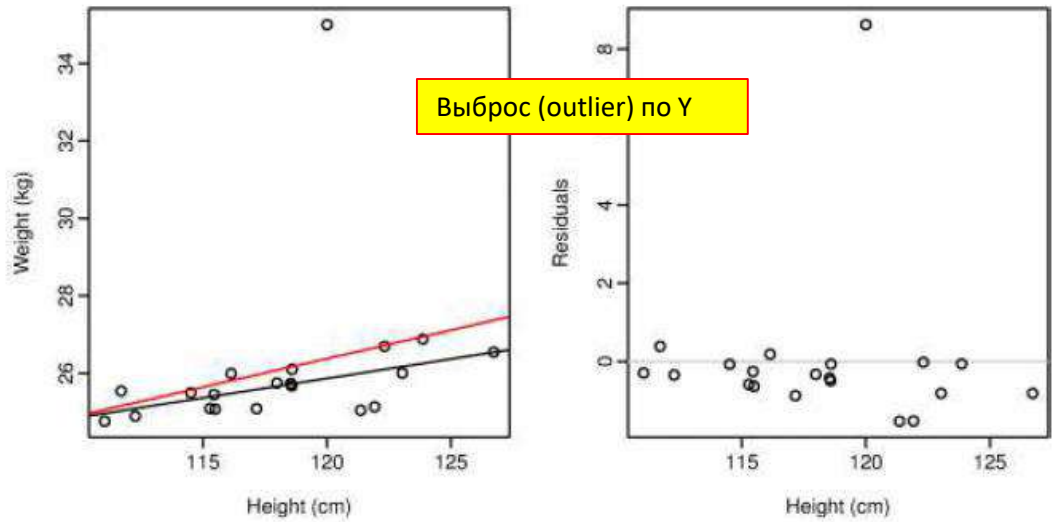




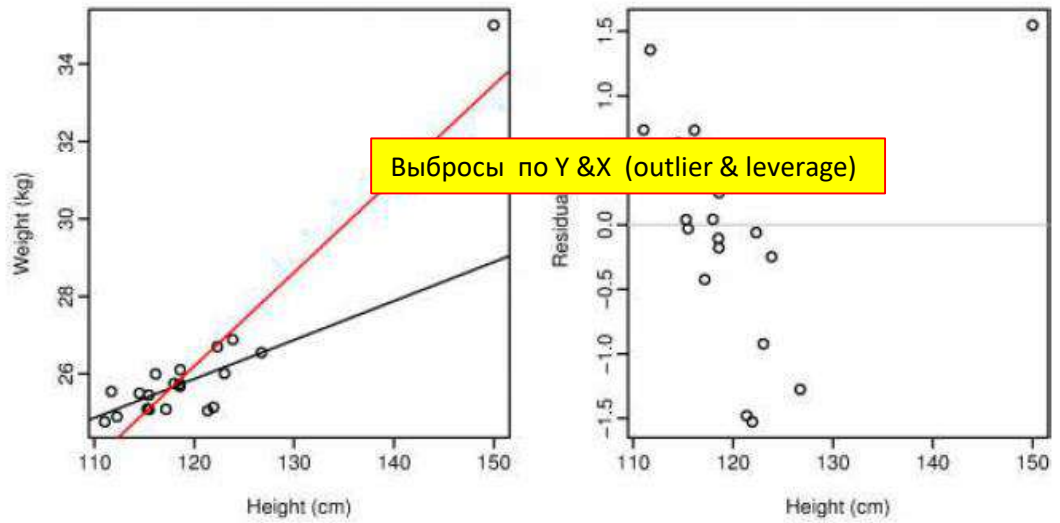
Гетероскедастичность (heteroscedasticity)!

Есть тренд в данных!

Регрессионный анализ: анализ остатков



Выброс (outlier) по Y



Выбросы по Y &X (outlier & leverage)

Расстояние Кука — мера воздействия i-го наблюдения на регрессионное уравнение:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{RSS\,(k+1)} = \frac{\hat{\varepsilon}_i^2}{RSS\,(k+1)} \frac{h_i}{(1-h_i)^2},$$

$\hat{y}_{j(i)}$ — предсказания модели, настроенной по наблюдениям $1, \ldots, i-1, i+1, \ldots, n$, для наблюдения j;
h_i — диагональный элемент матрицы $H = X\left(X^T X\right)^{-1} X^T$ (hat matrix).

Гетероскедастичность может быть следствием недоопределения модели.

Последствия гетероскедастичности:

- нарушаются предположения критериев Стьюдента и Фишера и методов построения доверительных интервалов для σ и β (независимо от объёма выборки);
- МНК-оценки β и R^2 остаются несмещёнными и состоятельными.

Варианты:

- переопределить модель, добавить признаки, преобразовать отклик;
- использовать модифицированные оценки дисперсии коэффициентов для оценки значимости;
- настроить параметры методом взвешенных наименьших квадратов.

Регрессионный анализ: другие функции потерь и робастная регрессия

Очень много условий, когда возможно корректное применение МНК оценок (нарушения предположений о линейности, гомоскедастичность, случайность выборки – некоррелированные наблюдения, случайность ошибок, отсутствие мультиколлинеарности, отсутствие выбросов как по предикторам, так и по предиктанту…)

В реальной жизни эти условия часто не выполняются!

Робастность в статистике – построение грубых, устойчивых, нечувствительных оценок

В вычислительных схемах:

для одномерных характеристик – методы, основанные на квантилях

В анализе связей между переменными – методы, основанные на ранговых оценках корреляции

В регрессионном анализе -?

$$y_i = x_i^T \beta + \varepsilon_i \rightarrow \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

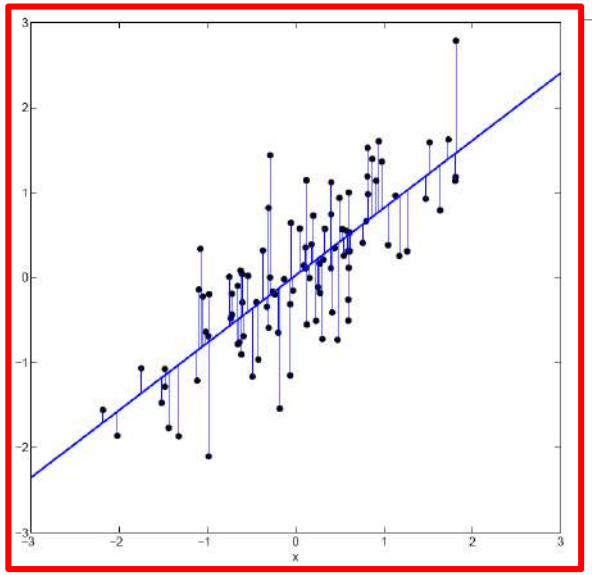
Функция потерь квадратична – «жестокое наказание» за большую невязку – Обычная линейная МНК регрессия

Надо вводить другие функции потерь!

Смягчать потери от больших выбросов – тогда оценки будут более устойчивыми, менее чувствительными к выбросам!

Регрессионный анализ – качество подгонки модели

Качество подгонки модели: Общая (total), объясненная моделью (explained) и остаточная (residual) суммы квадратов (sum of squares):



$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total Sum of Squares);}$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(Explained Sum of Squares);}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(Residual Sum of Squares);}$$

$$TSS = ESS + RSS.$$

Коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

$R^2 = r^2_{y\hat{}y}$ - **квадрат коэффициента множественной корреляции наблюденных и восстановленных (предсказанных моделью) значений y!!!!!!!!**

Если идеально восстановлены значения, R^2 равен 1

Регрессионный анализ: робастная регрессия

Есть множество альтернатив квадратичной функции потерь $\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$.

М-оценки Хубера:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T\beta}{\hat{\sigma}}\right)$$

ρ – робастная функция потерь,
если ρ(t)=1/2 t^2 - то это просто МНК функция потерь и МНК оценка

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^T\beta|$$

Минимум абсолютных отклонений – MAD – относительно хорошо справляется с выбросами по Y, но плохо – с выбросами по X (leverage)

$$\hat{\beta} = \arg\min_{\beta} \mathrm{Med}\{(y_i - x_i^T\beta)^2\}$$

Минимум медианы квадратичных отклонений – LMS – относительно хорошо справляется со значительным числом выбросов по Y, часто используется как начальное приближение для расчетов других робастных статистик

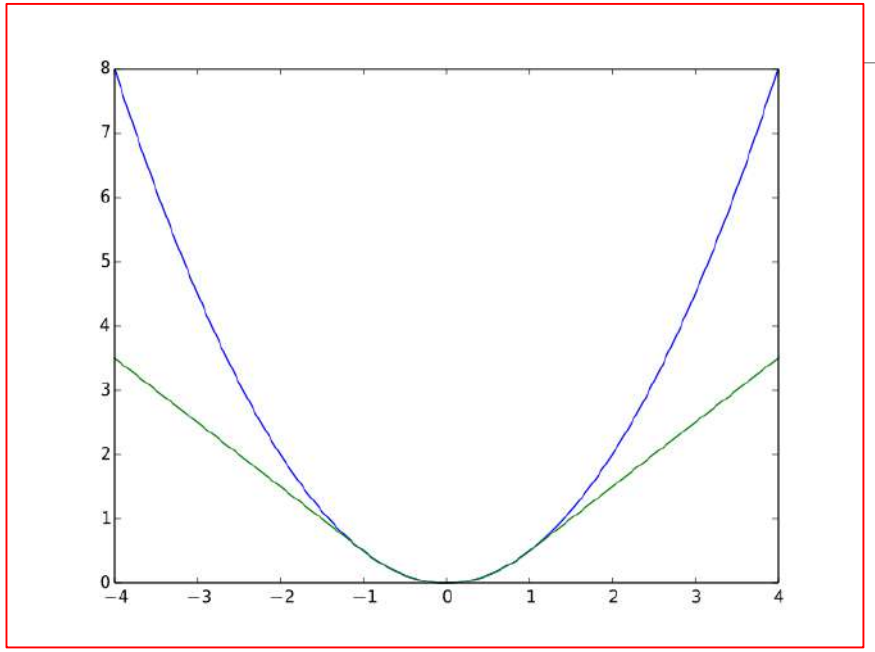
$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{q} r_{(i)}(\beta)^2$$

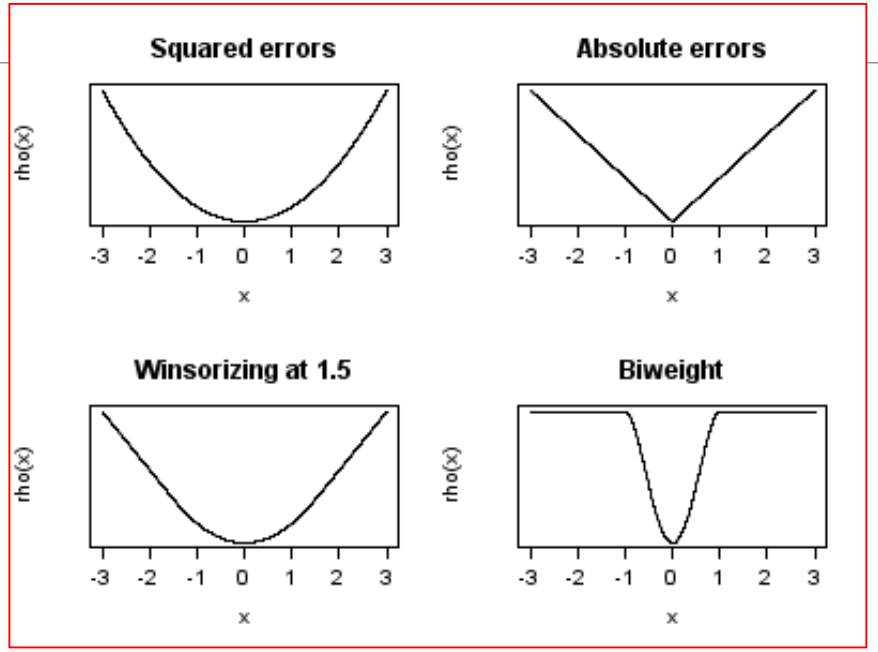
$$r_{(i)}(\beta) = y_{(i)} - x_{(i)}^T\beta,\ r_{(1)}(\beta)^2 \le \cdots \le r_{(q)}(\beta)^2$$

$$q = \lfloor n(1-\alpha) + 1 \rfloor$$

Урезанные наименьшие квадраты - LTS

Регрессионный анализ: робастная регрессия – виды функций потерь





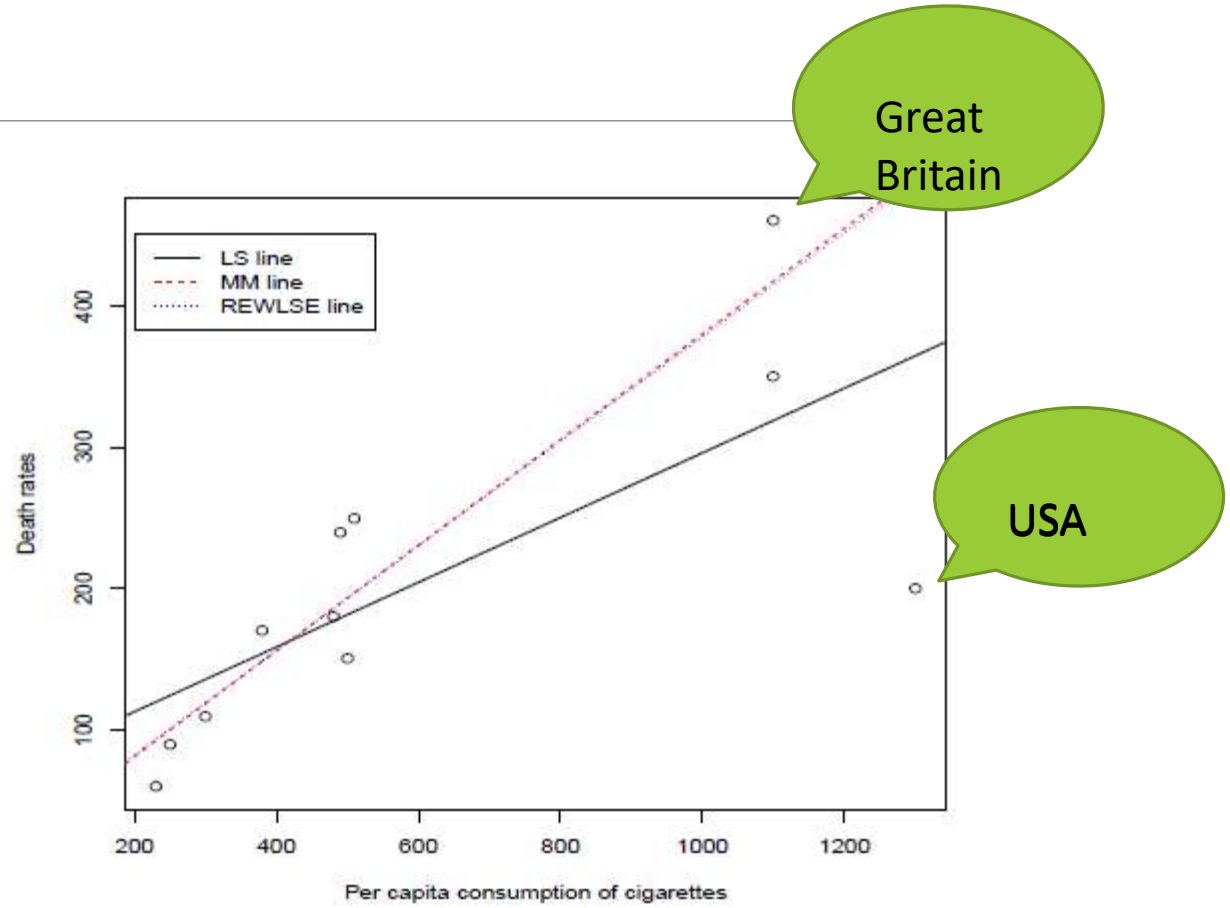
Регрессионный анализ: робастная регрессия –пример

связь между потреблением сигарет в 1930-е на душу населения, и смертностью на миллион населения от рака легких в 1950-е годы

Country	Per capita consumption of cigarette	Deaths rates
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1100	350
GreatBritain	1100	460
Iceland	230	060
Netherlands	490	240
Norway	250	090
Sweden	300	110
Switzerland	510	250
USA	1300	200

	Complete data		Data without USA	
Estimators	Intercept	Slope	Intercept	Slope
LS	67.5609	0.2284	9.1393	0.3687
MM	7.0639	0.3729	5.9414	0.3753
REWLSE	9.1393	0.3686	9.1393	0.3686

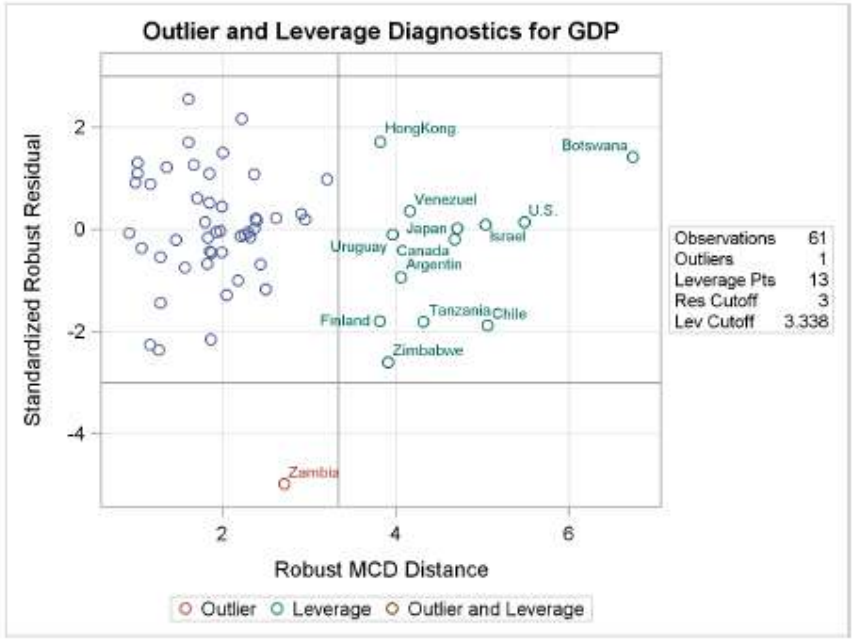
Обычная регрессия – МНК и два робастных метода: MM & REWLSE

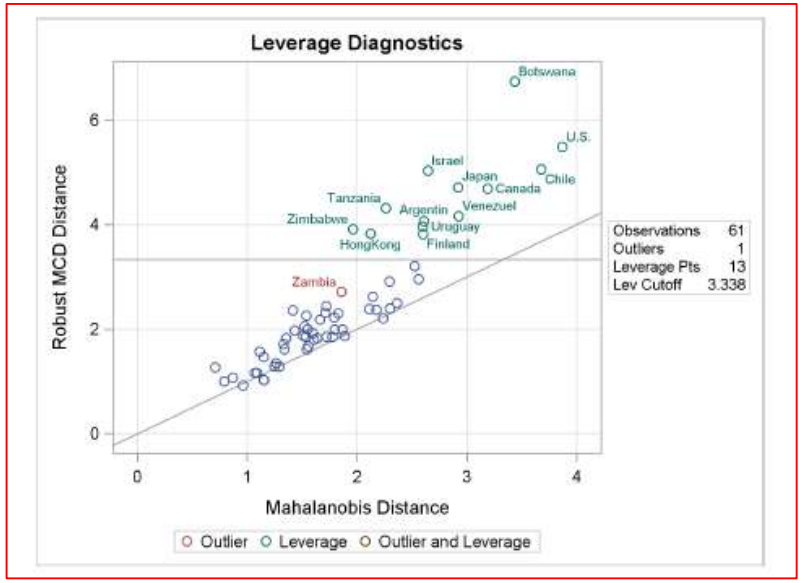


Регрессионный анализ: робастная регрессия

зависимость ВВП стран от ряда макроэкономических параметров

$$GDP = \beta_0 + \beta_1 LFG + \beta_2 GAP + \beta_3 EQP + \beta_4 NEQ + \epsilon$$





Diagnostics

Obs	country	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	Argentin	2.6083	4.0639	*	-0.9424	
5	Botswana	3.4351	6.7391	*	1.4200	
8	Canada	3.1876	4.6843	*	-0.1972	
9	Chile	3.6752	5.0599	*	-1.8784	
17	Finland	2.6024	3.8186	*	-1.7971	
23	HongKong	2.1225	3.8238	*	1.7161	
27	Israel	2.6461	5.0336	*	0.0909	
31	Japan	2.9179	4.7140	*	0.0216	
53	Tanzania	2.2600	4.3193	*	-1.8082	
57	U.S.	3.8701	5.4874	*	0.1448	
58	Uruguay	2.5953	3.9671	*	-0.0978	
59	Venezuel	2.9239	4.1663	*	0.3573	
60	Zambia	1.8562	2.7135		-4.9798	*
61	Zimbabwe	1.9634	3.9128	*	-2.5959	

GDP - Рост ВВП на одного работающего
LFG – Рост рабочей силы, Трудоспособного населения
GAP -Разрыв ВВП— разница между фактическим ВВП и потенциальным ВВП.
EQP – Инвестирование в оборудование
NEQ – Инвестирование в не-оборудование (человеческий фактор, R&D)

Квантильная регрессия

определение квантилей

Для данной случайной переменной Y, $\tau^{\ddot{y}}$ квантиль $(\text{для } 0 \le \tau \le 1)\ Q(\tau)$ это наименьшее значение y_τ такое, что доля значений Y меньших, или равных y_τ равна τ

Формально, если функция $F(y) = \Pr(Y \le y)$ распределения переменной, обратная функция:

$$Q(\tau) = \inf\left(\{y / F(y) \ge \tau\}\right)$$ определяет квантиль

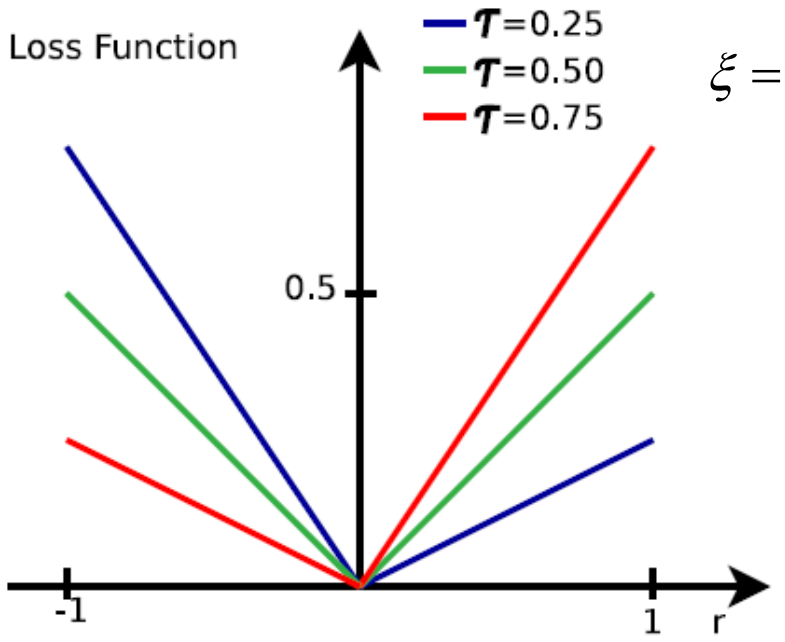
$y_{0.5}$ Называют медианой

$Q(\tau)$ можно легко получить, пересортировав значения переменной по возрастанию и пересчитывая их слева направо

По сравнению с единичной статистикой $E(Y)$ – средним значением Y, набор квантилей дает гораздо большую информацию о распределении Y

Квантильная регрессия – несимметричная функция потерь

Вычисление квантилей – оптимизационная задача



$$\xi = \arg\min_{\xi \in R} \left[\sum_{i:y_i \geq \xi} \tau \left| y_i - \xi \right| + \sum_{i:y_i < \xi} (1-\tau) \left| y_i - \xi \right| \right]$$

В частном случае для медианы – метод минимальных модулей:

$$\min_{\xi \in R} \sum_{i=1}^{n} |y_i - \xi|$$

Первая сумма – сумма штрафов за те значения элементов выборки, которые превосходят выбранное в качестве решения оптимизационной задачи значение.

Вторая сумма – сумма штрафов за те значения элементов выборки, которые меньше выбранного в качестве решения оптимизационной задачи значения.

Веса штрафов по каждому значению в общем случае не равны: с той стороны, где значений **должно быть больше** – веса меньше, и наоборот, с той стороны, где значений **должно быть меньше** – веса штрафов больше.

В частном случае для квантиля=0.5, веса одинаковы, и приходим к методу минимальных модулей, решение для него дает медиану.

Квантильная регрессия

От вычисления среднего – к МНК регрессии

$$\hat{\mu} = \text{argmin}_{\mu \in R} \sum_{i=1}^{n} (y_i - \mu)^2$$

Условное среднее для линейной регрессии:

$$E(Y|X = x) = x'\beta$$

$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

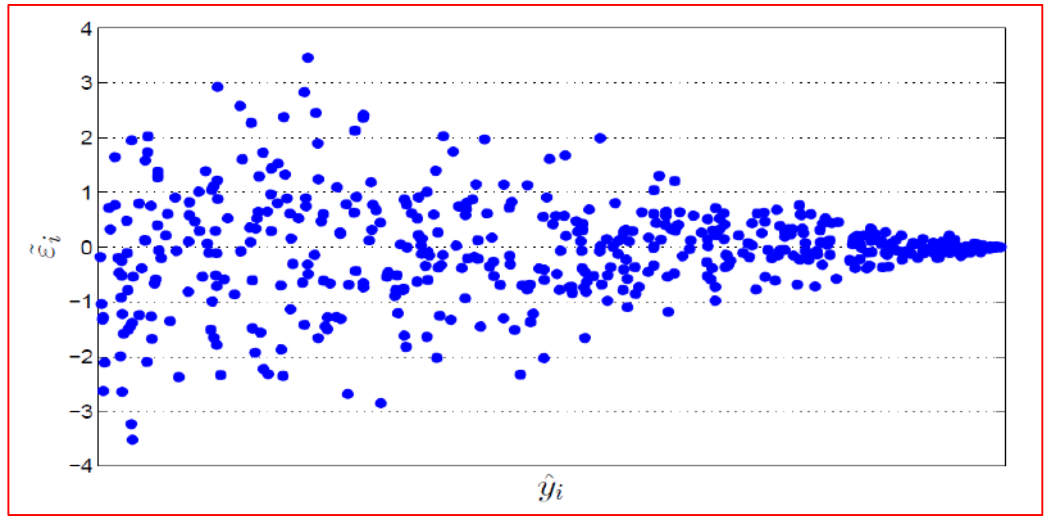
Нахождение решения для МНК регрессии – оптимизационная задача, решается аналитически

Особенности МНК регрессии

Используется информация о поведении только условного среднего, не используется информация о поведении всего распределения величины

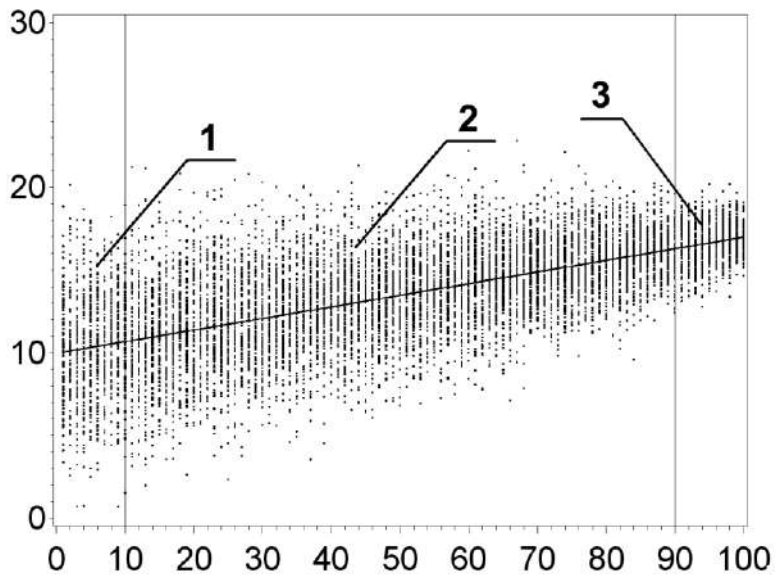
Для корректного расчета МНК регрессии необходимо, чтобы распределение величины не изменялось во времени и сами значения были независимы (Гомоскедастичность)

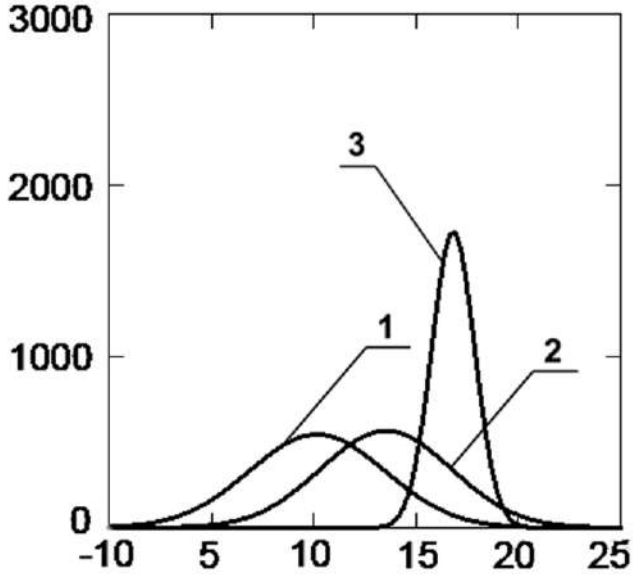
Чувствительность к выбросам



Квантильная регрессия

Особенности МНК регрессии





Традиционная МНК регрессия в средних не даёт никакой информации об изменении распределения.

Квантильная регрессия

В 1978 году была опубликована статья Роджера Коенкера (Koenker) и Гильберта Бассета-младшего (Bassett) “Квантильная регрессия” (REGRESSION QUANTILES).

Предложенный метод квантильной регрессии относится к робастным методам, так как является устойчивым к отклонениям от предположений классических моделей.

От нахождения квантилей – к квантильной регрессии:

Если ввести аргумент X и допустить, что $y=x'\beta+\varepsilon$

Квантиль $Q(\tau)$является решением оптимизационной задачи:

$$\xi = \arg\min_{\xi\in R}\left[\sum_{i:y_i\geq\xi}\tau\left|y_i-\xi\right|+\sum_{i:y_i<\xi}(1-\tau)\left|y_i-\xi\right|\right]$$

По аналогии с МНК регрессией:

$$E\left(Q(\tau)/X=x\right)=x'\beta_\tau.$$

β_τ -решение оптимизационной задачи:

$$\beta = \arg\min_{\beta\in R}\left[\sum_{i:y_i\geq x_i\beta}\tau\left|y_i-x_i\beta\right|+\sum_{i:y_i<x_i\beta}(1-\tau)\left|y_i-x_i\beta\right|\right]$$

Приведенная выше оптимизационная задача не может быть решена аналитически, но может быть решена как задача выпуклого линейного программирования, в том числе симплекс-методом

Квантильная регрессия и обычная МНК регрессия

Квантили:

$$F(y) = \text{Prob}\,(Y \le y)$$

$$Q(\tau) = \inf\{y : F(y) \ge \tau\}$$

$$0 < \tau < 1$$

$$\tilde{\xi} = \arg\min_{\xi \in R}\Big[\sum_{i:y_i \ge \xi} \tau\,|y_i - \xi| + \sum_{i:y_i < \xi} (1-\tau)\,|y_i - \xi|\Big]$$

$$\min_{\xi \in R} \sum_{i=1}^{n} |y_i - \xi|$$

Для медианы – метод минимальных модулей

МНК регрессия:

Среднее:

$$\hat{\mu} = \text{argmin}_{\mu \in R} \sum_{i=1}^{n} (y_i - \mu)^2$$

$$E(Y|X = x) = x'\beta$$

$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

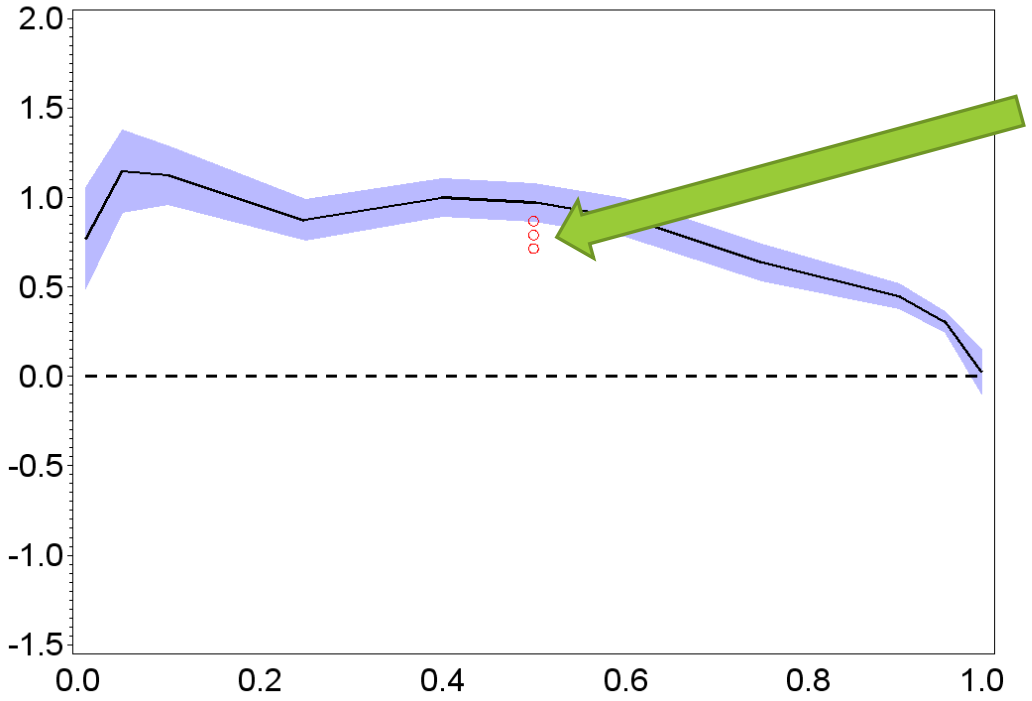
Может быть решена аналитически

Квантильная регрессия:

$$\min_{\beta \in R^p}\Big[\sum_{i \in \{i:y_i \ge x_i'\beta\}} \tau\,|y_i - x_i'\beta| + \sum_{i \in \{i:y_i < x_i'\beta\}} (1-\tau)|y_i - x_i'\beta|\Big]$$

Может быть решена как задача выпуклого линейного программирования, в том числе симплекс методом

Процесс - диаграммы для квантильной регрессии – климатические тренды



Что дает обычная МНК регрессия

Пример процесс-диаграммы, построенной по реальным данным (среднесуточная приземная температура для зимы (станция 34172)

Множественная регрессия – линейная модель

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i = \sum_{k=0}^{p-1} \beta_k x_{ik} + \varepsilon_i ,$$

Возможны матричные обозначения:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{*}\mathbf{y}$$

Если хотим построить полиномиальную модель – вводим новые переменные, являющиеся степенями исходных переменных

Спасибо за внимание!

Лекция -окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

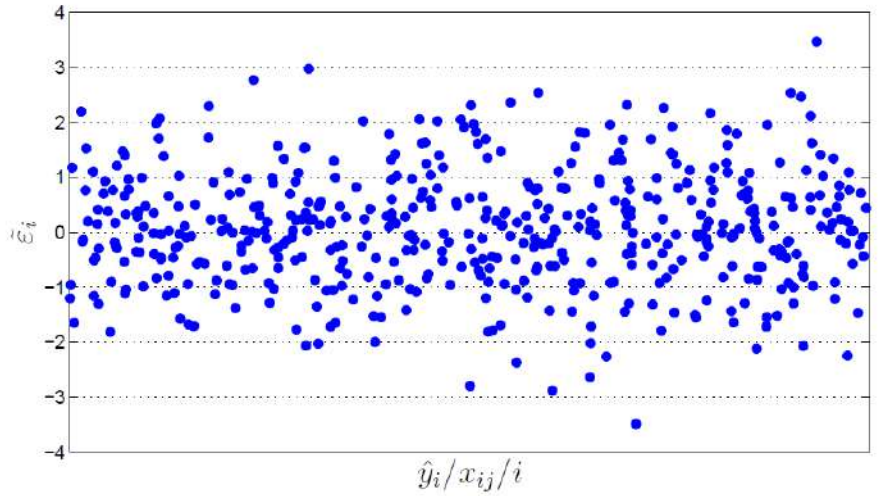
ТЕЛ. (484)3974658,

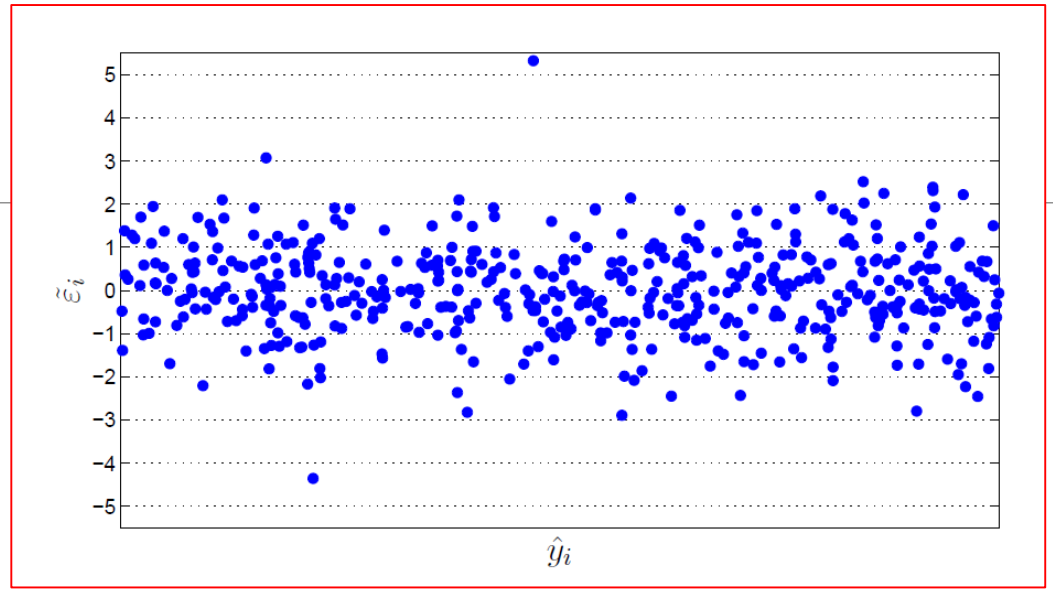
ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

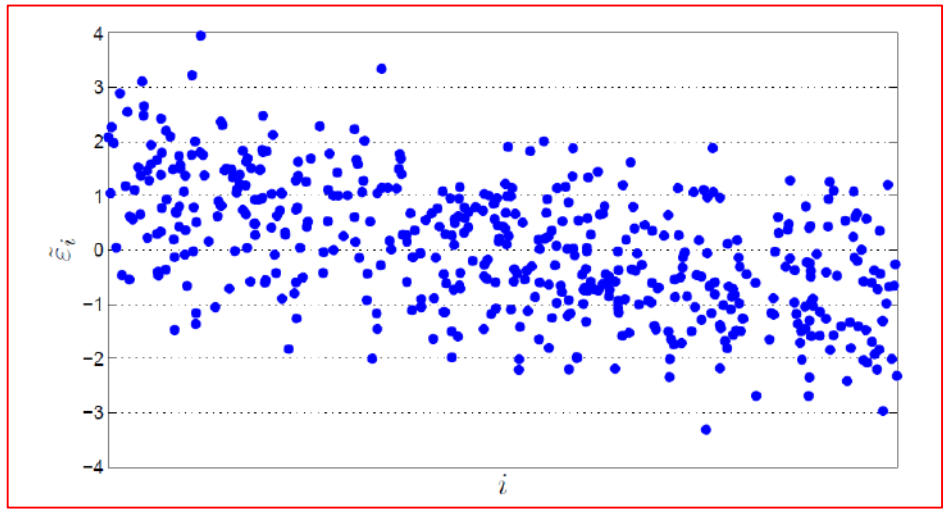
Регрессионный анализ - анализ остатков
Нужно посмотреть, как ведут себя остатки - графики!

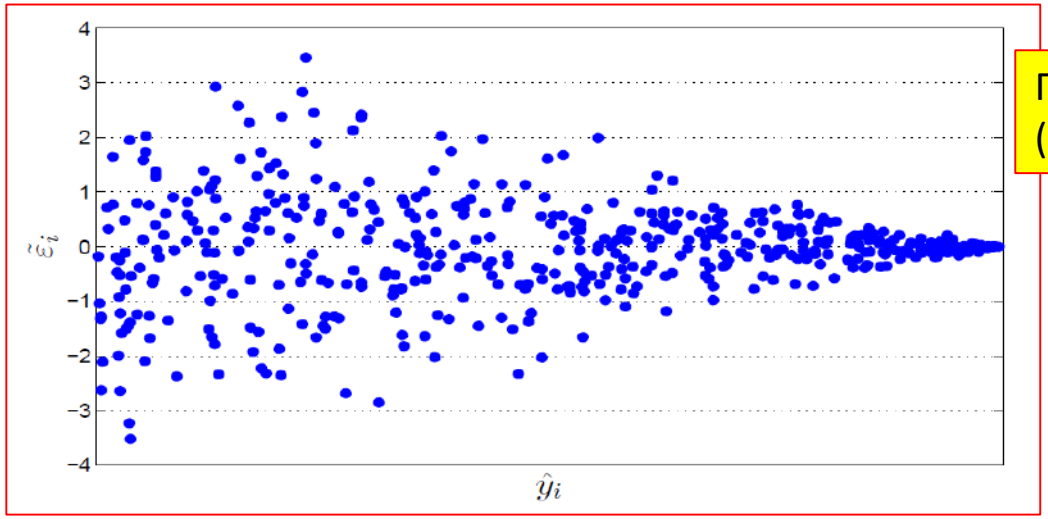
Строятся графики зависимости $\tilde{\varepsilon}_i$ от $\hat{y}_i, \quad x_{ij}, j = 1, \ldots, k, \quad i.$





Есть выбросы (outliers)!

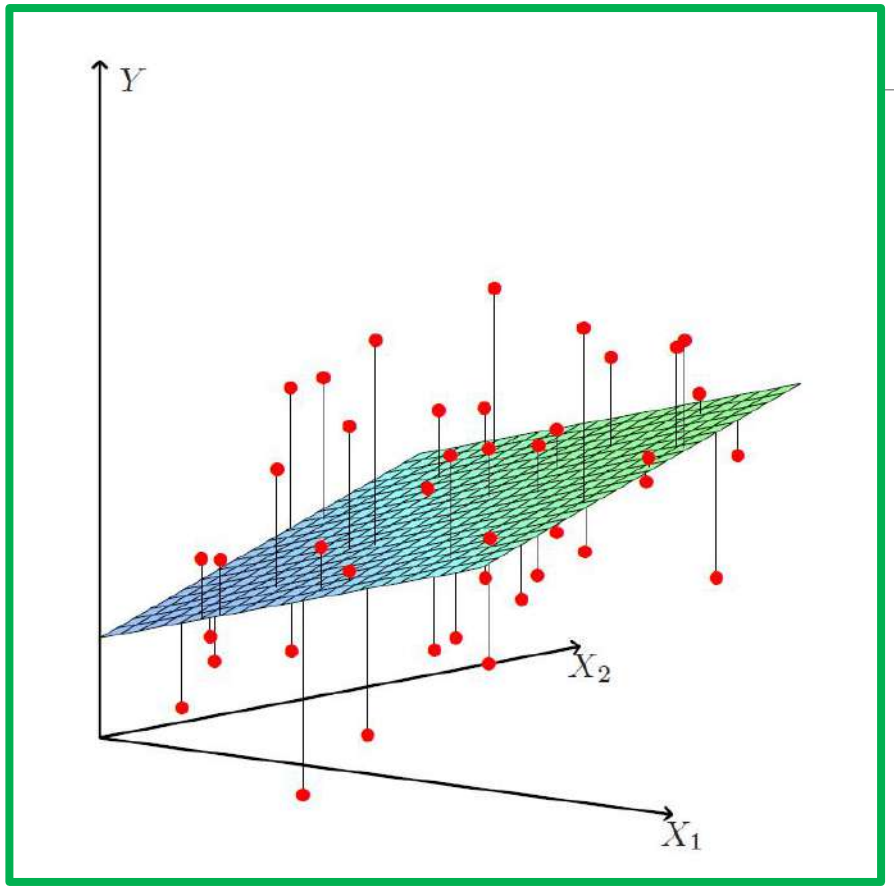


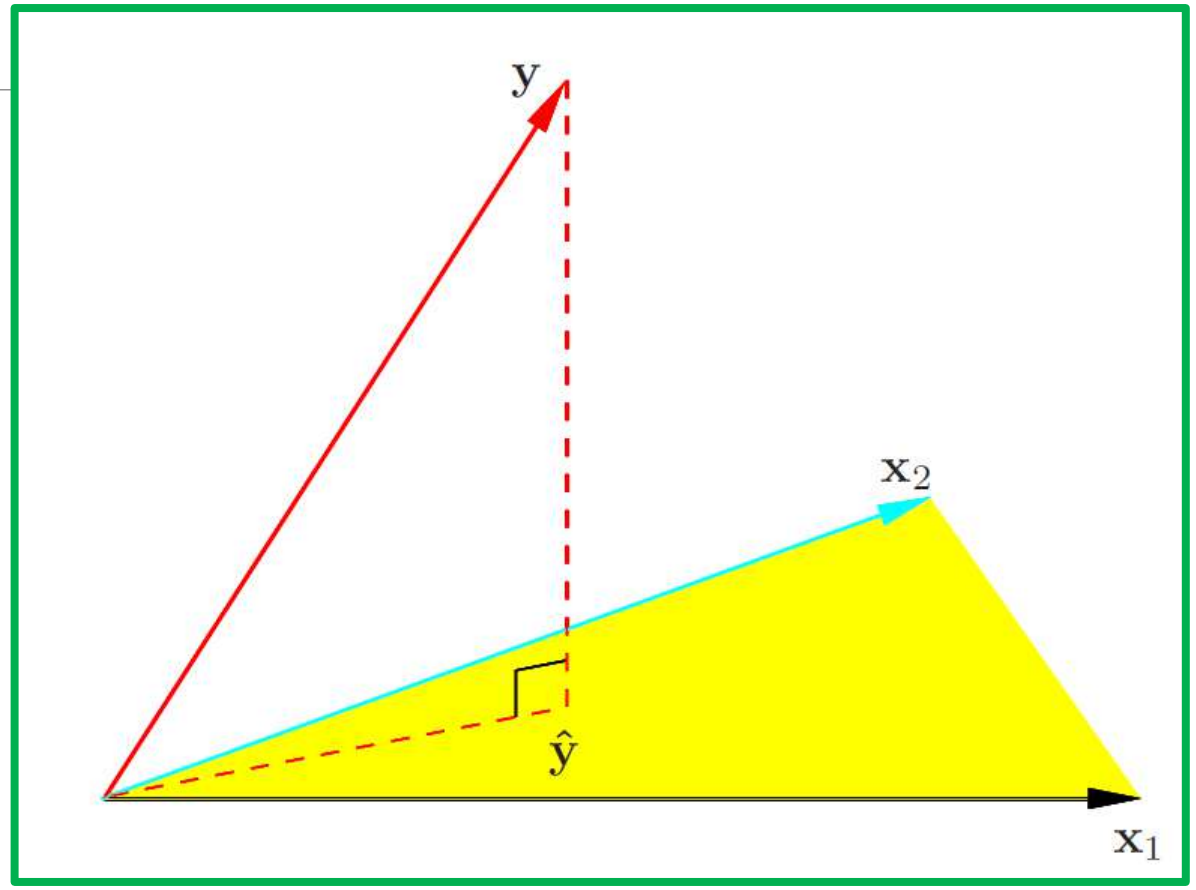


Гетероскедастичность (heteroscedasticity)!

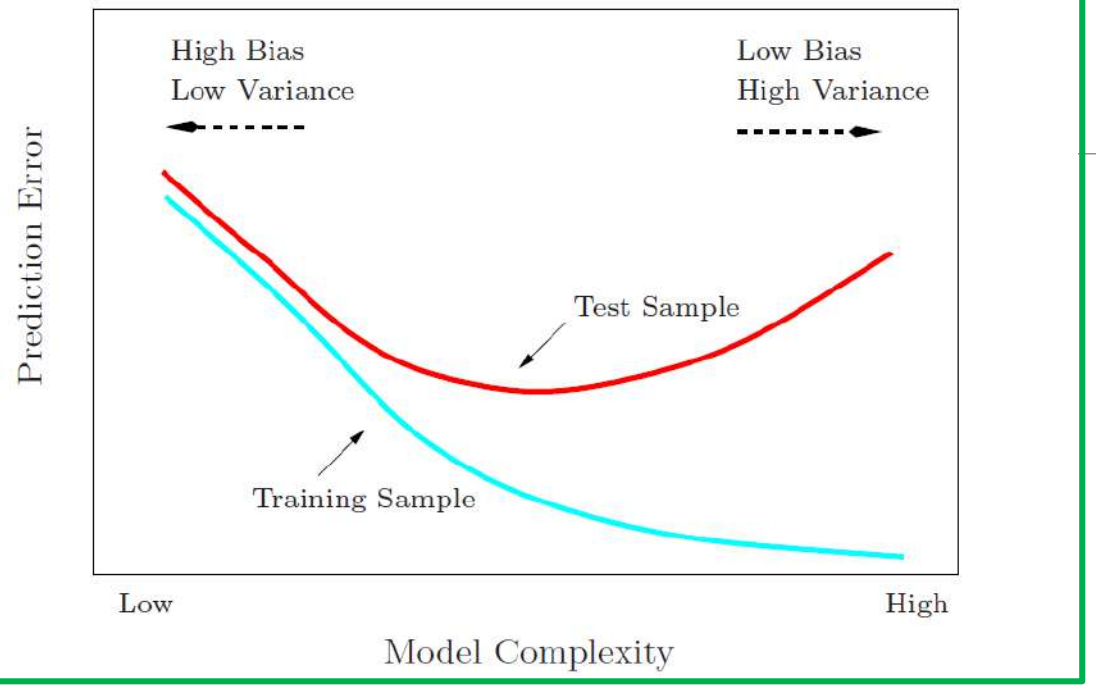
Есть тренд в данных!

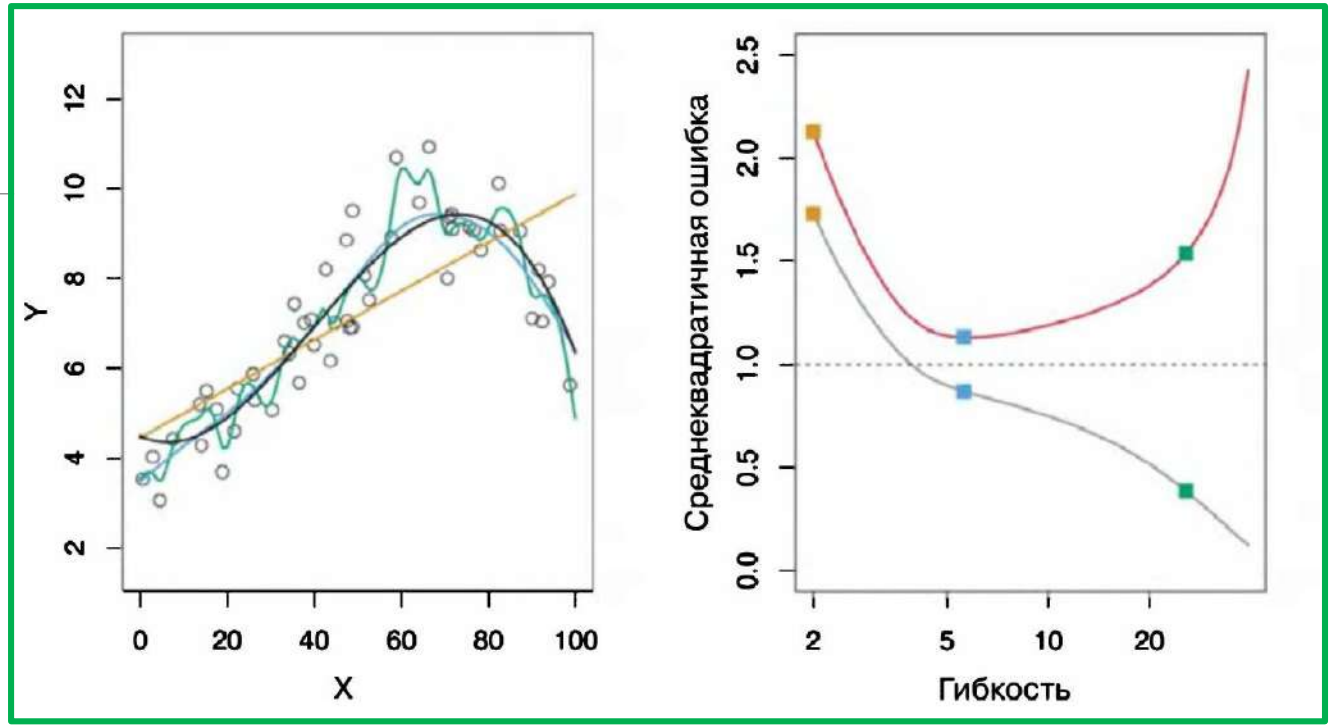
Регрессионный анализ – линейная регрессия с несколькими предикторами





Регрессионный анализ (и не только!) – что надо учитывать при построении моделей





По мере возрастания гибкости метода статистического обучения происходит монотонное снижение MSE на обучающих данных и U-образное изменение MSE на контрольных данных. Это является фундаментальным свойством статистического обучения, которое остается справедливым для любого имеющегося набора данных и для любого применяемого статистического метода.

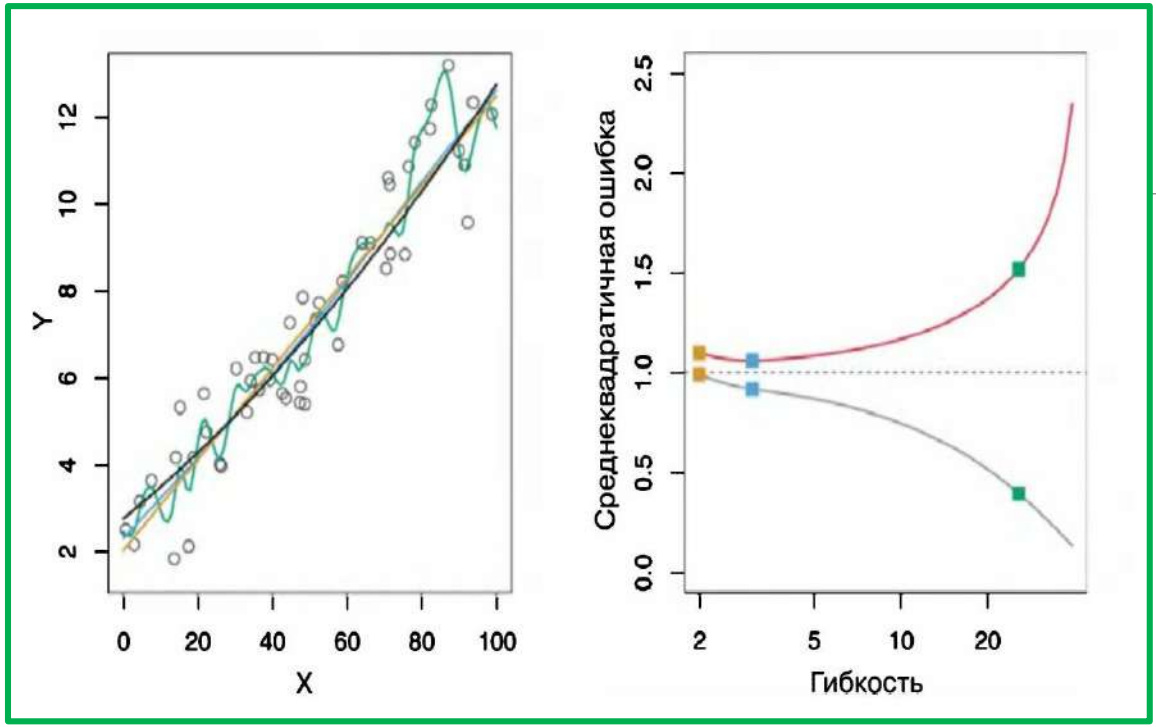
Переобучение – overfitting

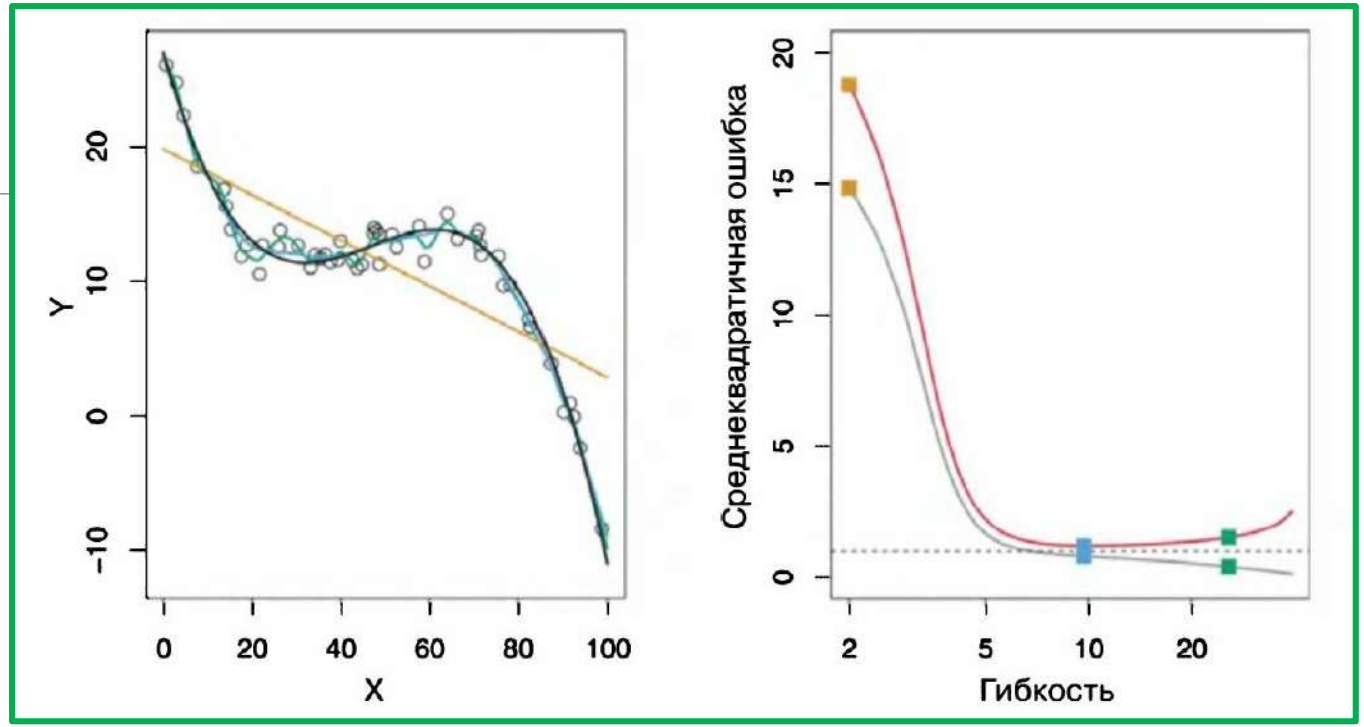
Оранжевая – линейная модель, синяя и зеленая – две подгонки сплайнами

На обучающих данных сложная модель находит «закономерности», которые могут оказаться просто результатом случайного воздействия. На контрольных (тестовых) данных она допускает большую ошибку, т.к. найденные «закономерности» в контрольных данных просто не существуют!

Случай 1.

Регрессионный анализ (и не только!) – что надо учитывать при построении моделей





Случай 2. Линейная модель подобрана удачно! Сложный сплайн – допускает переобучение!

Случай 3. Линейная модель подобрана неудачно…. Оба вида сплайнов подобраны удачно и хорошо описывают данные

$$\begin{aligned}\mathrm{EPE}_k(x_0) &= \mathrm{E}[(Y-\hat{f}_k(x_0))^2|X=x_0] \\ &= \sigma^2+[\mathrm{Bias}^2(\hat{f}_k(x_0))+\mathrm{Var}_{\mathcal{T}}(\hat{f}_k(x_0))] \\ &= \sigma^2+\left[f(x_0)-\frac{1}{k}\sum_{\ell=1}^{k}f(x_{(\ell)})\right]^2+\frac{\sigma^2}{k}.\end{aligned}$$

Противоречие между ограниченным объемом данных и сложностью модели: нужен компромисс между смещением и дисперсией!!

Регрессионный анализ (и не только!) – что надо учитывать при выборе и построении моделей:
Компромисс между смещением и дисперсией (Model Selection and the Bias–Variance Tradeoff)

$$\begin{aligned}\mathrm{EPE}_k(x_0) &= \mathrm{E}[(Y-\hat{f}_k(x_0))^2|X=x_0] \\ &= \sigma^2+[\mathrm{Bias}^2(\hat{f}_k(x_0))+\mathrm{Var}_{\mathcal{T}}(\hat{f}_k(x_0))] \\ &= \sigma^2+\left[f(x_0)-\frac{1}{k}\sum_{\ell=1}^{k}f(x_{(\ell)})\right]^2+\frac{\sigma^2}{k}.\end{aligned}$$

Левая часть представляет собой математическое ожидание ошибки на контрольной выборке и представляет собой среднее значение MSE, которое мы получили бы при многократном повторном оценивании f на основе большого числа обучающих выборок и вычислении ошибки для каждого контрольного значения x_0.

Более простая форма:

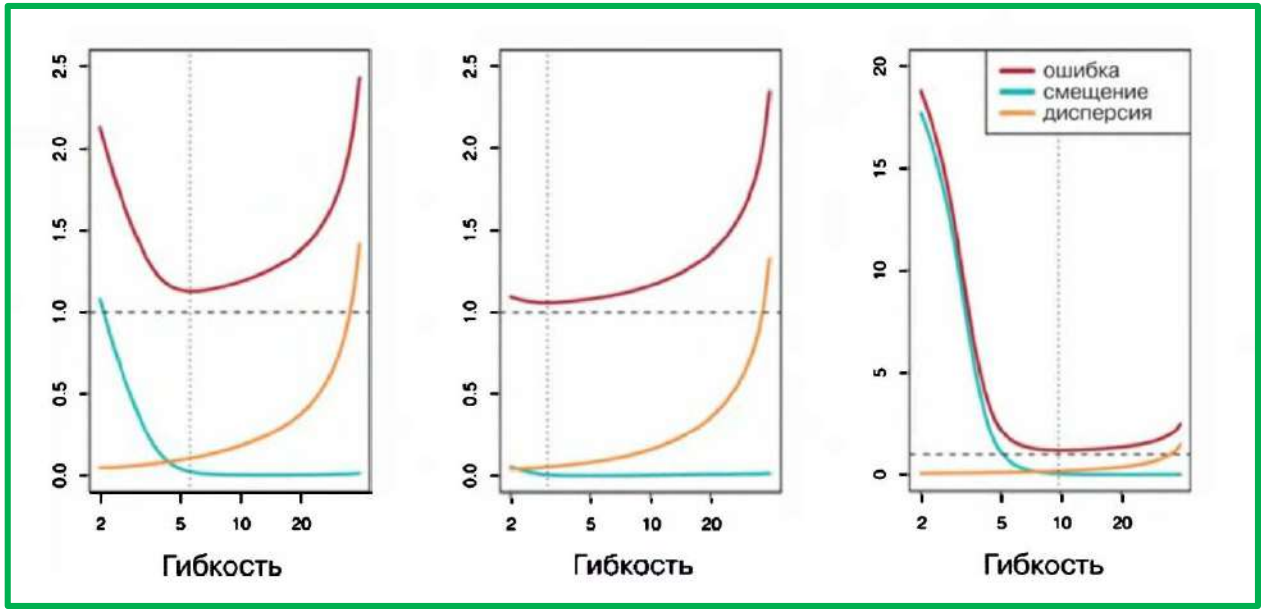
$$E\left(y_0-\hat{f}(x_0)\right)^2=\mathrm{Var}(\hat{f}(x_0))+[\mathrm{Bias}(\hat{f}(x_0))]^2+Var(\epsilon)$$

Дисперсия означает величину, на которую f изменилась бы при оценивании этой функции с использованием другой обучающей выборки.

Смещение означает ошибку, вводимую за счет аппроксимирования проблемы из реального мира, которая может оказаться чрезвычайно сложной, при помощи гораздо более простого метода. Например, линейная регрессия предполагает, что между Y и X_i, X_2...., Xp имеется линейная зависимость. Маловероятно, что какая-либо проблема из реального мира действительно описывается такой простой зависимостью, в связи с чем применение линейной регрессии несомненно приведет к определенному смещению оценки /.

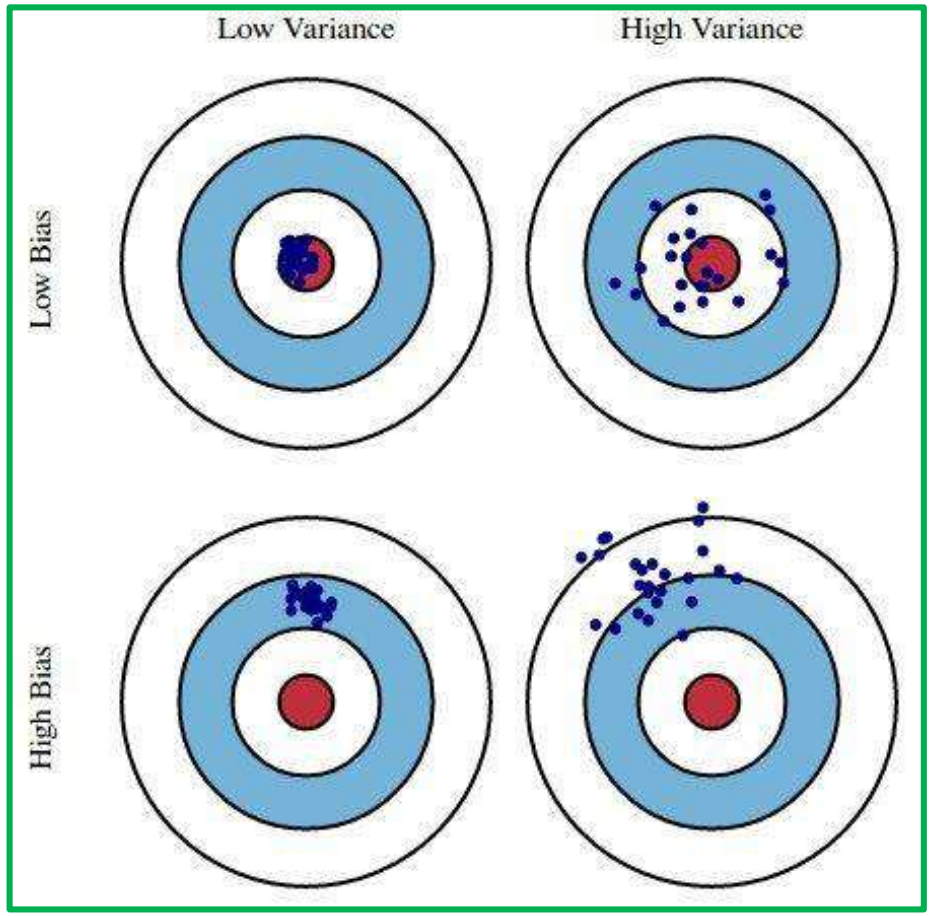
Как правило, при использовании более гибких методов дисперсия будет возрастать, а смещение — снижаться

Случай 1 | Случай 2 | Случай 3



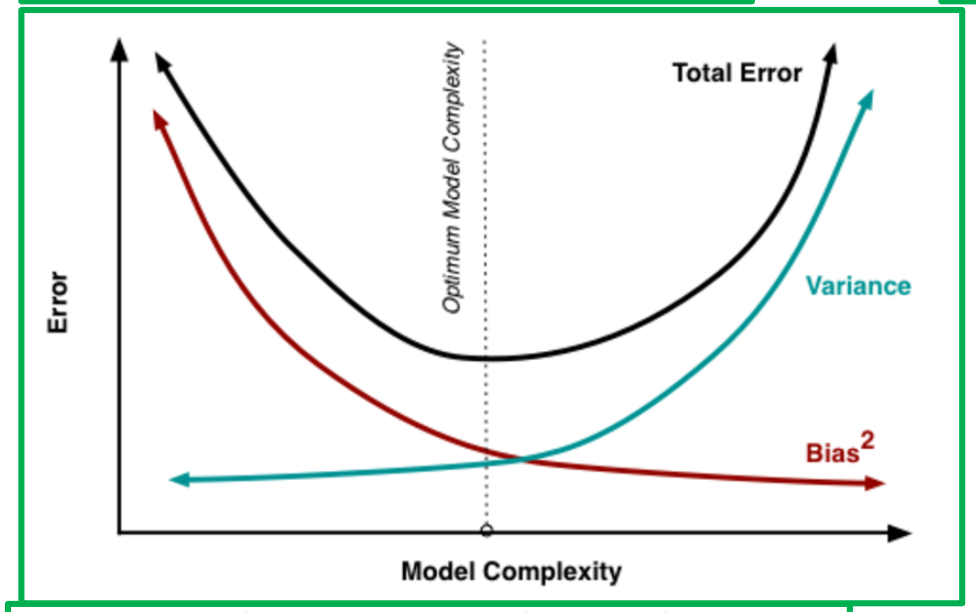
Компромисс между дисперсией и смещением (variance-bias trade-off) при использовании МНК оценок:

$$Bias(\hat{\beta}_{OLS}) = E(\hat{\beta}_{OLS}) - \beta.$$



$$Var(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1},$$

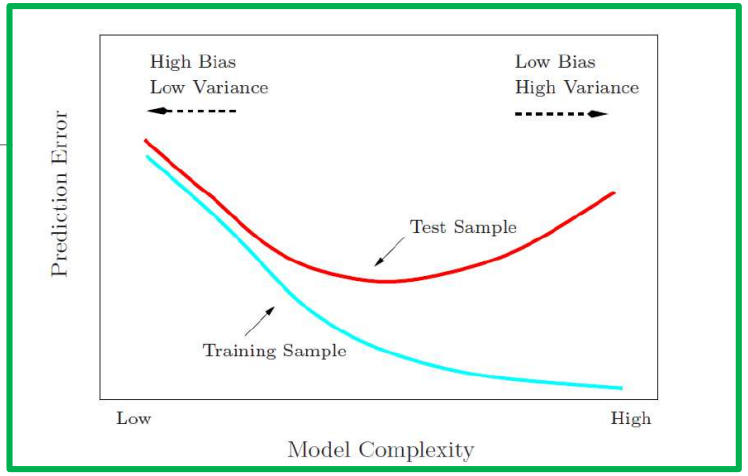
$$\hat{\sigma}^2 = \frac{e'e}{n-m},$$
$$e = y - X\hat{\beta}.$$



$$E(e) = (E(X\hat{\beta}) - X\beta)^2 + E(X\hat{\beta} - E(X\hat{\beta}))^2 + \sigma^2 = Bias^2 + Variance + \sigma^2$$

Это выражение уже нам встречалось!

Компромисс между дисперсией и смещением (variance-bias trade-off): что для линейных моделей, построенных методом наименьших квадратов?



n – число наблюдений

p – число переменных в линейной модели для МНК регрессии

Если истинная зависимость близка к линейной, то при n>> p

и дисперсия, и смещение – невелики (+++)

Если n> p, но незначительно (!), т.е. объем выборки ограничен, она (выборка) плохо представляет совокупность объектов, то модель демонстрирует большую дисперсию, она будет плохо работать на новых наблюдениях, не участвовавших в обучении!

Если n< p, то оценки параметров не единственные, МНК не применим, компьютер откажется строить модель!

Накладывая ограничения на оцененные коэффициенты при ограничении объема выборки, или «сжимая» их, мы можем существенно снизить дисперсию за счет незначительного увеличения смещения. Это может привести к существенным улучшениям точности предсказаний отклика на наблюдениях, которые не были использованы в обучении модели! Это т.н. сжатие коэффициентов!

Компромисс между дисперсией и смещением (variance-bias trade-off): что для линейных моделей, построенных методом наименьших квадратов?

---

n – число наблюдений

p – число переменных в линейной модели для МНК регрессии

Если истинная зависимость близка к линейной, то при n>> p и дисперсия, и смещение – невелики

Если n> p, но незначительно (!), т.е. объем выборки ограничен, она (выборка) плохо представляет совокупность объектов, то модель демонстрирует большую дисперсию, она будет плохо работать на новых наблюдениях, не участвовавших в обучении!

Если n< p, то оценки параметров не единственные, МНК не применим, компьютер откажется строить модель!

Многие переменные, включенные в регрессионную модель, в действительности не связаны с откликом. Включение таких несущественных переменных приводит к ненужной сложности итоговой модели.

Очень часто в модели пытаются использовать максимум наблюденных переменных! Удалив некоторые «ненужные» переменные, т. е. приравняв их коэффициенты к нулю, мы можем получить легко интерпретируемую модель. При использовании метода наименьших квадратов получение каких-либо коэффициентов, в точности равных нулю, крайне маловероятно.

Нужно смело (но обоснованно!) идти на выявление малоинформативных переменных и не включать их в модель!

Альтернативы использованию МНК при построении линейных моделей на исходном множестве признаков (предикторов, независимых переменных):

Отбор подмножества переменных: этот подход включает определение некоторого набора переменных, которые, как мы думаем, связаны с откликом. Далее мы подгоняем модель по методу наименьших квадратов с использованием этого уменьшенного набора переменных (без преобразований этих переменных).

Сжатие: Этот подход включает подгонку модели, содержащей все p предикторов. Однако, в отличие от коэффициентов, получаемых по методу наименьших квадратов, здесь оцененные коэффициенты «сжимаются» в направлении к нулю. Эффектом такого сжатия (также известно как «регуляризация») является снижение дисперсии. В зависимости от типа выполняемого сжатия оценки некоторых коэффициентов могут оказаться в точности равными нулю. Методы сжатия могут выполнять в ряде случаев и отбор переменных.

Снижение размерности: этот подход включает проецирование p предикторов на M-мерное подпространство, где $M < p$. Это достигается, в том числе, путем вычисления M различных линейных комбинаций, или проекций, переменных. Далее эти проекции используются в качестве предикторов для подгонки линейной регрессионной модели по методу наименьших квадратов. Очень часто в качестве новых переменных используются Главные Компоненты (PRINCIPAL COMPONENTS). Их построение и анализ является самостоятельным разделом прикладной статистики

Отбор подмножества переменных

Если p – исходное число переменных, то всего возможно рассмотреть 2^p вариантов подмножества переменных. Полный перебор в случаях больших p невозможен, несмотря на возможности вычислительных систем

Используют эвристические, но вычислительно реализуемые, подходы:

Пошаговое включение:

- Начинаем с пустого множества переменных, потом добавляем по одной переменной; на каждом шаге добавляется та переменная, которая обеспечивает наибольшее *приращение* качества модели. Если переменная попала в отбор, то она в дальнейшем не исключается

Пошаговое исключение:

- Отбор, противоположный пошаговому включению – этот метод начинает с полной модели, содержащей все p предикторов и подогнанной по методу наименьших квадратов, а затем последовательно, по одному за раз пробно исключается по одной переменной, в итоге исключается та, за счет исключения которой достигается наибольшее *приращение* качества модели. Если переменная попала в исключенные, она не может быть возвращена на последующих циклах.

Комбинированный поиск:

Похож на пошаговое включение и исключение, но при этом удаленная переменная на последующих этапах может быть возвращена, а включенная – может быть в последующем исключена. Это более длительная и ресурсоемкая процедура, но она все же значительно экономнее полного перебора!

Отбор подмножества переменных: критерии отбора

Модель, включающая все предикторы, всегда будет обладать минимальной RSS и максимальным значением R^2, поскольку эти величины тесно связаны с ошибкой обучения. По мере добавления переменных в модель ошибка на обучающей выборке будет снижаться, тогда как ошибка на контрольной выборке не обязательно будет вести себя тем же образом

Однако мы хотим выбрать модель с наименьшей ошибкой не на обучающей, а на контрольной выборке!!
Ошибка на обучающей выборке может оказаться плохой оценкой ошибки на контрольной выборке. Следовательно, RSS и R^2 не подходят для выбора оптимальной модели среди нескольких моделей с разным числом входящих в них предикторов!

Используют:

Статистики Ср (статистика Мэллоу), AIC (Информационный критерий Акайке), BIC (Байесовский информационный критерий)

Откорректированный R^2 (adjusted R-SQUARE): $$R_a^2 = \frac{ESS/(n-k-1)}{TSS/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1}$$

Непосредственно оценку ошибки на контрольных данных, применив для этого либо метод проверочной выборки (нужно разбить имеющиеся данные на обучающую и проверочную выборки), либо метод перекрестной проверки (cross-validation)

Сжатие:

«The simplest explanation is best“
Occam Razor (Бритва Оккама):
Из двух одинаковых по результативности моделей лучше та, которая проще

Регуляризация – прием, состоящий в том, что к целевой функции добавляется второе слагаемое, за счет которого предпочтительнее окажутся модели, имеющие меньшие квадраты значений коэффициентов. При этом второе слагаемое – функция коэффициентов модели, а не целевой переменной Y!

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\sum_{j=1}^{m} w_j^2$$

Это т.н. регуляризация по Тихонову.

Минимизация такого функционала приводит к так называемой гребневой, или ридж- регрессии (ridge regression).

где λ> 0 это *гиперпараметр,* который находят отдельно. Если он равен нулю – обычная оценка МНК

Другой прием:

LASSO регрессия (“Least Absolute Shrinkage and Selection Operator") – минимизация дополнительного слагаемого из коэффициентов в метрике L1 вместо L2 метрики для гребневой регрессии «ridge regression».

$$J(w,t) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

при условии:

$$\sum_{j=1}^{m}|w_j| \le t.$$

Сжатие: LASSO или гребневая регрессия?

Основная причина **преимущества гребневой регрессии** над методом наименьших квадратов заключается в обеспечении более оптимального баланса между смещением и дисперсией. При увеличении λ гибкость гребневой регрессии снижается, приводя при этом к более низкой дисперсии, но более высокому смещению.

В сравнении с методом отбора оптимального подмножества переменных, который требует поиска среди 2^p моделей, гребневая регрессия обладает также существенными вычислительными преимуществами. Для каждого значения λ нужно построить только одну модель.

Тем не менее:

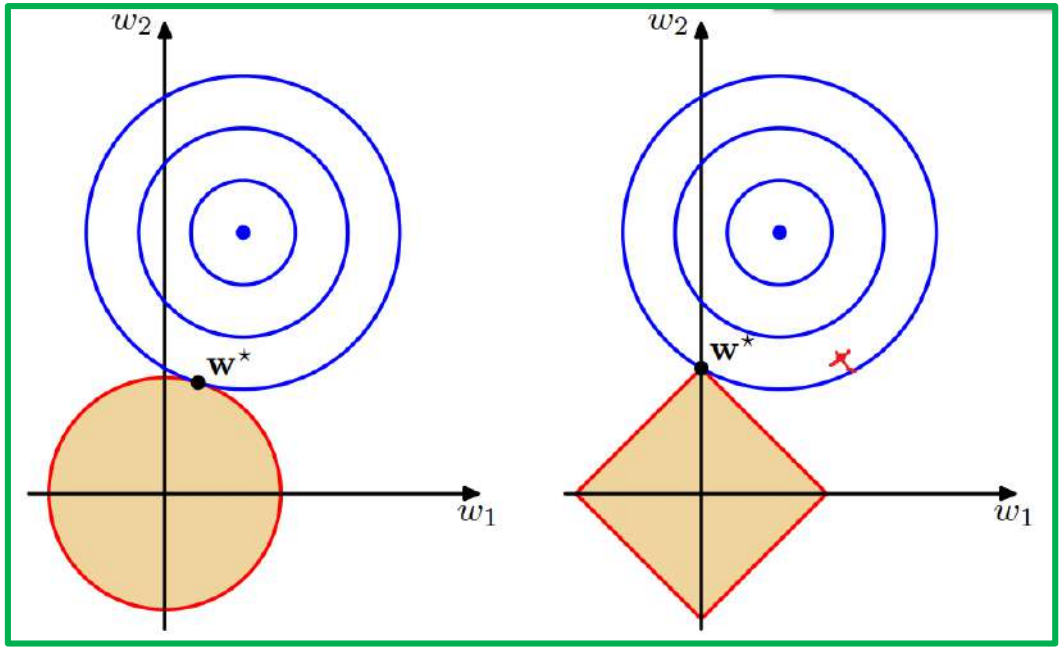
Штрафное слагаемое в выражении оптимизируемой суммы будет сжимать все коэффициенты в направлении к нулю, но никогда не приравняет ни один из этих коэффициентов в точности к нулю. Интерпретация результатов при этом не всегда облегчается.

Преимущества метода LASSO:

Однако в случае с LASSO -штраф при достаточно большом параметре λ делает некоторые оценки коэффициентов в точности равными нулю. Аналогично методу отбора оптимального подмножества переменных, метод LASSO выполняет отбор переменных.

В результате этого модели, полученные при помощи метода LASSO, обычно бывает легче интерпретировать, чем модели, полученные при помощи гребневой регрессии. Мы говорим, что LASSO дает разреженные модели, т. е. модели, которые включают только некоторое ограниченное подмножество переменных.

Сжатие: LASSO или гребневая регрессия?



LASSO или гребневая регрессия?
Слева – гребневая регрессия
Справа LASSO
В красной фигуре – пространство варьирования коэффициентов в штрафном слагаемом

Существует еще выпуклая комбинация между ридж-регрессией и LASSO – т.н. ELASTIC NET:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|\right),$$

где α - параметр смешивания ридж-регрессии и LASSO: ридж ($\alpha = 0$) и LASSO ($\alpha = 1$)

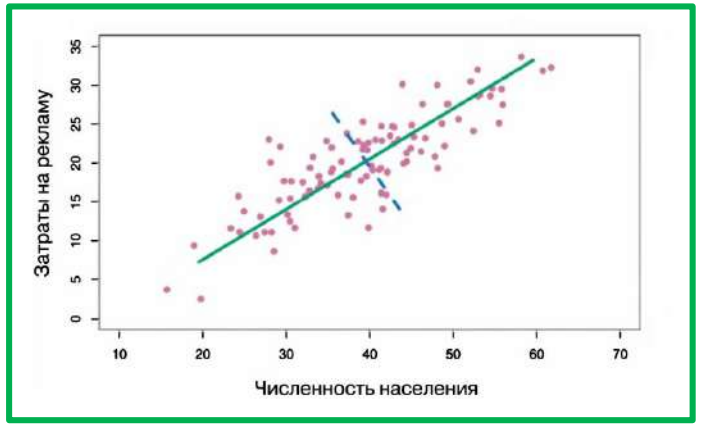
Рассмотрели две альтернативы обычному использованию МНК (пошаговый отбор предикторов и сжатие).
Если при пошаговом отборе и отбрасывании некоторых переменных мы фактически делали равными нулю коэффициенты при этих переменных, то при сжатии мы ограничивали коэффициенты по абсолютной величине. Делали это по-разному, в том числе некоторые делали точно равными нулю (как в LASSO!)

Снижение размерности

Регрессия не на исходные независимые переменные, а на главные компоненты

Главные компоненты – новые переменные, являющиеся ЛИНЕЙНЫМИ КОМБИНАЦИЯМИ исходных переменных

Это самостоятельный раздел прикладной статистики, который будем рассматривать отдельно



Метод регрессии на главные компоненты (PCR Principal Component Regression) включает вычисление первых M главных компонент Z_1,..., Z_m и последующее использование этих компонент в качестве предикторов в линейной регрессионной модели, которая подгоняется по методу наименьших квадратов. Ключевая идея заключается в том, что для объяснения основной доли дисперсии в данных, а также их связи с откликом часто достаточным оказывается применение лишь небольшого числа главных компонент.

Спасибо за внимание!

Лекция -окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

ТЕЛ. (484)3974658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

Раздел анализа – классификация – позиционирование *где его место?*

- Пусть обучающая выборка объектов выборка объема (размера) N .
  Обозначим:

$$x_1, \ldots, x_N = X_{train}, \{y_1, \ldots, y_N\} = Y_{train}$$

Тогда выделим следующие возможные типы задач:

- Обучение с учителем (supervised learning): Известны X_{train}, Y_{train}
- Обучение без учителя (unsupervised learning): Известно только X_{train}
- Частичное обучение (semi-supervised learning): Известно для всех ***N*** объектов X_{train} и для некоторых (!!!!) ***l*** объектов (***l<N)*** объектов из X_{train} - известна целевая переменная ***y***

Раздел анализа – классификация – позиционирование
где его место?

Имеется множество объектов. Каждый объект описывается вектором его наблюдаемых характеристик (признаков) $x \in X$ и скрытых характеристик $y \in Y$ (целевая переменная).

Существует (на всем множестве – на генеральной совокупности) некоторая функция $f : X \to Y$

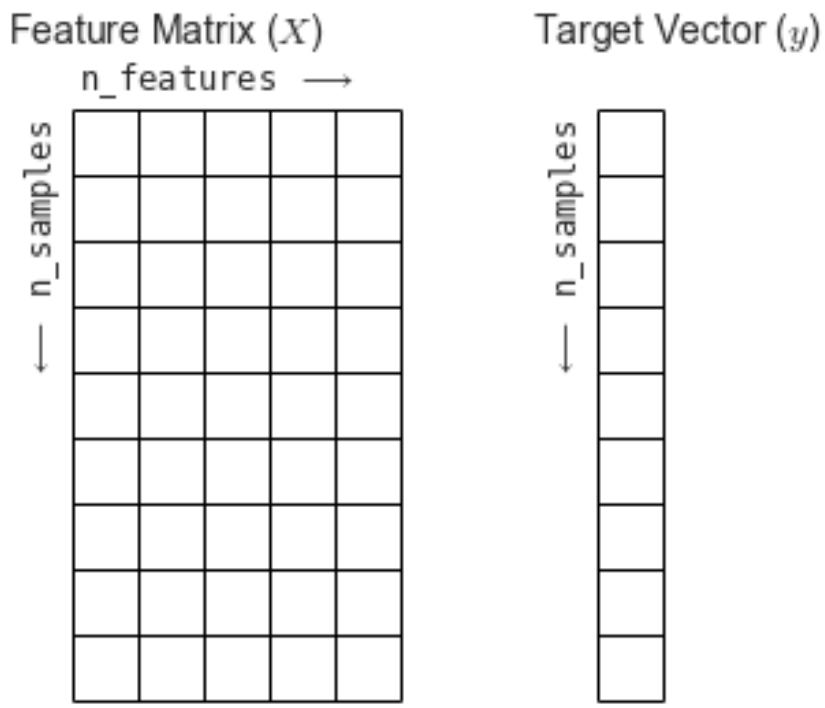
Задача: имея **ограниченный** набор объектов (обучающая выборка), построить функцию $a: X \to Y$, приближающую f на всем множестве объектов (на генеральной совокупности).

Какие бывают признаки (переменные в анализе, характеристики объектов, и т.д.)?

- Вещественный признак – принимает вещественные значения
- **Бинарный признак - может принимать 2 значения**
- **Категориальный признак - может принимать конечное число значений**
- Порядковый признак – упорядоченный категориальный признак

Раздел анализа – классификация – позиционирование *где его место?*

- ***В любом варианте задача подразумевает наличие матрицы данных, содержащей признаки X – ее строки – объекты, ее столбцы- признаки***



Если вариант задачи – обучение с учителем, то подразумевается наличие вектора значений целевой переменной Y

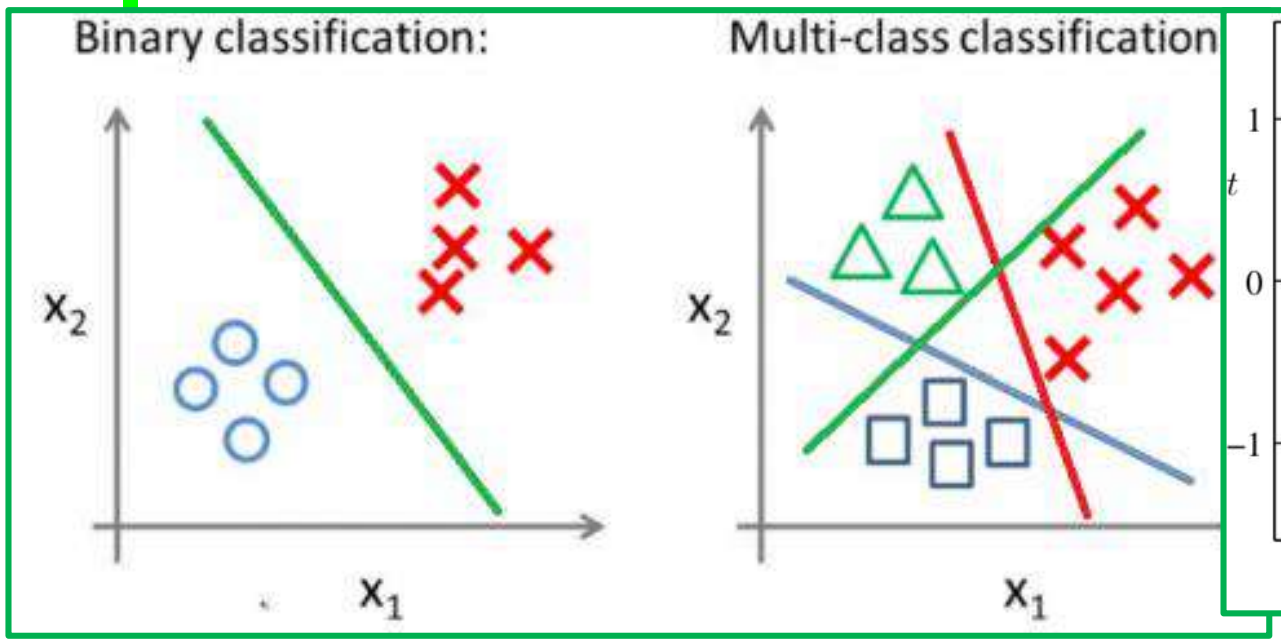
Y – значения из конечного множества чисел - классификация
Y - произвольные значения из множества вещественных чисел - регрессия

ANOVA и другие модели

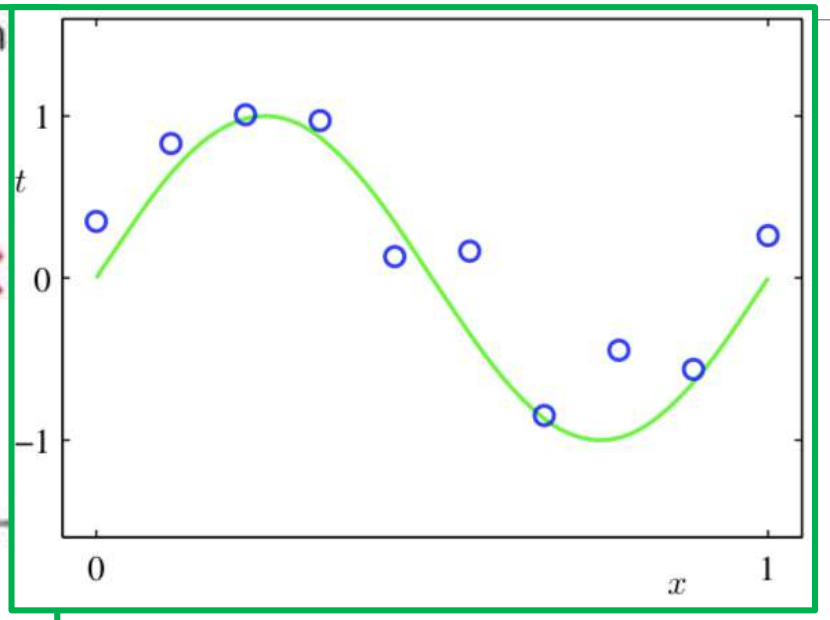
	ТИПЫ ПРЕДИКТОРОВ		
	КАТЕГОРИАЛЬНЫЕ	НЕПРЕРЫВНЫЕ	КАТЕГОРИАЛЬНЫЕ И НЕПРЕРЫВНЫЕ
ТИП ПРЕДИКТАНТА			
Непрерывные	Дисперсионный анализ Analysis of Variance (ANOVA)	«Обычная» (МНК) (OLS) регрессия и ее модификации и альтернативы метода МНК	Ковариационный анализ (ANCOVA)
Категориальные	Анализ таблиц сопряженности или **Логистическая регрессия**	**Логистическая регрессия**	**Логистическая регрессия**

Раздел анализа – классификация – позиционирование *где его место?*

- ***При обучении с учителем:***



Классификация



Регрессия

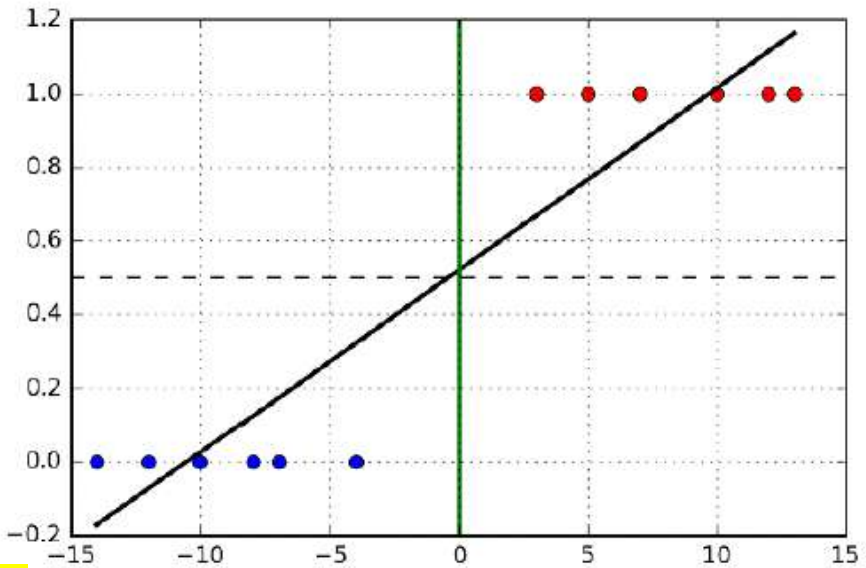
Раздел анализа – классификация – позиционирование
где его место?

- ***При обучении классификации с учителем:***

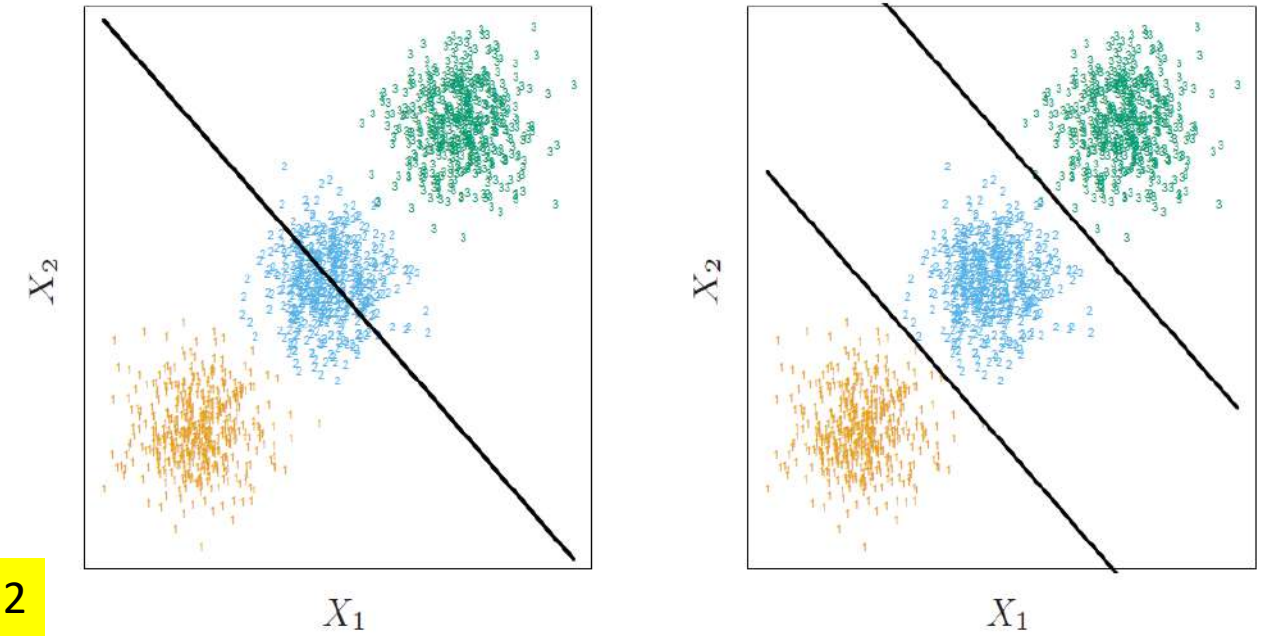
Примеры задач:

1. Пациент поступает в отделение клиники с набором симптомов, которые могут быть связаны с одним из заболеваний. Каким из заболеваний болен этот человек? (это дифференциальная диагностика,
2. Но как другой вариант – задача отличить больного от здорового человека, болен – не болен конкретной болезнью)
3. Сервис банковского обслуживания через Интернет должен иметь возможность определить, является ли та или иная онлайн- транзакция мошеннической?
4. Выдача кредита – будет ли дефолт у потенциального получателя кредита?
5. Для всех постановок задачи – нужна выборка с предикторами и с категоризирующей переменной

Классификация – обучение с учителем-использование регрессии бессмысленно!



1



2

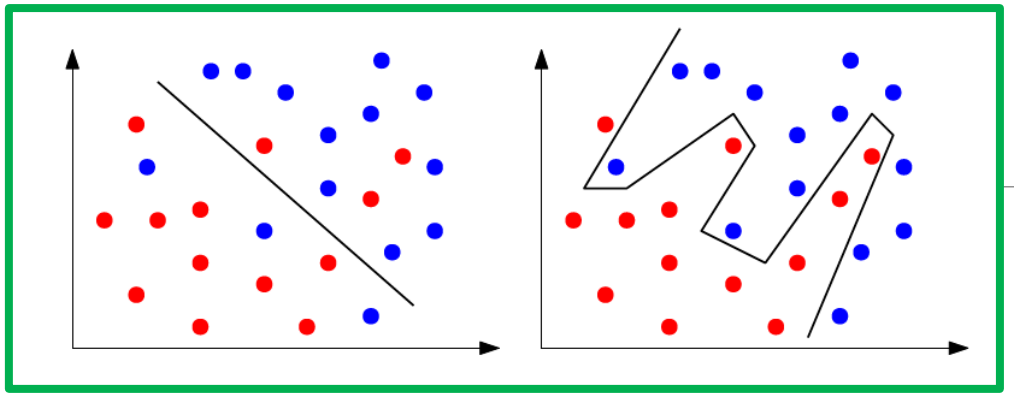
Бессмысленность использования линейной регрессии для классификации – восстановления значения категориальной переменной (классификация с учителем):

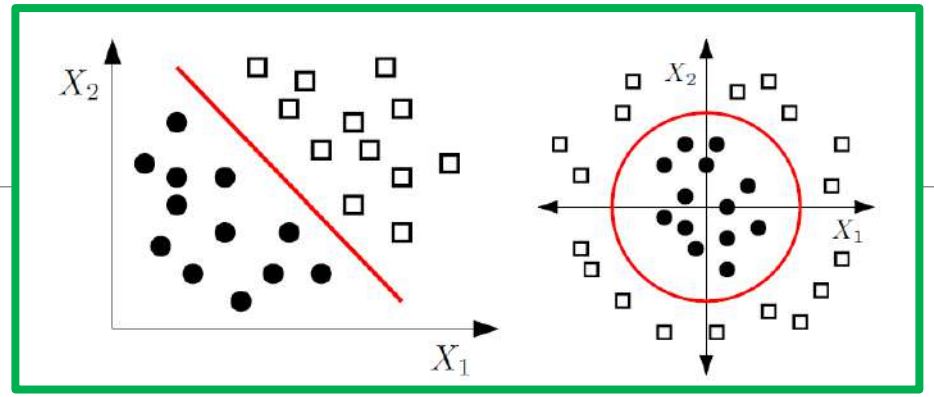
1 –черная линия – линейная регрессия, зеленая – разделяющая прямая. Восстановит значения больше 1 и меньше 0.

2 – Слева - линейная регрессия (черная линия)для случая **<u>классифицирующая переменная принимает три значения: 1 (желтые),2 (синие), 3 (зеленые).</u>**

Линейная регрессия (слева) дает абсурдный результат для переменной, равной 2 (синие). Справа – разделяющие поверхности

Классификация – обучение с учителем





Линейный классификатор (слева) –допускает ошибки классификации на обучающей выборке – **слишком** простая модель!
Сложная модель (справа) не допускает ошибки на обучающей выборке, но… это типичный пример переобучения!

Линейный классификатор (слева) – решает задачу классификации без ошибок на обучающей выборке (простая ситуация – линейно разделимый случай в пространстве переменных X_1 и X_2.
Сложная структура данных – линейная модель (справа) не подходит для классификации этих данных. Подходит в качестве разделяющей поверхности окружность (выделена красным). Есть другой вариант – введение дополнительной координаты X_3, тогда возможно линейное разделение!

Способы решения задачи классификации с учителем

1. **Логистическая регрессия**
2. **Разделяющие поверхности – линейные, квадратичные и т.д.**
3. **Байесова классификация**

Есть и другие методы (нейронные сети, деревья классификации, машины опорных векторов (SVM – Support Vector Machine), и др.

Все эти методы предполагают наличие обучающей выборки, по которой строится решающая функция (поверхность, правило) -функция разделения объектов на классы

Низкий процент ошибок найденной функции на обучающей выборке не есть показатель успеха решения задачи.

Качество найденной функции проверяется одним из существующих способов.

Функция с приемлемым качеством запоминается, и затем используется для классификации новых объектов

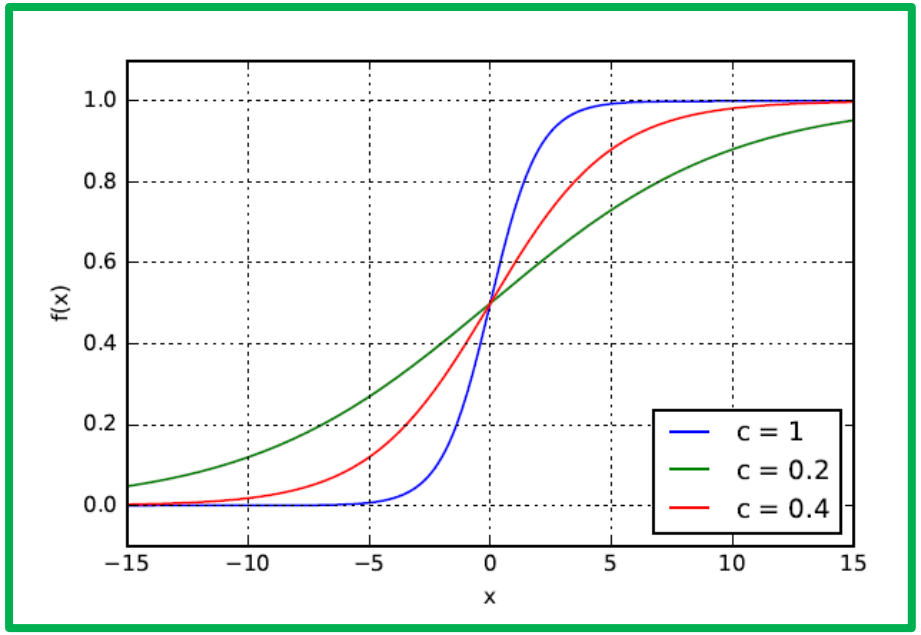
Если есть несколько моделей, построенных разными способами, то их качество может сопоставляться

Логистическая регрессия

Представим, что мы хотим преобразовать вещественную переменную X, заданную на интервале от минус до плюс бесконечности, в значения на отрезке [0,1] – сопоставить ее значениям значения вероятности P:

Для этого подходит, в частности,
Функция *f(x)*
Это т.н. логит-преобразование:

$$f(x) = \frac{1}{1 + e^{-cx}}$$

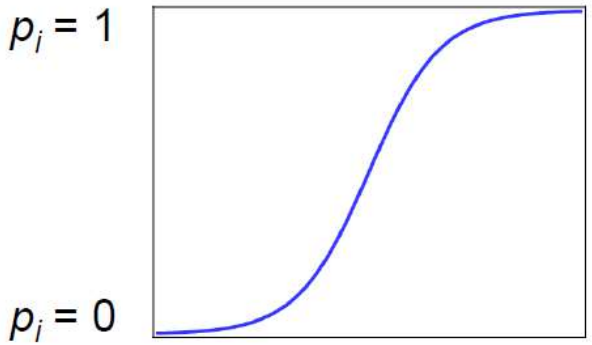


C – модулирующий параметр
C близко к 0 - пологая

Логистическая регрессия

posterior probability

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} = \eta$$



$$\Leftrightarrow p_i = \frac{1}{1+e^{-\eta}}$$

ЛОГИТ – логарифм отношения шансов – отношения вероятностей: вероятность того, что событие произойдет, к вероятности того, что событие не произойдет

Поскольку логистическое преобразование решает проблему об ограничении на 0-1 границы для первоначальной зависимой переменной (вероятности), то эти преобразованные значения можно использовать в обычном линейном регрессионном уравнении. А именно, если произвести логистическое преобразование обеих частей описанного выше уравнения, мы получим стандартную модель линейной регрессии!

Логистическая регрессия: некоторые соотношения и особенности - более детально

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Вычислим, чему равно 1-p(X), и разделим одну вероятность (того, что событие состоится), на другую (того, что событие не состоится).

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

Левая часть – РИСК СОБЫТИЯ (ODDS). Он изменяется от нуля до бесконечности
Возьмем логарифм обеих частей этого равенства:

$$\log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$

Левая часть – логит – линеен по X! В логистической регрессионной модели увеличение X на одну единицу изменяет логарифм риска коэффициент при X, или, что то же самое, умножает риск на экспоненту коэффициента при X
Можем расширить на несколько предикторов:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Логистическая и линейная регрессия – сходства и различия

Математическая модель логистической регрессии

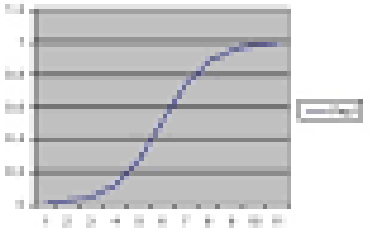
Линейная регрессия

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = \beta^T x$$

Логистическая регрессия

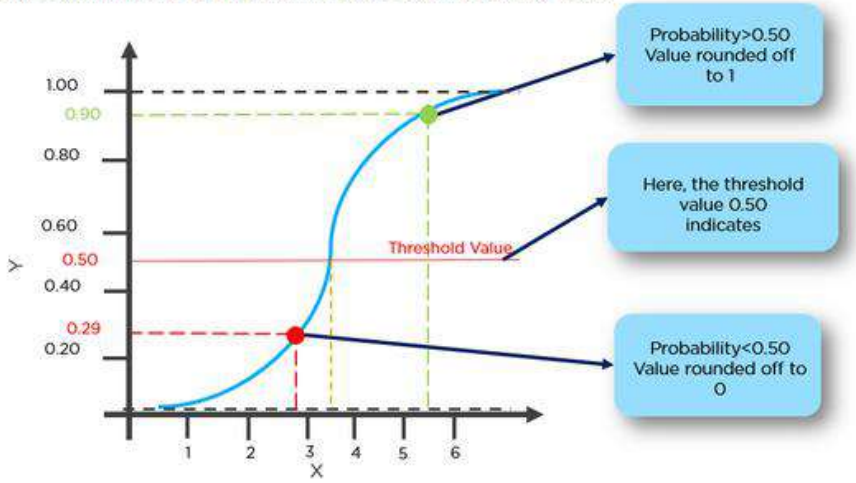
$$h_\beta(x) = g(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = g(\beta^T x)$$

где $g(z) = \frac{1}{1+e^{-z}}$ - сигмоида



Получаем $h_\beta(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_k x_k)}}$

Ex=we can make classification of tumour having more than 50% chance of having cancer as 1 and tumour having less than 50% chance of having cancer as 0.



Here red point will be predicted as 0 whereas blue point will be predicted as 1.

Логистическая регрессия

В логистической регрессии порог отсечения изменяется от 0 до 1 – это и есть расчетное значение уравнения регрессии. **Будем называть его рейтингом**.

Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности, или матрицу сопряженности (confusion matrix), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам.

	Фактически	
Модель	**положительно**	**отрицательно**
положительно	TP	FP
отрицательно	FN	TN

Это матрица 2x2

	Фактически	
Модель	**положительно**	**отрицательно**
положительно	TP	FP
отрицательно	FN	TN

TP (*True Positives*) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);
TN (*True Negatives*) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);
FN (*False Negatives*) – положительные примеры, классифицированные как отрицательные **(ошибка I рода).** Это так называемый "ложный пропуск" (или «пропуск цели»
– когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);
FP (*False Positives*) – отрицательные примеры, классифицированные как положительные **(ошибка II рода);** Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи, «ложная тревога»).

Логистическая регрессия

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи. Например, если мы прогнозируем вероятность наличия заболевания, то положительным исходом будет класс "Больной пациент", отрицательным – "Здоровый пациент". И наоборот, если мы ходим определить вероятность того, что человек здоров, то положительным исходом будет класс "Здоровый пациент", и так далее.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями (rates), выраженными в процентах

Доля **истинно положительных примеров** (True Positives Rate – это и есть чувствительность, как будет показано ниже):

$$TPR = \frac{TP}{TP + FN} \cdot 100\,\%$$

Доля **ложно положительных примеров** (False Positives Rate, это и есть 100-специфичность, как будет показано ниже):

$$FPR = \frac{FP}{TN + FP} \cdot 100\,\%$$

	Фактически	
Модель	**положительно**	**отрицательно**
положительно	TP	FP
отрицательно	FN	TN

Логистическая регрессия

Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (Sensitivity) – это доля истинно положительных случаев, которые правильно идентифицированы :

$$S_e = TPR = \frac{TP}{TP + FN} \cdot 100\,\%$$

Специфичность (Specificity) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$S_p = \frac{TN}{TN + FP} \cdot 100\,\%$$

Заметим, что FPR=100%-Sp.

	Фактически	
Модель	**положительно**	**отрицательно**
положительно	TP	FP
отрицательно	FN	TN

Логистическая регрессия

Заметим, что FPR=100%-Sp.

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры).

Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

Если рассуждать в терминах медицины:

Задачи диагностики заболевания, где модель классификации пациентов на больных и здоровых называется диагностическим тестом, то получится следующее:

Чувствительный диагностический тест проявляется в гипердиагностике – максимальном предотвращении пропуска больных (когда важнее всего не пропустить опасное заболевание);

Специфичный диагностический тест диагностирует только доподлинно больных. Это важно в случае, когда, например, лечение больного связано с серьезными побочными эффектами и гипердиагностика пациентов не желательна.

ROC -кривые

ROC-кривая (Receiver Operator Characteristic) – кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении. Название пришло из систем обработки сигналов.

Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами. Какой-то один исход обязательно нужно считать положительным, а другой – отрицательным.

ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров.

В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. При этом предполагается, что у классификатора имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение на два класса. Этот параметр часто называют порогом, или точкой отсечения (cut-off value).

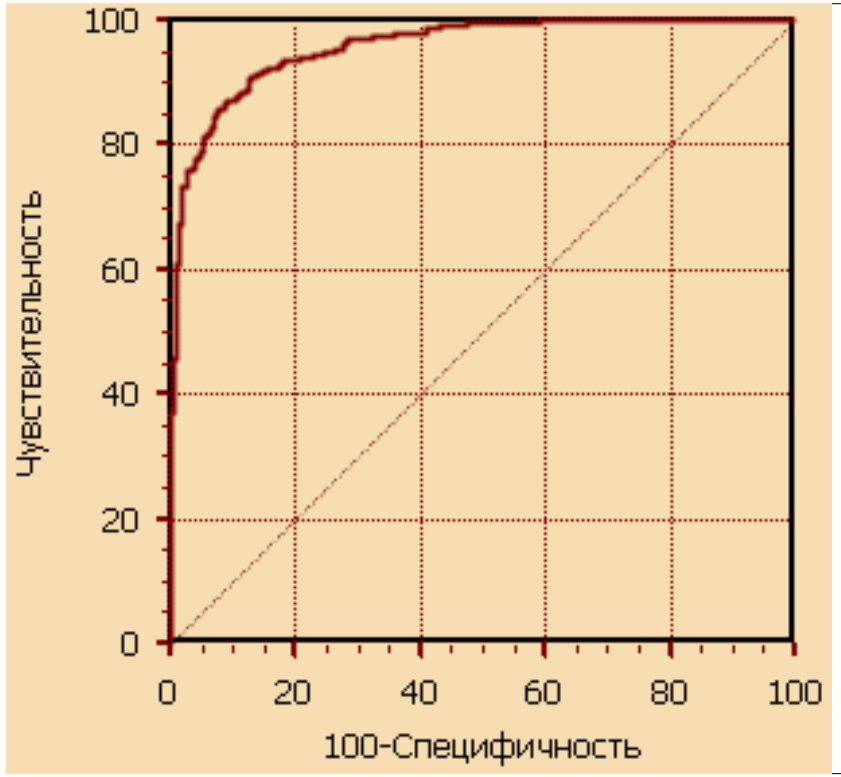
В зависимости от него будут получаться различные величины ***ошибок I и II рода***.

Логистическая регрессия: ROC-кривая

ROC-кривая получается следующим образом:

Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом dx (например, 0.01) рассчитываются значения чувствительности Se и специфичности Sp. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.

Строится график зависимости: по оси Y откладывается чувствительность Se, по оси X – 100%–Sp (сто процентов минус специфичность), или, что то же самое, FPR – доля ложно положительных случаев.

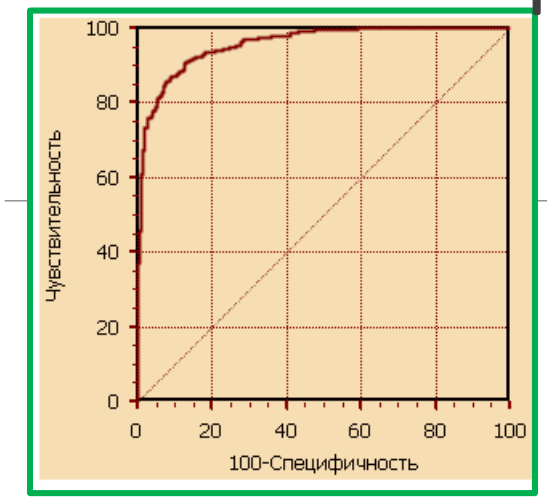


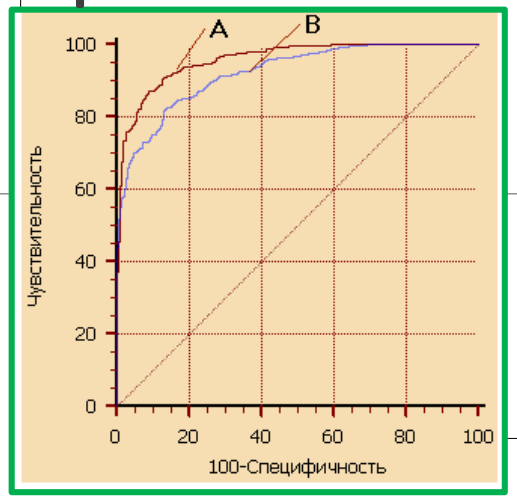
Для идеального классификатора график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100% или 1.0 (идеальная чувствительность), а доля ложно положительных примеров равна нулю. Поэтому чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели.

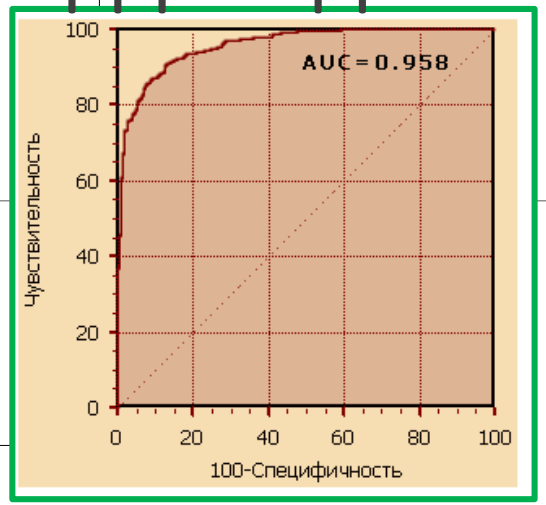
Наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагональной прямой, тем менее эффективна модель.

Диагональная линия соответствует "бесполезному" классификатору, т.е. полной неразличимости двух классов – просто «бросание монетки».

Логистическая регрессия – площадь под ROC кривой







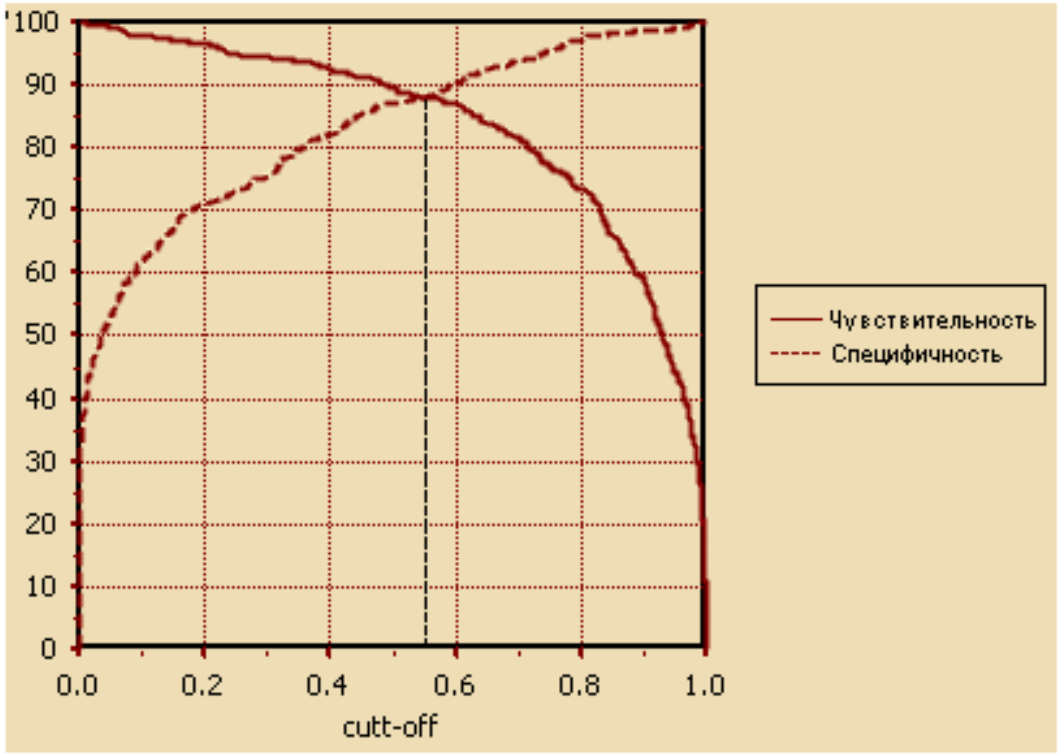
AUC – Area Under Curve

Для расчета AUC можно использовать формулу метода трапеций

Интервал AUC	Качество модели
0.9-1.0	Отличное
0.8-0.9	Очень хорошее
0.7-0.8	Хорошее
0.6-0.7	Среднее
0.5-0.6	Неудовлетворительное

Идеальная модель обладает 100% чувствительностью и специфичностью. Однако на практике добиться этого невозможно, более того, невозможно одновременно повысить и чувствительность, и специфичность модели.

Компромисс находится с помощью порога отсечения, т.к. пороговое значение влияет на соотношение Se и Sp. Можно говорить о задаче нахождения *оптимального порога отсечения* (optimal cut-off value).



"Точка баланса" между чувствительностью и специфичностью

Что может быть критерием выбора:

- Требование минимальной величины чувствительности (специфичности) модели. Например, нужно обеспечить чувствительность теста не менее 80%. В этом случае оптимальным порогом будет максимальная специфичность (чувствительность), которая достигается при 80% (или значение, близкое к нему "справа" из-за дискретности ряда) чувствительности (специфичности).
- Требование максимальной суммарной чувствительности и специфичности модели, т.е.

$$\text{Cutt\_ off}_o = \max_k (\text{Se}_k + \text{Sp}_k)$$

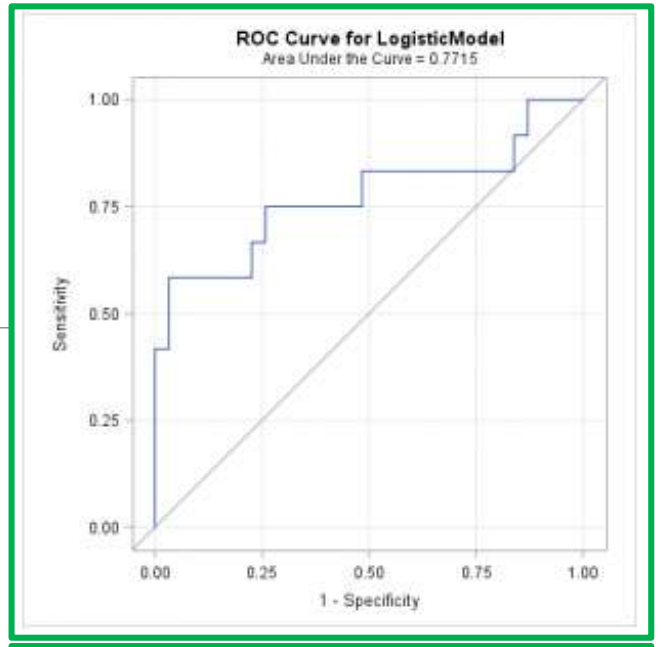
- Требование баланса между чувствительностью и специфичностью, т.е. когда $\text{Se} \approx \text{Sp}$:

$$\text{Cutt\_ off}_o = \min_k |\text{Se}_k - \text{Sp}_k|$$

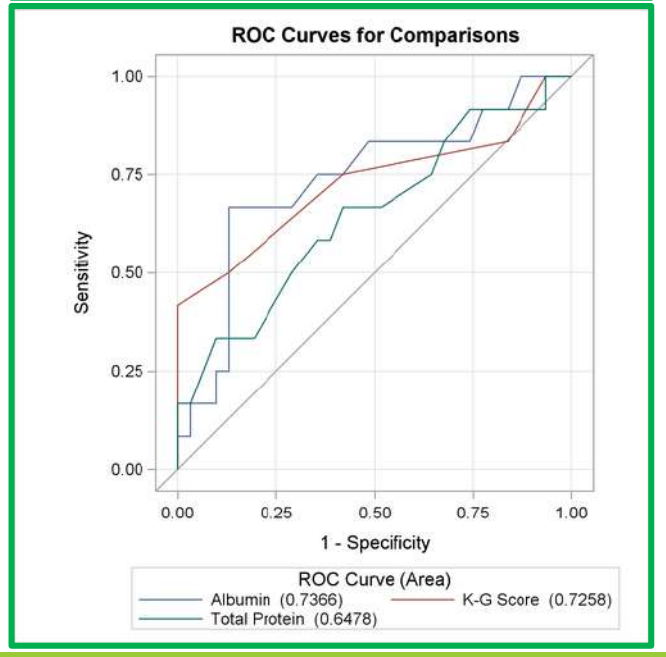
Реализация в SAS: PROC LOGISTIC

```
data roc;
   input alb tp totscore popind @@;
   totscore = 10 - totscore;
   datalines;
3.0 5.8 10 0   3.2 6.3  5 1   3.9 6.8  3 1   2.8 4.8  6 0
3.2 5.8  3 1   0.9 4.0  5 0   2.5 5.7  8 0   1.6 5.6  5 1
3.8 5.7  5 1   3.7 6.7  6 1   3.2 5.4  4 1   3.8 6.6  6 1
4.1 6.6  5 1   3.6 5.7  5 1   4.3 7.0  4 1   3.6 6.7  4 0
2.3 4.4  6 1   4.2 7.6  4 0   4.0 6.6  6 0   3.5 5.8  6 1
3.8 6.8  7 1   3.0 4.7  8 0   4.5 7.4  5 1   3.7 7.4  5 1
3.1 6.6  6 1   4.1 8.2  6 1   4.3 7.0  5 1   4.3 6.5  4 1
3.2 5.1  5 1   2.6 4.7  6 1   3.3 6.8  6 0   1.7 4.0  7 0
3.7 6.1  5 1   3.3 6.3  7 1   4.2 7.7  6 1   3.5 6.2  5 1
2.9 5.7  9 0   2.1 4.8  7 1   2.8 6.2  8 0   4.0 7.0  7 1
3.3 5.7  6 1   3.7 6.9  5 1   3.6 6.6  5 1
;

ods graphics on;
proc logistic data=roc plots(only)=roc;
   LogisticModel: model popind(event='0') = alb tp totscore;
   output out=LogiOut predicted=LogiPred;       /* output predicted value, to be used later */
run;
```




```
ods graphics on;
proc logistic data=roc plots=roc(id=prob);
   model popind(event='0') = alb tp totscore / nofit;
   roc 'Albumin' alb;
   roc 'K-G Score' totscore;
   roc 'Total Protein' tp;
   roccontrast reference('K-G Score') / estimate e;
run;
```



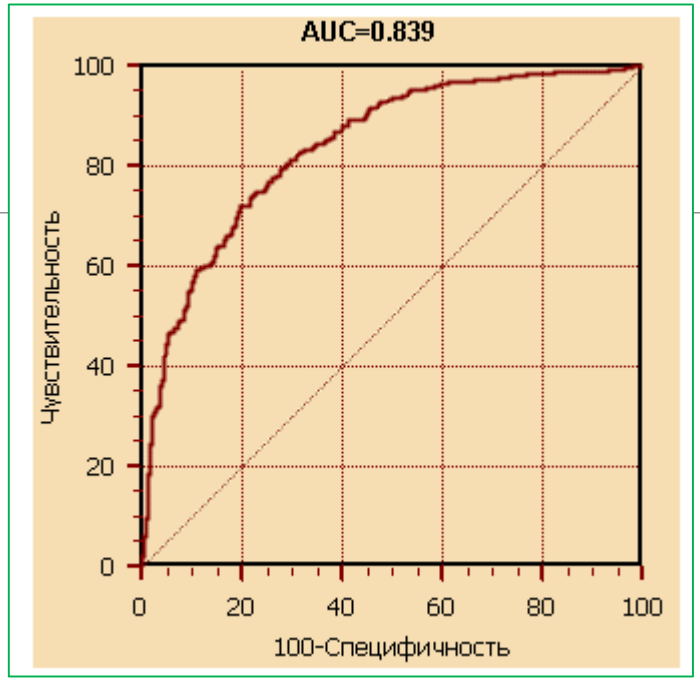
Пример - диагностика диабета: Обучающая выборка содержит 768 записей со следующими полями:

1. Концентрация глюкозы;
2. Число случаев беременности;
3. Артериальное диастолическое давление, мм. рт. ст.;
4. Толщина кожной складки трехглавой мышцы, мм.;
5. 2-х часовой сывороточный инсулин;
6. Индекс массы тела;
7. Числовой параметр наследственности диабета;
8. Возраст, лет;
9. Зависимая переменная (1 – наличие заболевания, 0 – отсутствие).

Распределение зависимой переменной следующее: 500 случаев отсутствия заболевания, 268 – его наличие.

Рассчитанные коэффициенты логистической регрессии приведены в таблице

Независимая переменная	Коэффициент
1	0.1232
2	0.0352
3	-0.0133
4	0.0006
5	-0.0012
6	0.0897
7	0.9452
8	0.0149
Константа	-8.4047



фрагмент массива точек "Чувствительность-Специфичность"

Порог	Se, %	Sp, %	Se+Sp	|Se-Sp|
...	...	...	...	...
0.25	84.3	65.0	149.3	19.3
0.26	83.6	65.6	149.2	18.0
0.27	83.2	67.4	150.6	15.8
...	...	...	...	...
0.31	78.0	73.0	151.0	5.0
0.32	76.1	75.0	151.1	1.1
0.33	75.4	75.6	151.0	**0.2**
0.34	75.0	76.8	151.8	1.8
0.35	74.3	77.8	**152.1**	3.5
0.36	72.0	79.2	151.2	7.2
0.37	70.9	80.2	151.1	9.3
0.38	69.4	80.8	150.2	11.4
0.39	69.3	81.2	150.5	11.9
0.40	67.2	82.0	149.2	14.8
...	...	...	...	...
0.49	58.6	88.8	147.4	30.2
0.50	58.2	89.0	147.2	30.8
0.51	57.8	89.2	147.0	31.4
...	...	...	...	...

Оптимальным порогом классификации, обеспечивающим максимум чувствительности и специфичности теста (или минимум ошибок I и II рода), является точка 0.35. В ней чувствительность равна 74.3%, что означает: у 74.3% пациентов с наличием диабета диагностический тест будет положителен. Специфичность равна 77.8%, следовательно, у 77.8% пациентов, у которых нет диабета, результаты теста отрицательны.
Точкой баланса, в которой чувствительность и специфичность примерно совпадают, является 0.33.

Если мы выберем порог 0.25, в котором чувствительность теста очень высокая (>84%), то получим гипердиагностику пациентов. А если зафиксировать порог на уровне 0.5, то будем диагностировать только доподлинно больных (специфичность 89%). Что считать здесь оптимальным порогом? Все зависит от конкретной задачи, универсальных рецептов нет. В диагностике диабета, наверное, следует выбрать наиболее чувствительный тест: ложноположительный результат может угрожать, например, лишь дополнительным визитом к врачу, а ложноотрицательный – не выявлением опасной, но излечимой болезни.

Коэффициент Джини (Gini coefficient)

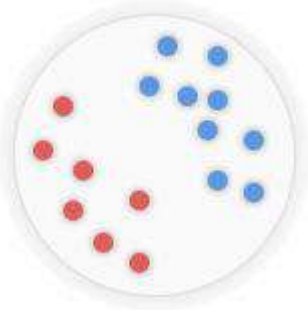
Коэффициент Джини (**Gini Coefficient):** 2*AUC – 1,

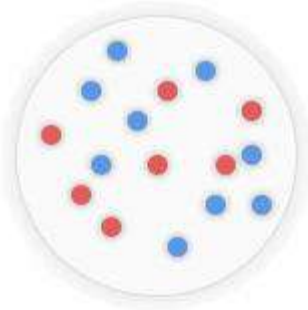
Для нормализации площади под кривой (AUC),

Изменяется в пределах [-1, 1].

В экономике – используется для оценки неравномерности сосредоточения богатств у разных групп населения

Метрические алгоритмы:
Гипотезы непрерывности и компактности





Непрерывность (для регрессии):
близким объектам X соответствуют, как правило, близкие Y
Слева – выполняется, справа – не выполняется

Компактность (для классификации):
близкие объекты X принадлежат, как правило, одному классу (им соответствуют одинаковые Y)
Слева – выполняется, справа – не выполняется

Метрические алгоритмы – зачем они нужны?

$$a(x, X_{train}) = \underset{c}{\operatorname{argmax}} \sum_{i=1}^{N} w(x, x_i) I[y_i = c], x_i \in X_{train}$$

Классификация

$$a(x, X_{train}) = \frac{\sum_{i=1}^{N} w(x, x_i) y_i}{\sum_{i=1}^{N} w(x, x_i)}, x_i \in X_{train}$$

Регрессия

Можем построить непараметрический, **ленивый** алгоритм:

- Если ненулевой вес только у ближайшего объекта, то алгоритм называют алгоритмом ближайшего соседа
- Если ненулевые веса для k ближайших объектов, то алгоритм называют алгоритмом k ближайших соседей (k-nearest neighbors, knn, kNN).
- X - объект с неизвестным значением Y появляется, мы находим для него одного (или k) ближайших соседей в обучающей выборке, и по значениям Y для найденных соседей - определяем Y для этого объекта

Надо научиться измерять расстояния!
Такой алгоритм «ничему не учится»!

Метрические алгоритмы

Алгоритм ближайшего соседа – просто нужно запомнить всю обучающую выборку, а затем найти расстояния от заданного X ко всем элементам X обучающей выборки, и классифицировать X так, как классифицирован ближайший к нему вектор обучающей выборки.

Его достоинство – идейная простота, насчет простоты реализации – вопрос…

Множество недостатков: он чувствителен к погрешностям (если ближайший сосед – нетипичный выброс, то и классификация X будет также нетипичным выбросом). Он ничему не «учится» - нет параметров, настраиваемых по выборке, есть только зависимость от успешного или неуспешного выбора метрики близости. В результате - низкое качество классификации

Алгоритм k ближайших соседей (k nearest neighbors, kNN). Чтобы сгладить влияние выбросов, будем относить классифицируемый объект X к тому классу, элементов которого окажется больше среди k ближайших соседей $x(i)$, $i = 1, \ldots, k$

K=1 или l – два противоположных вырожденных случая, оба нежелательны

На практике оптимальное значение параметра k определяют по критерию скользящего контроля с исключением объектов по одному (leave-one-out, LOO).

Метрические алгоритмы

Для произвольного $x \in X$ отранжируем объекты $x_1, \ldots, x_\ell$:

$$\rho(x, x^{(1)}) \leqslant \rho(x, x^{(2)}) \leqslant \cdots \leqslant \rho(x, x^{(\ell)}),$$

$x^{(i)}$ — i-й сосед объекта x среди $x_1, \ldots, x_\ell$;
$y^{(i)}$ — ответ на i-м соседе объекта x.

Метрический алгоритм классификации:

$$a(x; X^\ell) = \arg\max_{y \in Y} \underbrace{\sum_{i=1}^{\ell} [y^{(i)} = y]\, w(i, x)}_{\Gamma_y(x)},$$

$w(i, x)$ — вес (степень важности) i-го соседа объекта x,
неотрицателен, не возрастает по i.
$\Gamma_y(x)$ — *оценка близости* объекта x к классу y.

Метрические алгоритмы

Метод k ближайших соседей (k nearest neighbors, kNN)

$w(i,x) = [i \leqslant 1]$ — метод ближайшего соседа
$w(i,x) = [i \leqslant k]$ — метод k ближайших соседей

Преимущества:

- простота реализации (lazy learning);
- параметр k можно оптимизировать по критерию скользящего контроля (leave-one-out):

$$\mathrm{LOO}(k, X^\ell) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \setminus \{x_i\}, k\big) \neq y_i \Big] \to \min_k .$$

Недостатки:

- неоднозначность классификации при $\Gamma_y(x) = \Gamma_s(x), \quad y \neq s$
- не учитываются значения расстояний

Метрические алгоритмы

Метод k взвешенных ближайших соседей

$w(i,x) = [i \leqslant k]w_i$,

где w_i — вес, зависящий только от номера соседа;

Возможные эвристики:

$w_i = \frac{k+1-i}{k}$ — линейное убывающие веса;

$w_i = q^i$ — экспоненциально убывающие веса, $0 < q < 1$;

Проблемы:

- как более обоснованно задать веса?
- возможно, было бы лучше, если бы вес $w(i,x)$ зависел не от порядкового номера соседа i, а от расстояния до него $\rho(x, x^{(i)})$.

Метод k взвешенных ближайших соседей – переход к Парзеновскому окну и к методу потенциальных функций

Проверка качества классификации. Случайное разбиение выборки – всегда ли хорошо?

Качество зависит от объектов в валидации!
Решение — скользящий контроль (cross-validation).
В пределе, когда только 1 объект в тесте — LOO (leave one out).

Нельзя никогда забывать, какую задачу мы решаем!
Если выборка маленькая, то нужно сохранять баланс классов — stratified валидация. Каждая градация классифицирующей переменной должна быть представлена!

Как сделать валидацию в случае:
• Спам-фильтра (примеров спама всегда намного меньше, чем не-спама, должна быть обеспечена **стратифицированная балансировка)**
• Предсказания объема продаж на следующую неделю (эта задача – **экстраполяция**, нельзя ее сводить к задаче **интерполяции!**)
• Предсказания стоимости квартир для всего дома целиком

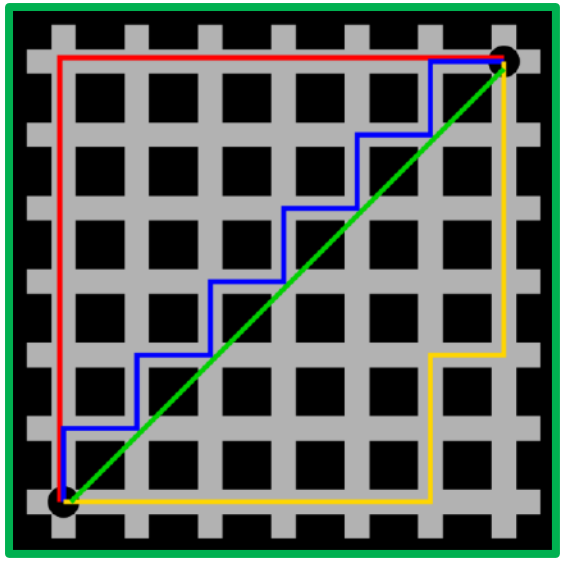
Метрики

$$\rho(x, y) = \left(\sum_{j=1}^{D} |x_j - y_j|^P\right)^{\frac{1}{P}}$$

Нельзя так считать, если признаки разных масштабов!!!!!!

D – число признаков

P=2 – Евклидово расстояние
P=1 - Манхэттенское расстояние
P=бесконечности - Расстояние Чебышева – максимальное расстояние между признаками



Если признаки разных масштабов – нужна стандартизация или нормализация

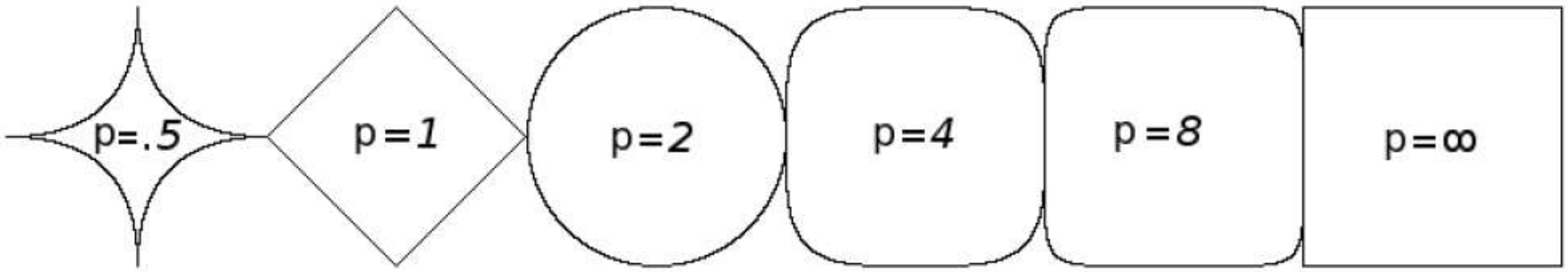
$$x^j = \frac{x^j - mean(x^j)}{std(x^j)}$$

$$x^j = \frac{x^j - min(x^j)}{max(x^j) - min(x^j)}$$

Метрики

$$\rho(x, x_i) = \left(\sum_{j=1}^{n} |x^j - x_i^j|^2 \right)^{1/2} \qquad \rho(x, x_i) = \left(\sum_{j=1}^{n} w_j |x^j - x_i^j|^p \right)^{1/p}$$

Евклидова метрика (p=2) и обобщенная метрика Минковского (произвольное p) для вычисления расстояния между объектами X и X_i



Границы областей близости между двумя объектами для разных p в метрике Минковского

Другие метрики

Расстояния на категориальных признаках:

1. Расстояние Хэмминга — число категориальных признаков, которые имеют разные значения.

2. Счетчики — среднее значение признака/целевой переменной с такой категорией

Косинусное расстояние:

По определению скалярного произведения считаем угол между векторами

$$\rho(x, y) = \alpha = arccos \frac{x \cdot y}{|x||y|}$$

$$sim(x, y) = \frac{x \cdot y}{|x||y|}$$

$$rho(x, y) = 1 - sim(x, y)$$

Другие метрики

Расстояние Джаккарда:

Как померить расстояние между множествами? Например, предложение — мешок (множество) слов. Есть два множества слов – два предложения ***X, Y:***

$$\rho(X, Y) = 1 - \frac{X \cap Y}{X \cup Y}$$

Расстояние Левенштейна:

Минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую. Надо сделать три операции:

$$\rho(kitten, sitting) = 3$$

Для строк - редакторское расстояние Левенштейна:
CTGGGCTAAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC
CTGGGACTAAA....CCTTAGCCTATTTACAAAAATGGGCCATTAGG...TTGC

Качество классификационных алгоритмов

•Accuracy (точность) — процент правильно классифицированных объектов

• Precision (аккуратность) —процент правильно классифицированных объектов класса 1 среди всех объектов, которым алгоритм присвоил метку 1.

• Recall (полнота) — процент правильно классифицированных объектов класса 1 среди всех объектов класса 1

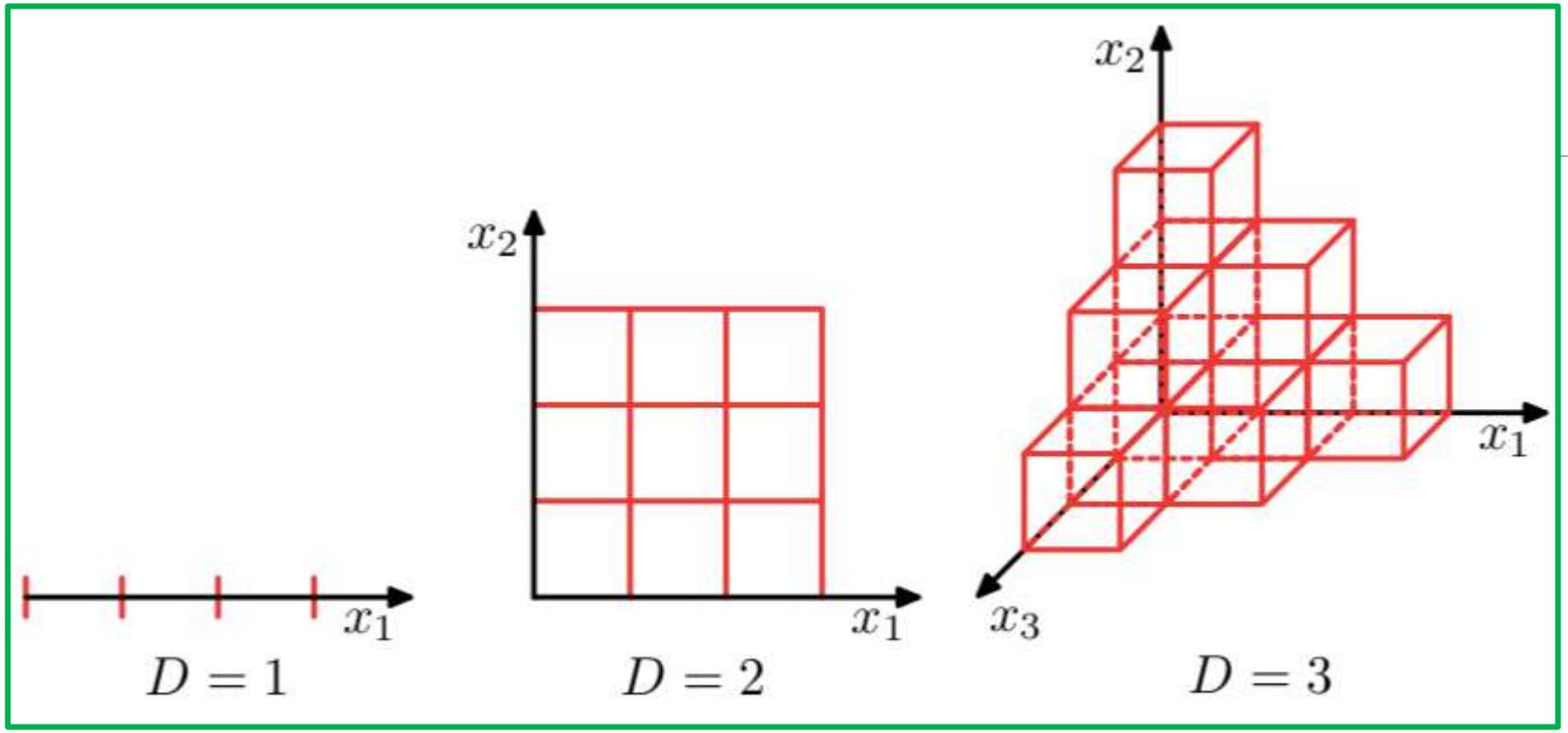
• F1-score — среднее гармоническое Precision и Recall

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Тестовая выборка содержит 10 объектов класса 1 и 990 объектов класса 0. Какая точность у константного алгоритма?

Почему именно среднее гармоническое? Среднее гармоническое сильнее реагирует, когда сомножители в числителе резко отличаются!

Проклятие размерности – curse of dimensionality



В пространстве большой размерности объекты сильно удалены друг от друга!!!

Curse of dimensionality - проблема, связанная с экспоненциальным возрастанием количества данных из-за увеличения размерности пространства. Термин «проклятие размерности» был введен Ричардом Беллманом в 1961 году. Проблема «проклятия размерности» часто возникает в машинном обучении, например, при применении метода ближайших соседей

Что же делать?????

Отбор признаков (для регрессии – пошаговая регрессия)

Переходить в «синтезированное» пространство меньшей размерности с минимальной потерей информативности и решать задачи в новом пространстве (для регрессии – регрессия на главные компоненты)

Схожие решения - и для классификации!

Проклятие размерности – curse of dimensionality

«Проклятие размерности» особенно явно проявляется при работе со сложными системами, которые описываются большим числом параметров.
Это влечет за собой следующие трудности:
•Трудоемкость вычислений
•Необходимость хранения огромного количества данных
•Увеличение доли шумов
•В линейных классификаторах увеличение числа признаков ведет к проблемам мультиколлинеарности и переобучения.
•В метрических классификаторах расстояния обычно вычисляются как средний модуль разностей по всем признакам.
Согласно Закону Больших Чисел, сумма n слагаемых стремится в некоторому фиксированному пределу при $n \to \infty$.
Таким образом, расстояния во всех парах объектов стремятся к одному и тому же значению, а значит, становятся неинформативными.

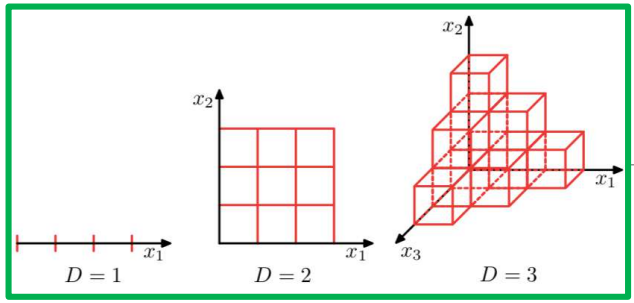
Пример

Рассмотрим единичный интервал [0,1]. 100 равномерно разбросанных точек будет достаточно, чтобы покрыть этот интервал с частотой не менее 0,01.
Теперь рассмотрим 10-мерный куб. Для достижения той же степени покрытия потребуется уже 10^{20} точек. То есть, по сравнению с одномерным пространством, требуется в 10^{18} раз больше точек.
Поэтому, например, использование переборных алгоритмов становится неэффективным при возрастании размерности системы.
И это показывает фундаментальную проблему размерности при использовании алгоритма k-ближайших соседей; с увеличением числа измерений и приближением отношения ближайшего к среднему расстоянию предсказательная сила алгоритма уменьшается. Если ближайшая точка находится почти так же далеко, как средняя точка, то она обладает лишь немного большей предсказательной силой, чем средняя точка.

Способы устранения «проклятия размерности»

Основная идея при решении проблемы — понизить размерность пространства, а именно спроецировать данные на подпространство меньшей размерности.

Проклятие размерности – curse of dimensionality… что делать?



Отбор признаков:

Задача: найти и удалить вредные признаки.

Какие признаки для нас вредные?

- Перебрать все варианты и посмотреть качество (лучший, если признаков мало)
- Посчитать корреляцию с целевой функцией и удалить шумные
- Посчитать корреляцию всех пар признаков и удалить скоррелированные
- Последовательно удалять худшие
- Последовательно добавлять лучшие

Плюсы и минусы метрических алгоритмов

Алгоритм (метрический):

- Наглядный, понятный
- Идеально работает, если правильно выбрана метрика
- Ленивый алгоритм, совсем не учится
- Позволяет делать беспризнаковое распознавание (например, распознавание личной подписи человека)
- На признаковом распознавании, как правило, работает хуже других алгоритмов

Сложность обучения — (запоминаем выборку)

Сложность предсказания — (считаем все расстояния)

В таком виде это в real time системах это работать не будет (даже если это алгоритм одного ближайшего соседа)!

А если kNN – k ближайших соседей….

Метрические алгоритмы – применяют ли их на практике?

Применяют!

Все большие поисковые/рекомендательные системы состоят из двух компонент:

- Начальный (первый) этап-Грубый отбор Объектов - Кандидатов
- Второй –финальный этап- Использование финальной модели

Быстрый приближенный поиск ближайших соседей идеально подходит под задачу выборов Объектов – Кандидатов

Найденные по итогам грубого отбора расстояния передаются в финальную модель для более тонких методов обработки и анализа

Приближенный поиск ближайших соседей – грубый отбор кандидатов – существует несколько методов, в том числе применимых в BIG DATA

Линейный дискриминантный анализ: для чего он?

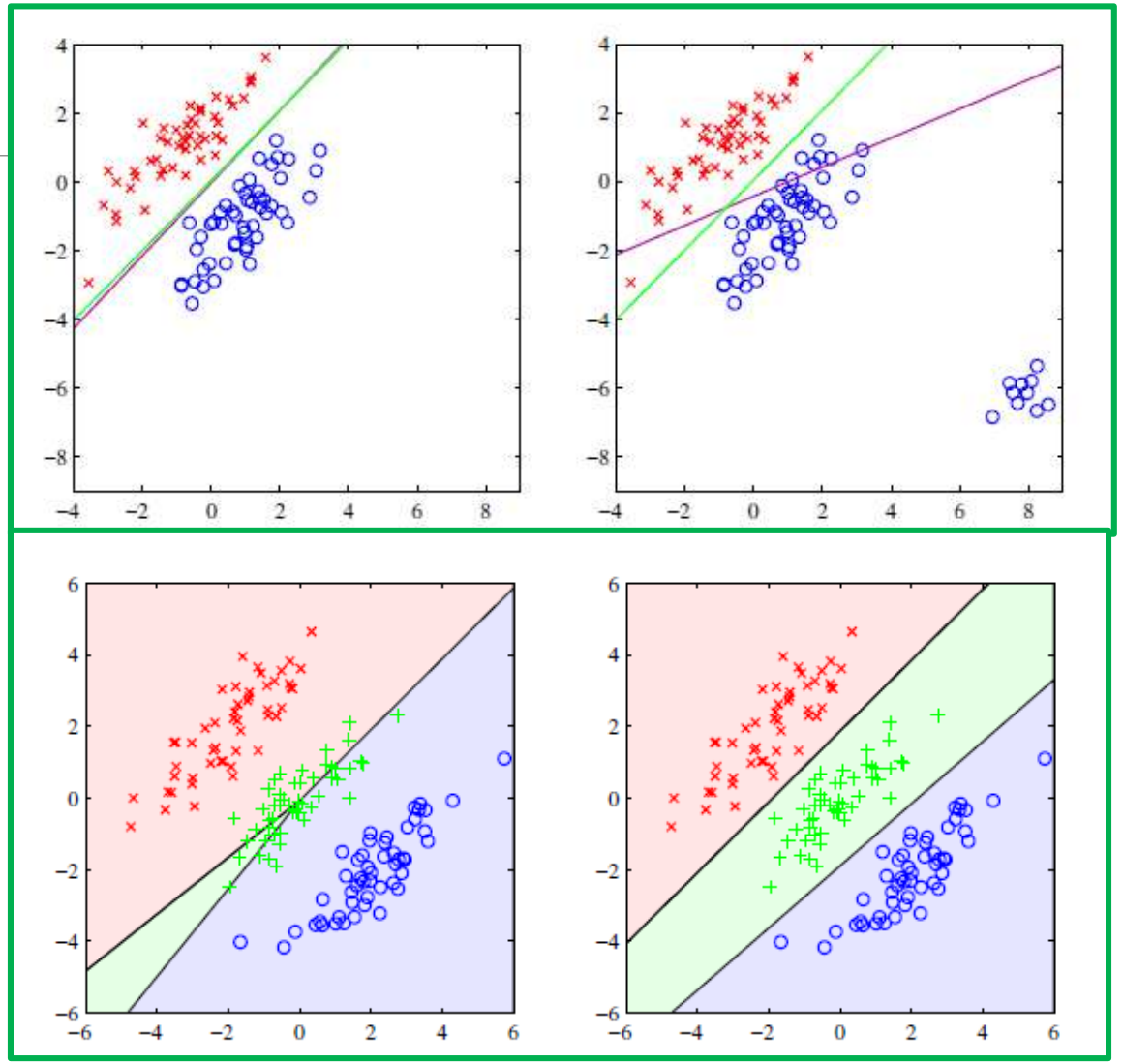
Есть логистическая регрессия! Но:

Когда классы хорошо разделены, оценки параметров логистической регрессии на удивление очень нестабильны. Линейный дискриминантный анализ не страдает от этой проблемы.

При малом объеме выборки и примерно нормально распределенных предикторах во всех классах линейный дискриминантный анализ, опять-таки, более устойчив, чем логистическая регрессионная модель.

Как было отмечено, линейный дискриминантный анализ популярен в ситуациях, когда зависимая переменная имеет больше двух классов. Геометрическая интерпретация проще!

Нельзя при решении задачи классификации действовать формально по аналогии с регрессией по методу наименьших квадратов – за большие отклонения от линии регрессии давать большой штраф



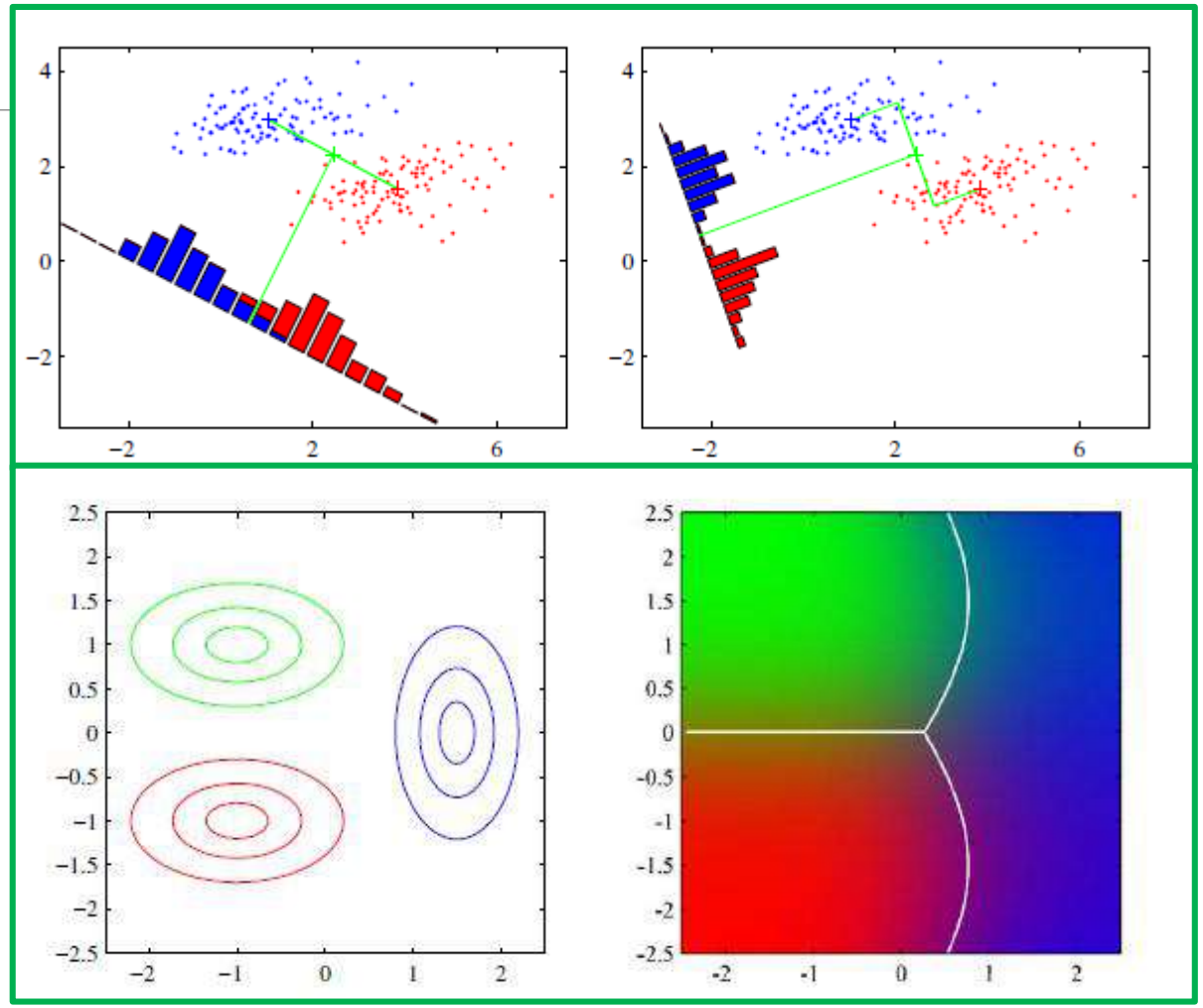
Линейный дискриминантный анализ:
для чего он?

Надо стремиться к тому, чтобы разделяющая поверхность (гиперплоскость) максимально правильно разделяла обучающую выборку

Как должен выглядеть ее направляющий вектор?

Нельзя чтобы гиперплоскость была параллельна прямой, соединяющей ЦЕНТРЫ множеств разных классов!

Три подмножества имеют одинаковую общую дисперсию, но разные ковариации, как тогда проходят разделяющие поверхности?



Классификация – дискриминантный анализ:

Байесовский классификатор (x –многомерные объекты, y – принадлежность классу)

$$h(x) = \arg\max_{y \in Y} p(y \mid x).$$

Одним из распространенных методов решения задач классификации является так называемый байесовский подход, при котором максимизируется апостериорная вероятность класса. В этом случае решающее правило имеет вид:

$$h(x) = \arg\max_{y \in Y} p(y \mid x) = \arg\max_{y \in Y} \frac{p(x \mid y)p(y)}{p(x)} = \arg\max_{y \in Y} p(x \mid y)p(y).$$

По теореме Байеса решающее правило можно переписать в виде (знаменатель нас не интересует)

Пусть объект x описывается n признаковыми функциями $f_1(x); : : : ; f_n(x)$, значения которых равны $x_1; : : : ; x_n$ (мы отождествляем объект и его признаковое описание), предположим, что значения данных функций являются независимыми случайными величинами. В этом случае мы приходим к решающему правилу:

$$h(x) = \arg\max_{y \in Y} \prod_{i=1}^{n} p(f_i(x) = x_i \mid y)p(y).$$

Данный классификатор называется **наивным** байесовским, для его обучения требуется оценить априорные вероятности классов $p(y)$ и условные плотности $p(x_i \mid y)!$

Классификация – дискриминантный анализ:

Байесовский классификатор (*x* –многомерные объекты, *y* – принадлежность классу)

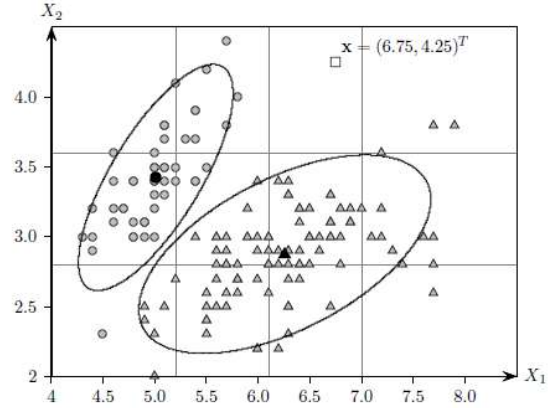
$$\hat{y} = \arg\max_i \{P(c_i|\mathbf{x})\} \qquad P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i)\cdot P(c_i)}{P(\mathbf{x})} \qquad P(\mathbf{x}) = \sum_{j=1}^{k} P(\mathbf{x}|c_j)\cdot P(c_j)$$

$$\hat{y} = \arg\max_i \{P(c_i|\mathbf{x})\}$$

$$= \arg\max_i \left\{\frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})}\right\} = \arg\max_i \{P(\mathbf{x}|c_i)P(c_i)\} \qquad \hat{P}(c_i) = \frac{n_i}{n}$$

$$f_i(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(\sqrt{2\pi})^d\sqrt{|\boldsymbol{\Sigma}_i|}}\exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right\} \qquad P(\mathbf{x}|c_i) = 2\epsilon\cdot f_i(\mathbf{x})$$

$$P(c_i|\mathbf{x}) = \frac{2\epsilon\cdot f_i(\mathbf{x})P(c_i)}{\sum_{i=1}^{k} 2\epsilon\cdot f_i(\mathbf{x})P(c_i)} = \frac{f_i(\mathbf{x})P(c_i)}{\sum_{i=1}^{k} f_i(\mathbf{x})P(c_i)} \qquad \hat{y} = \arg\max_i \left\{f_i(\mathbf{x})P(c_i)\right\} \qquad \hat{\boldsymbol{\mu}}_i = \frac{1}{n_i}\sum_{\mathbf{x}_j\in\mathbf{D}_i}\mathbf{x}_j \qquad \hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\mathbf{Z}_i^T\mathbf{Z}_i$$



По теореме Байеса решающее правило можно переписать в виде (знаменатель нас не интересует)

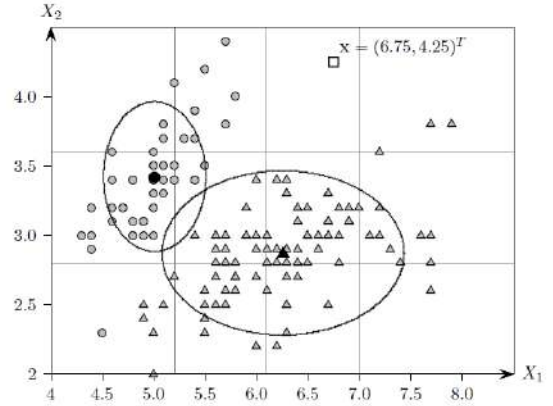
Пусть объект x описывается n признаковыми функциями $f_1(x); : : : ; f_n(x)$, значения которых равны $x_1; : : : ; x_n$ (мы отождествляем объект и его признаковое описание), предположим, что значения данных функций являются независимыми случайными величинами. В этом случае мы приходим к решающему правилу:

$$\hat{P}(c_1|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)\hat{P}(c_1) = (4.914\times10^{-7})\times 0.33 = 1.622\times10^{-7}$$
$$\hat{P}(c_2|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2)\hat{P}(c_2) = (2.589\times10^{-5})\times 0.67 = 1.735\times10^{-5}$$

Классификация – дискриминантный анализ:

Байесовский классификатор (x –многомерные объекты, y – принадлежность классу) Наивный Байесовский классификатор

$$\hat{y} = \arg\max_i \{P(c_i|\mathbf{x})\} \qquad P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i)\cdot P(c_i)}{P(\mathbf{x})} \qquad P(\mathbf{x}) = \sum_{j=1}^{k} P(\mathbf{x}|c_j)\cdot P(c_j)$$



$$\hat{y} = \arg\max_i \{P(c_i|\mathbf{x})\} = \arg\max_i \left\{\frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})}\right\} = \arg\max_i \{P(\mathbf{x}|c_i)P(c_i)\}$$

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{i1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{i2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_{id}^2 \end{pmatrix}$$

$$P(\mathbf{x}|c_i) = P(x_1, x_2, \cdots, x_d|c_i) = \prod_{j=1}^{d} P(x_j|c_i)$$

$$P(x_j|c_i) \propto f(x_j|\mu_{ij}, \sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

$$|\boldsymbol{\Sigma}_i| = \det(\boldsymbol{\Sigma}_i) = \sigma_{i1}^2\sigma_{i2}^2\cdots\sigma_{id}^2 = \prod_{j=1}^{d}\sigma_{ij}^2$$

$$\boldsymbol{\Sigma}_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{i1}^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{i2}^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \frac{1}{\sigma_{id}^2} \end{pmatrix}$$

$$(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = \sum_{j=1}^{d}\frac{(x_j-\mu_{ij})^2}{\sigma_{ij}^2}$$

$$P(\mathbf{x}|c_i) = \frac{1}{(\sqrt{2\pi})^d\sqrt{\prod_{j=1}^{d}\sigma_{ij}^2}}\exp\left\{-\sum_{j=1}^{d}\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

$$= \prod_{j=1}^{d}\left(\frac{1}{\sqrt{2\pi}\,\sigma_{ij}}\exp\left\{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}\right)$$

$$= \prod_{j=1}^{d} P(x_j|c_i)$$

Да
ба
ou
условные плотности $p(x_i \mid y)$!

Классификация – дискриминантный анализ

$$h(x) = \arg\max_{y \in Y} \prod_{i=1}^{n} p(f_i(x) = x_i \mid y) p(y).$$

Априорные вероятности классов можно оценить их частотами в обучающей выборке.

Но приходится решать задачу восстановления плотностей при условиях принадлежности объектов разным классам, а это более сложная задача, чем сама классификация - избыточная задача!
Восстановление n одномерных плотностей - намного более простая задача, чем одной n-мерной.

На практике делают предположения о принадлежности многомерных распределений в каждом классе некоторому параметрическому семейству, например, нормальному (гауссову) многомерному распределению

Тогда нужно определить параметры этого распределения в каждом классе по имеющимся подвыборкам объектов
Опасности из-за дисбаланса представленности классов в выборке!

Дискриминантный анализ – как его дифференцировать

Может быть параметрическим или непараметрическим

Параметрический более не менее приемлем только тогда, когда внутриклассовые распределения близки к нормальному (гауссовому). В этом случае, если внутриклассовые дисперсии равны – получаем линейный дискриминантный анализ. Если внутриклассовые дисперсии не равны – получаем квадратичный дискриминантный анализ

Непараметрический – свободен от ограничений – требования нормальности внутриклассовых распределений. В этой ситуации используются ядерные методы, либо уже известные нам метрические методы, в частности, метод n ближайших соседей.

Дискриминантный анализ -проблемы

При проведении ДА возникают проблемы, аналогичные линейной регрессии:

Мультиколлинеарность

Наличие бесполезных переменных, мешающих анализу

Переобучение

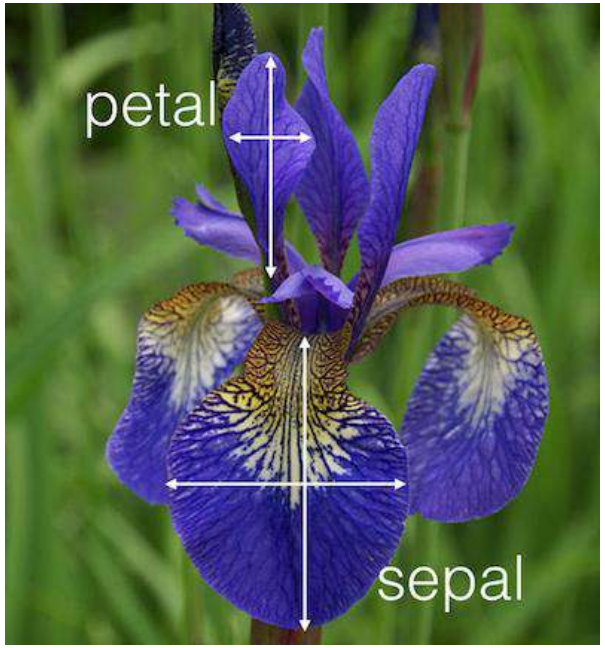
В условиях ограниченности выборки – дисбаланс количества объектов в классах

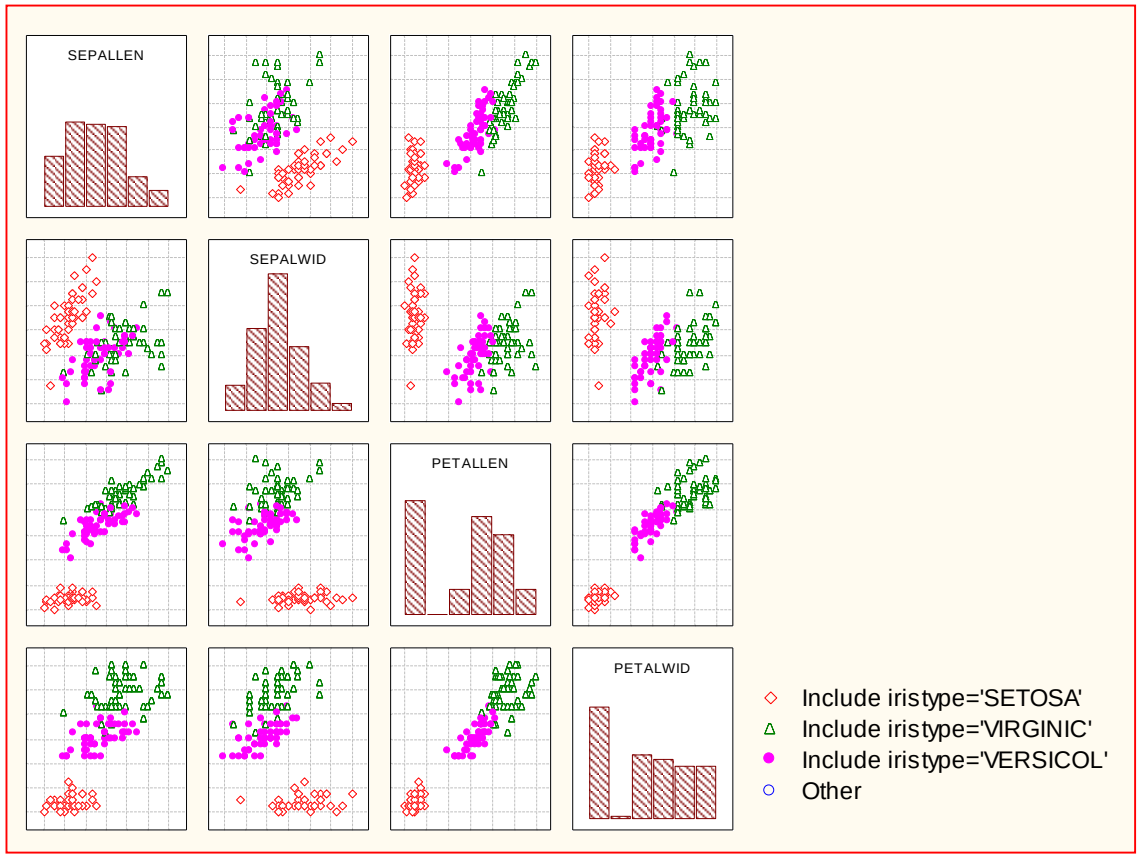
Как с этим бороться?

Пошаговый дискриминантный анализ – исключения и включения предикторов, переборы

Преобразование координат – линейное преобразование переменных – канонический дискриминантный анализ

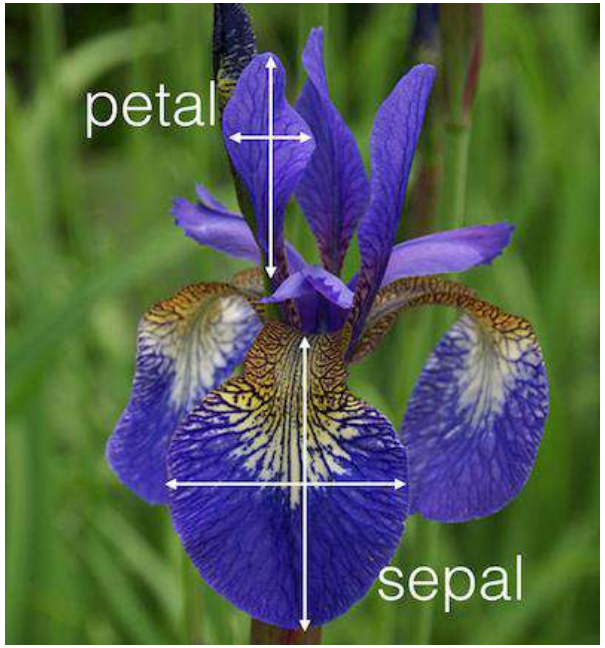
Классификация – дискриминантный анализ
Ирис Фишера – эталон массива данных для дискриминантного анализа

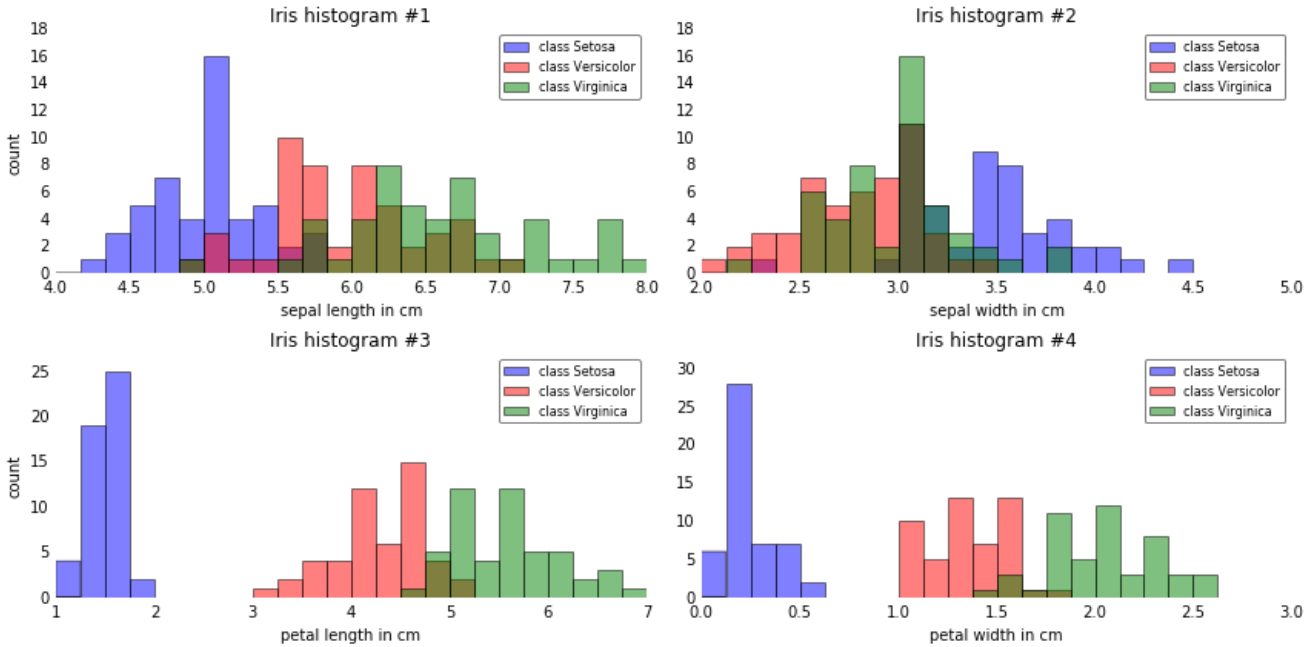




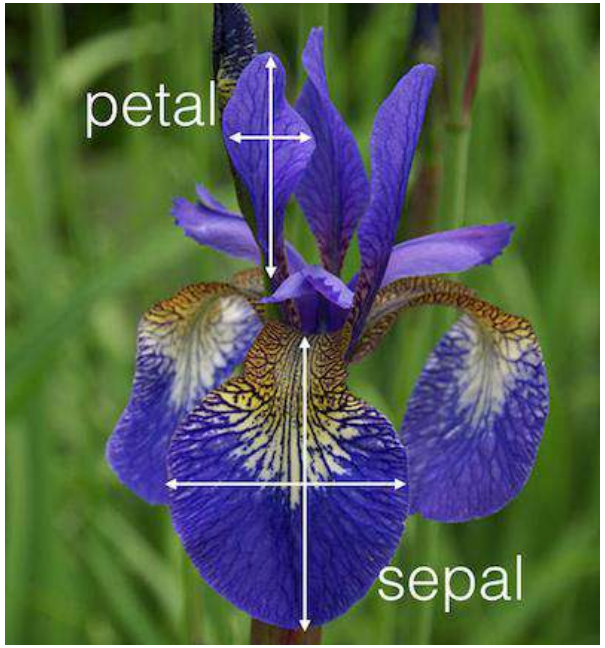
На матричной диаграмме рассеяния видно, что по НЕКОТОРЫМ переменным класс SETOSA отделим уже без привлечения остальных переменных.
По классам VIRGINIC , VERSICOL – разделение значительно труднее!

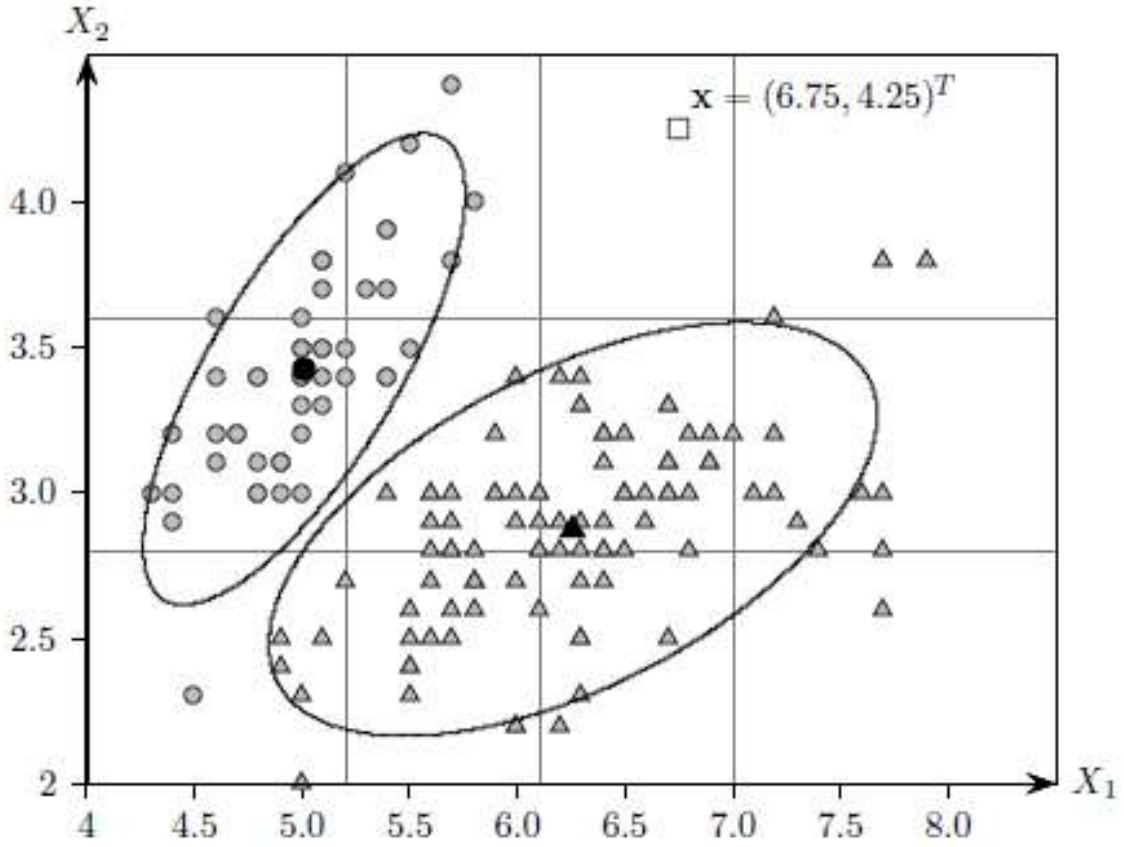
Классификация – дискриминантный анализ
Ирис Фишера – эталон массива данных для дискриминантного анализа



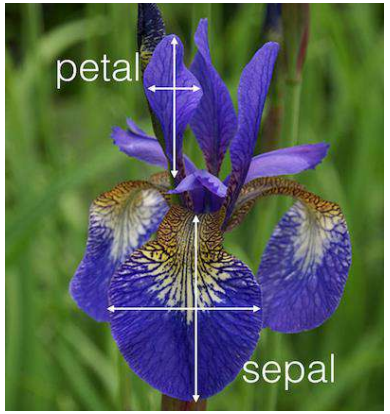


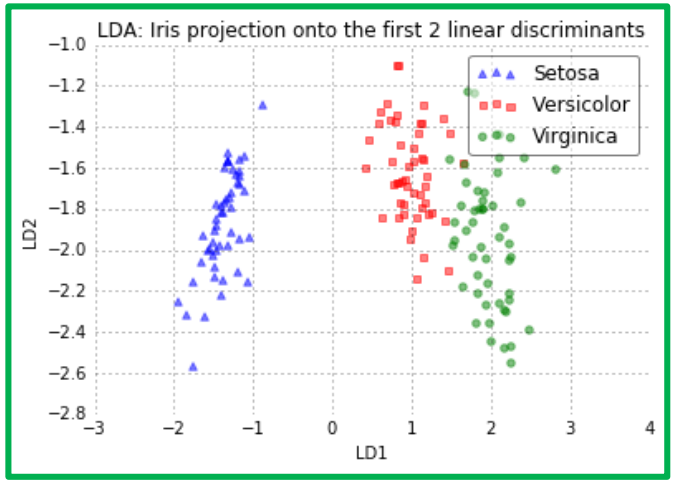
Классификация – дискриминантный анализ
Ирис Фишера – эталон массива данных для дискриминантного анализа

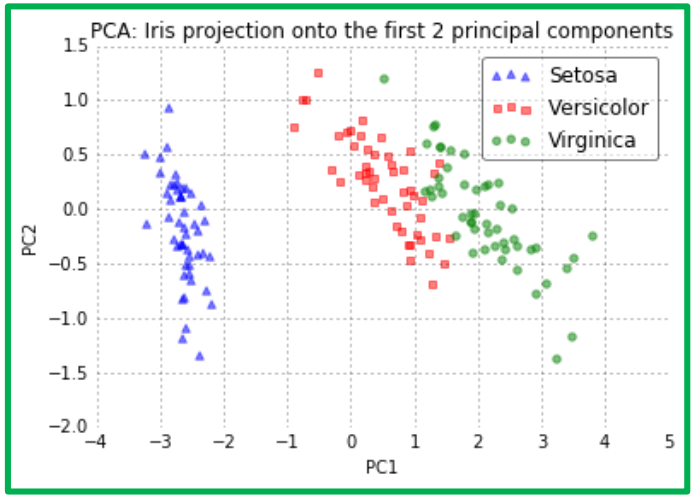


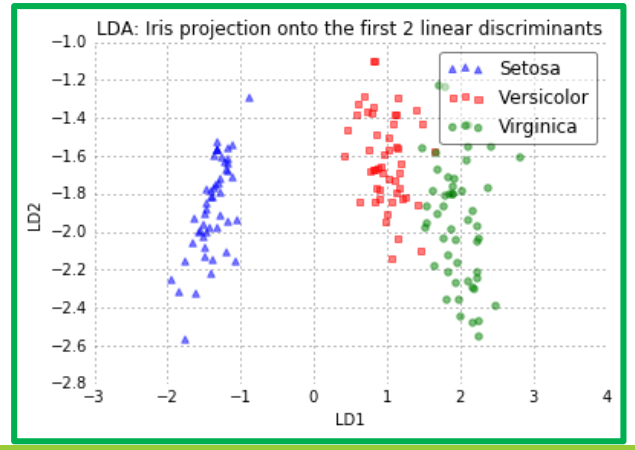


Классификация – дискриминантный анализ





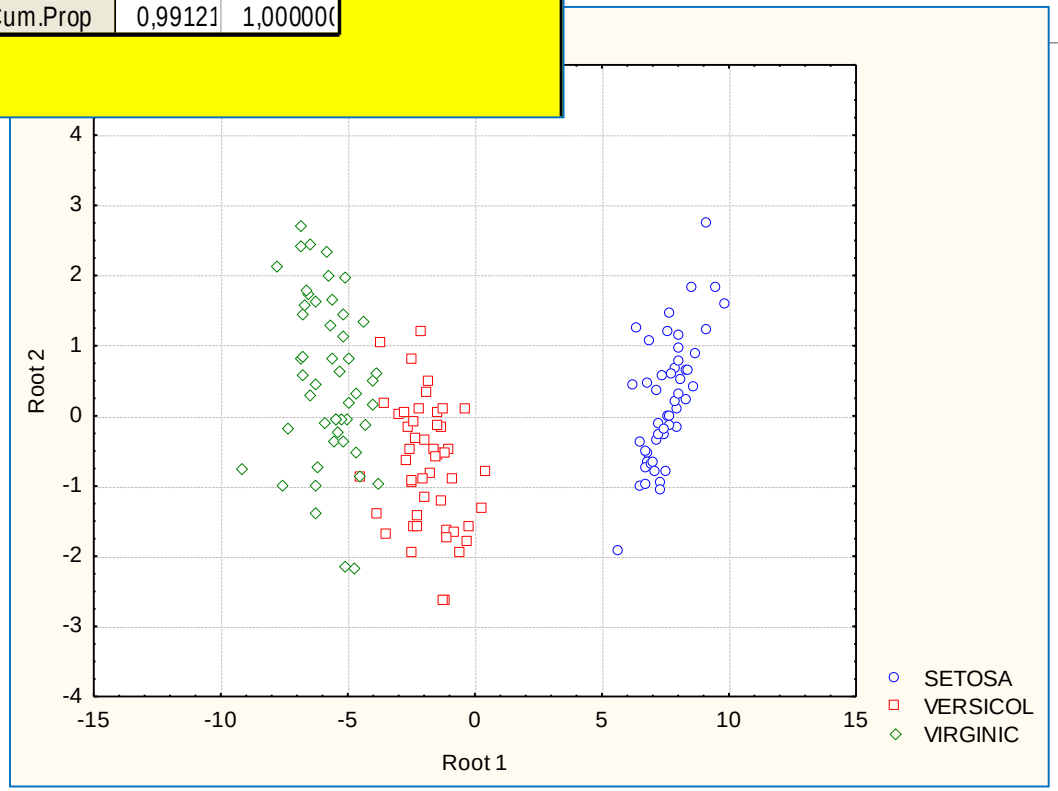


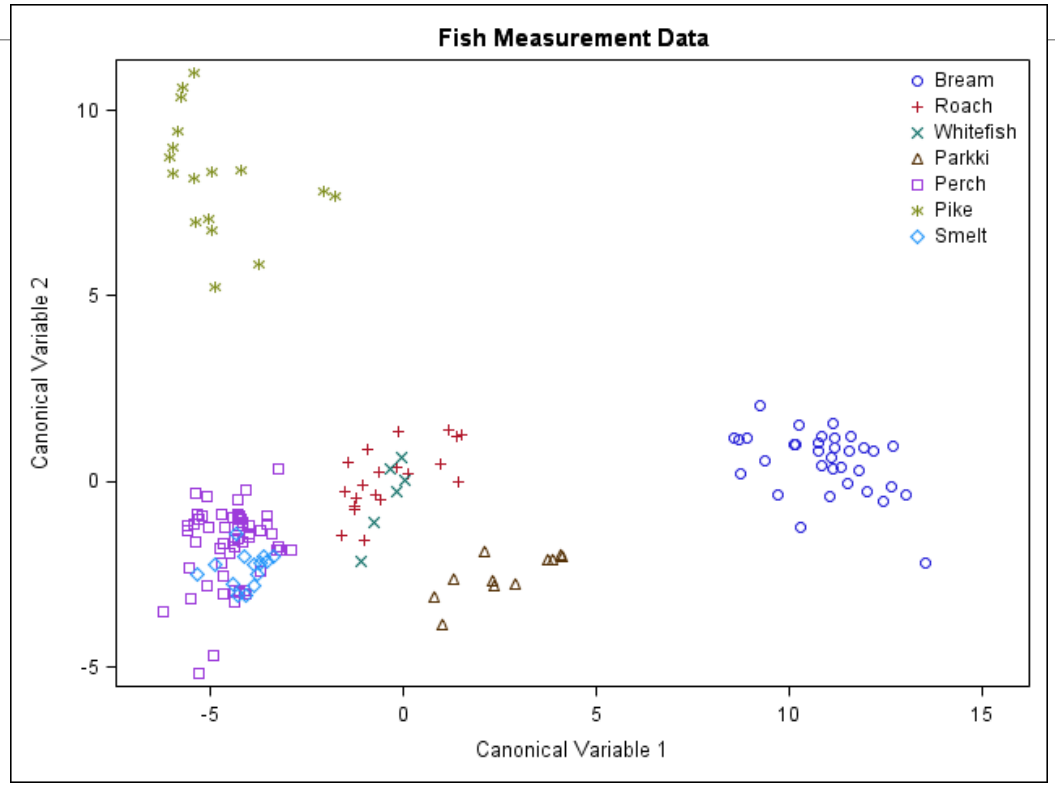


Поворот системы координат – линейные преобразование переменных
Может быть преобразование – методом главных компонент – наилучшим образом объяснить ОБЩУЮ дисперсию
Может быть преобразование – методом канонического дискриминантного анализа – наилучшим образом объяснить МЕЖГРУППОВУЮ дисперсию

Variable	Standardized Coefficients (Irisdat) for Canonical Variables	
	Root 1	Root 2
SEPALLEN	0,42695	0,012408
SEPALWID	0,52124	0,735261
PETALLEN	-0,9472(	-0,401038
PETALWID	-0,5751(	0,581040
Eigenval	32,19193	0,285391
Cum.Prop	0,99121	1,000000

Дискриминантный анализ (Discriminant analysis)





Канонический дискриминантный анализ: уже первые несколько новых (канонических) переменных хорошо объясняют максимальную долю дисперсии – дисперсии МЕЖКЛАССОВОЙ!!! Всего же канонических переменных – минимум из: числа классов минус 1 и числа исходных переменных

Спасибо за внимание!

Лекция -окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

ТЕЛ. (484)3974658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

Преобразование переменных

•Исходные переменные далеко не всегда подходят для выполнения анализа
•Различные по своей сложности и эффективности методы и приемы преобразований

•Простейшие преобразования – преобразования отдельных переменных, стандартизация, нормализация, логарифмирование, разбиения на градации и т.д.

•Уменьшения размерности за счет пошаговых процедур – есть и в регрессионном анализе и в классификации

•Более сложные – повороты исходного пространства переменных – линейные преобразования
•Более сложные – кластеризация переменных (не наблюдений, а переменных!!!!)

•Увеличение размерности – умышленное – иногда полезно, например, для разделения линейно неразделяемых подмножеств

•Уменьшение размерности за счет линейных преобразований - в первую очередь – метод главных компонент

Анализ главных компонент
(Principal component analysis) –для чего он нужен?

•Когда много переменных, мы не обладаем априорными сведениями о том, какие из них важны
•В задачах анализа многомерных данных трудно выработать какие-то предположения о механизмах, порождающих такие данные: могут быть сложные механизмы, а могут быть простые, но замаскированные высокой размерностью
•Некоторые входные переменные (переменные – предикторы) в регрессионном анализе могут быть высоко коррелированы или просто являться линейными комбинациями других
•При этом возникает мультиколлинеарность – одна из основных «бед» регрессионного анализа
•В Разведке Данных DATA MINING очень часто невозможно разобраться в исходном большом количестве исходных переменных, особенно когда имеем дело с БОЛЬШИМИ ДАННЫМИ, и связи между переменными неясны
•Речь идет о том, чтобы найти новые переменные, являющиеся линейными комбинациями исходных переменных, и дать новую интерпретацию им в терминах вариаций (дисперсий)
• **Забегая вперед:**
•Такой вариант нахождения новых переменных , перехода к ним от старых (исходных) переменных, - является типичным для стратегий DATA MINING
•Используется в DATA MINING для **МОДИФИКАЦИИ ПЕРЕМЕННЫХ**
•Используется в изощренных схемах машинного обучения как один из методов построения кодировщиков и декодировщиков

Анализ главных компонент (Principal component analysis)

•Для группировки объектов используется **кластерный** анализ

•Формально **кластерный** анализ может быть применен и к задаче группировки переменных

•Метод **главных компонент** (Principal Components Analysis Method) основан на других идеях

•Это метод, в отличие от многих других методов анализа, **имеющий строгое математическое обоснование**

•Иногда его называют **Методом Эмпирических Ортогональных Функций (ЭОФ) – Empirical Orthogonal Functions (EOF)**

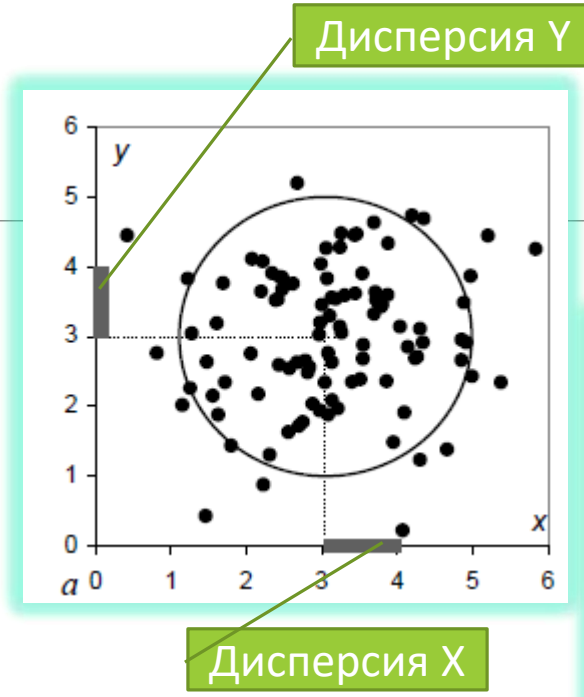
•Почему Эмпирических?

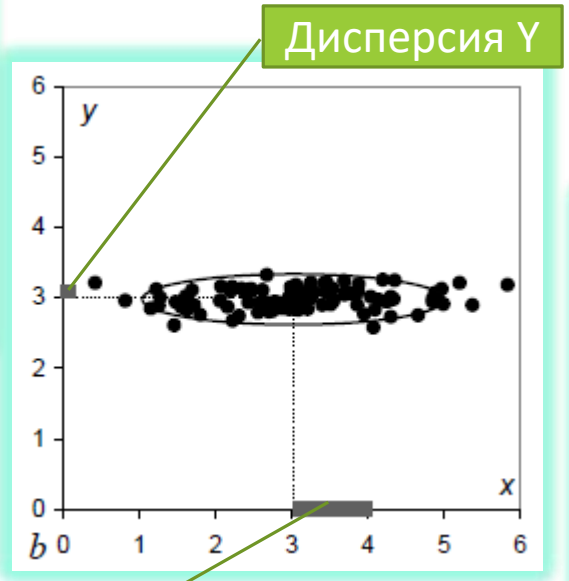
•Почему Ортогональных?

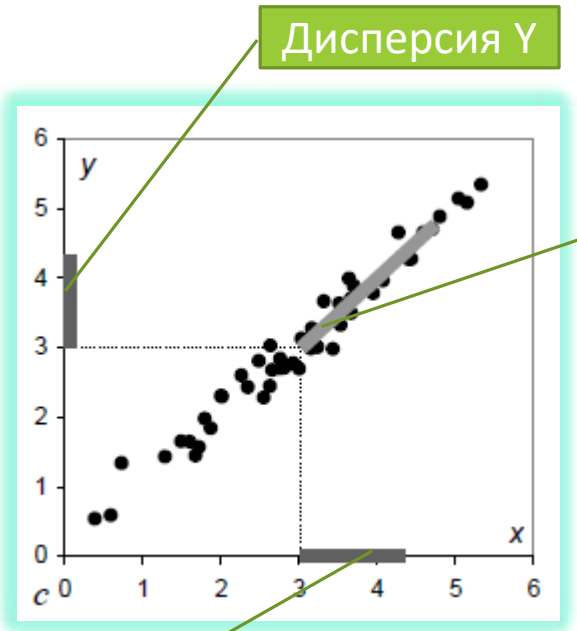
•Введен в 1901 г Пирсоном, развит в 1933 г Хотеллингом

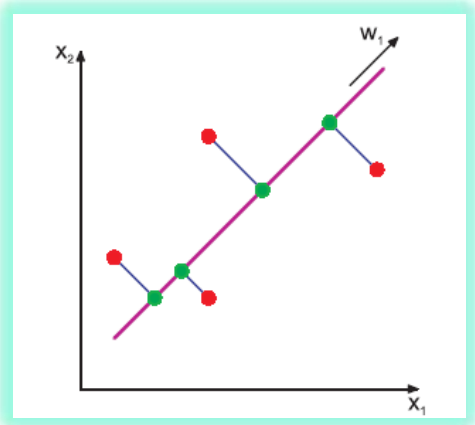
•Иногда в теории информации называют Преобразованием Карунена-Лоэва

Анализ главных компонент (Principal component analysis)



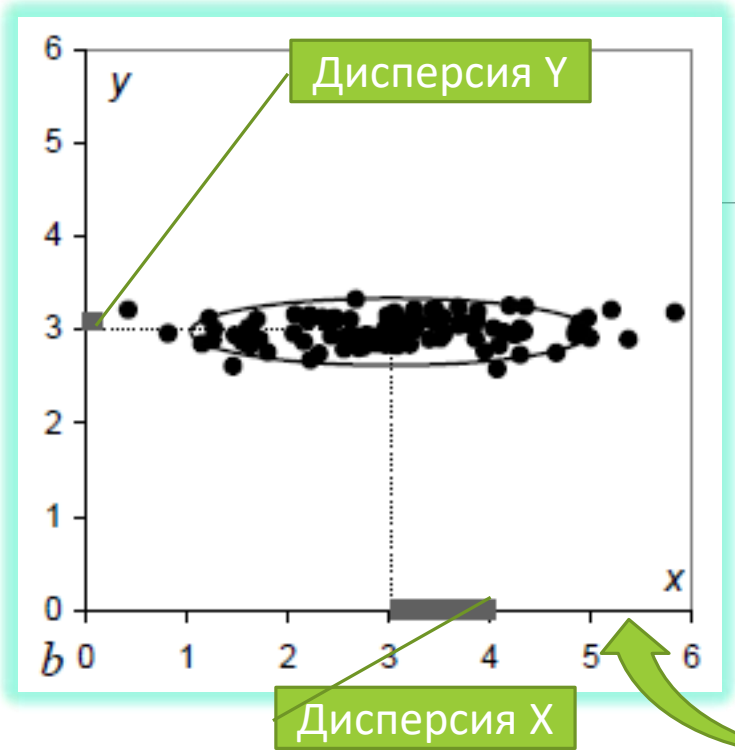


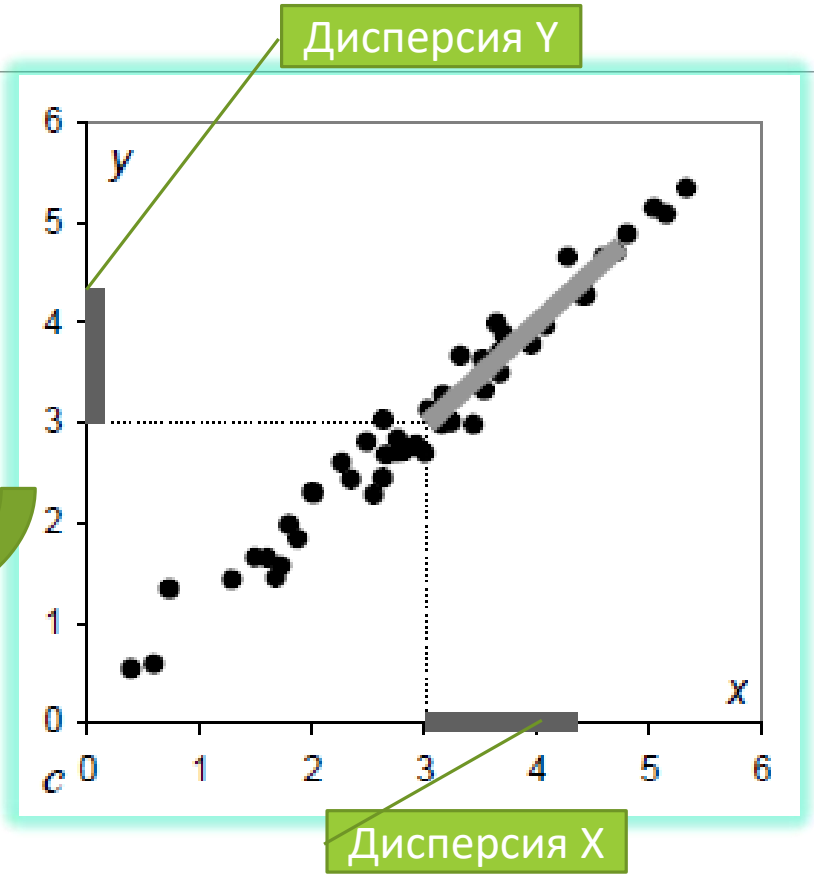




•Речь идет о повороте исходных переменных – выборе новых переменных, являющихся линейными комбинациями старых переменных

Анализ главных компонент (Principal component analysis)





Надо как-то повернуть оси, чтобы вторая картинка стала похожа на первую
Иначе говоря, сделать линейное преобразование исходных переменных –
В новых координатах будет выглядеть как на рис. 2

Анализ главных компонент (Principal component analysis)

•Речь идет о повороте исходных переменных – выборе новых переменных, являющихся линейными комбинациями старых переменных

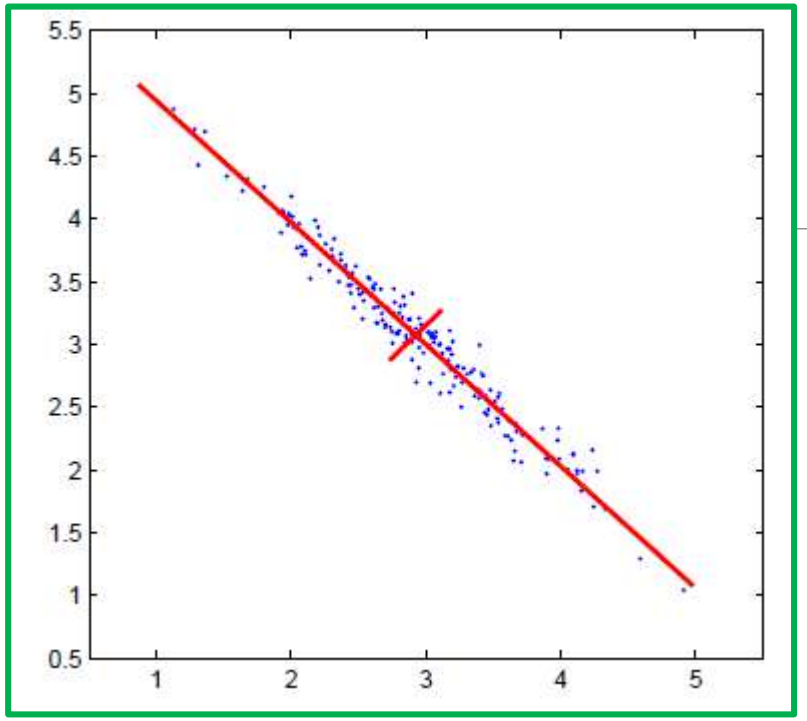
•Имея массив данных с p числовыми переменными, можно вычислить p главных компонент (PC)

•Каждая PC представляет собой линейную комбинацию исходных переменных , **с коэффициентами, равными собственным векторам (eigenvectors) корреляционной или ковариационной матрицы**

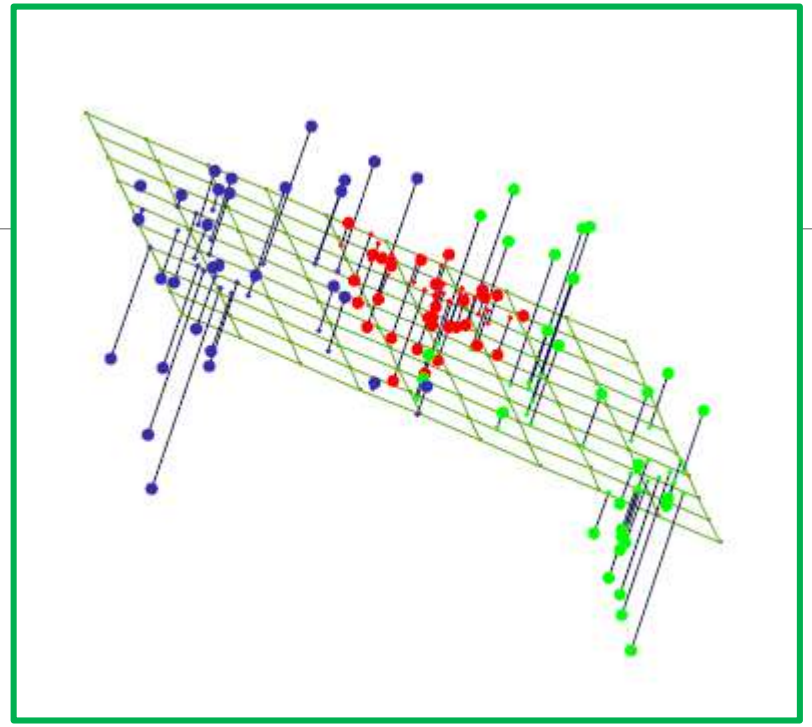
• Собственные векторы часто приводятся к единичной длине

•Главные компоненты упорядочиваются по убыванию собственных значений (eigenvalues) корреляционной или ковариационной матрицы, которые равны дисперсиям компонент.

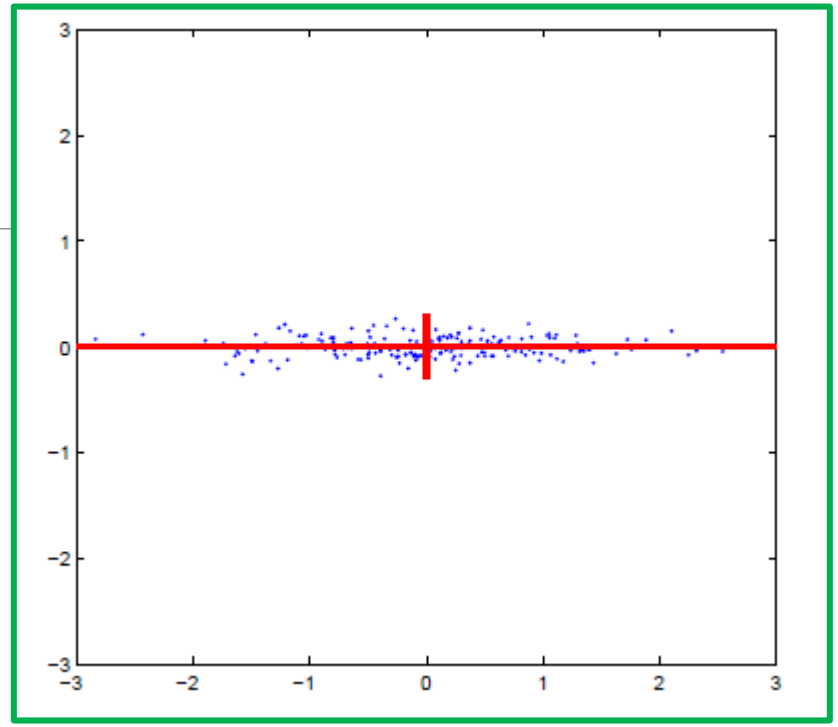
Анализ главных компонент (Principal component analysis)



Исходное пространство
переменные сильно коррелируют
Дисперсии старых переменных близки



Ищем гиперплоскость данной размерности, такую, чтобы ошибка проектирования всей выборки на эту гиперплоскость была бы минимальной



Преобразованное пространство – новые координаты
Сдвинули начало координат в центр выборки
Повернули оси так, что новые переменные не коррелированы
В новых координатах можем избавиться от переменных с малой дисперсией

Метод главных компонент Principal Component Analysis) – немного математики

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Элементы ковариационной матрицы S

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{(n-1)s_i s_j} = \frac{s_{ij}}{s_i s_j}$$

Элементы корреляционной матрицы

$$\mathbf{x}_c = \mathbf{x} - \bar{\mathbf{x}}$$

Центрирование вектора X

$$z_i = \mathbf{u}_i'(\mathbf{x} - \bar{\mathbf{x}})$$

i-я главная компонента – линейная комбинация центрированных координат вектора X

$$(\mathbf{S} - \lambda_i \mathbf{I})\,\mathbf{u}_i = 0$$

i-я главная компонента определяется из этого условия , **I** – единичная матрица размерности d, где d – количество исходных числовых переменных

Метод главных компонент
(Principal Component Analysis)

$$|\mathbf{S} - \lambda\,\mathbf{I}| = 0$$

Требуется решать детерминантное уравнение, имеется d скаляров λi , являющихся решениями этого уравнения. Эти скаляры называются собственными значениями (eigenvalues) или характеристическими значениями (characteristic values) матрицы S, они представляют дисперсии для новых переменных Zi. Решив систему для разных собственных значений λi , получаем семейство собственных векторов (eigenvectors) или характеристических векторов (characteristic vectors) **u**i.

$$\forall\ i, j\ \ \mathbf{u}_i'\mathbf{u}_j = 0$$

Это ортогональная система некоррелированных векторов

Метод главных компонент
(Principal Component Analysis)
Свойства Главных Компонент:

•Собственные векторы взаимно ортогональны, поэтому главные компоненты представляют собой взаимно перпендикулярные направления в пространстве исходных переменных

• Новые переменные – главные компоненты взаимно не коррелированы, их значения - метки (scores) главных компонент -

• Первая главная компонента имеет наибольшую дисперсию среди всех возможных линейных комбинаций единичной длины, составленных из исходных переменных.

• j –я главная компонента имеет наибольшую дисперсию всех возможных линейных комбинаций единичной длины, ортогональных к первым j -1 главным компонентам

•Последняя, p - я главная компонента имеет наименьшую дисперсию из всех возможных линейных комбинаций исходных переменных

Метод главных компонент
(Principal Component Analysis)

$$|\mathbf{S} - \lambda\,\mathbf{I}| = 0$$

Требуется решать детерминантное уравнение, имеется d скаляров λi , являющихся решениями этого уравнения. Эти скаляры называются собственными значениями (eigenvalues) или характеристическими значениями (characteristic values) матрицы S, они представляют дисперсии для новых переменных Zi. Решив систему для разных собственных значений λi , получаем семейство собственных векторов (eigenvectors) или характеристических векторов (characteristic vectors) **u**i.

$$\forall\ i, j\ \ \mathbf{u}_i'\mathbf{u}_j = 0$$

Это ортогональная система некоррелированных векторов

Метод главных компонент
(Principal Component Analysis)
Свойства Главных Компонент:

•Собственные векторы взаимно ортогональны, поэтому главные компоненты представляют собой взаимно перпендикулярные направления в пространстве исходных переменных

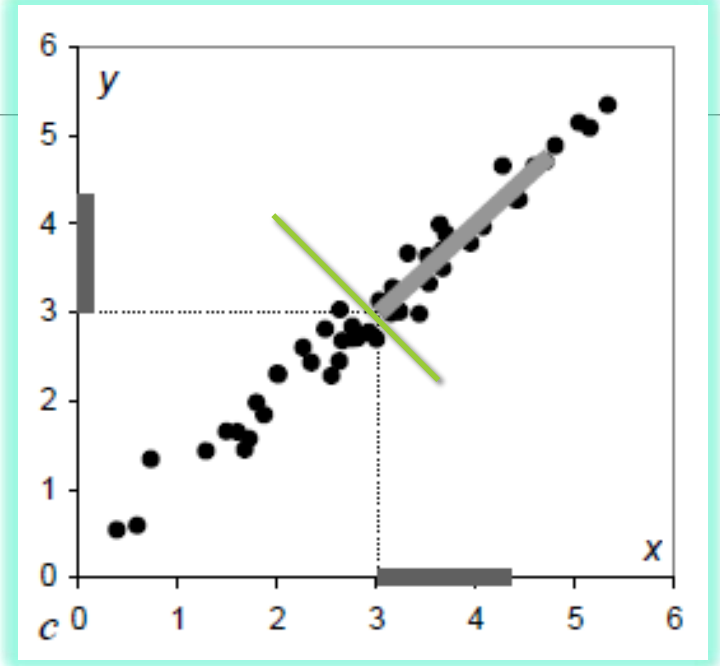
• Новые переменные – главные компоненты взаимно не коррелированы, их значения - метки (scores) главных компонент -

• Первая главная компонента имеет наибольшую дисперсию среди всех возможных линейных комбинаций единичной длины, составленных из исходных переменных.

• j –я главная компонента имеет наибольшую дисперсию всех возможных линейных комбинаций единичной длины, ортогональных к первым j -1 главным компонентам

•Последняя, p - я главная компонента имеет наименьшую дисперсию из всех возможных линейных комбинаций исходных переменных

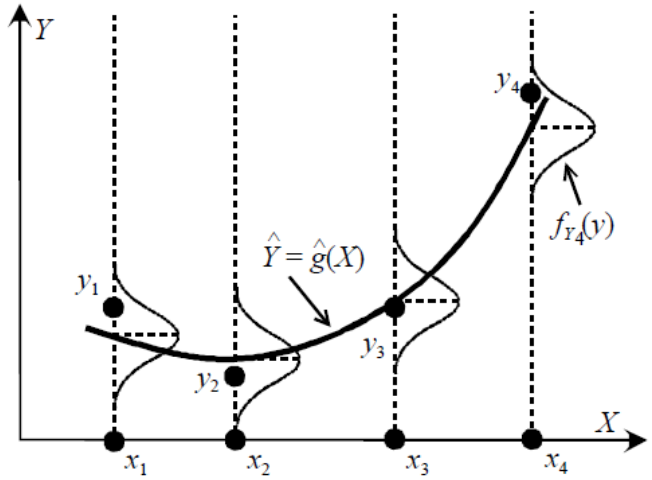
Метод главных компонент
(Principal Component Analysis)
Свойства Главных Компонент:

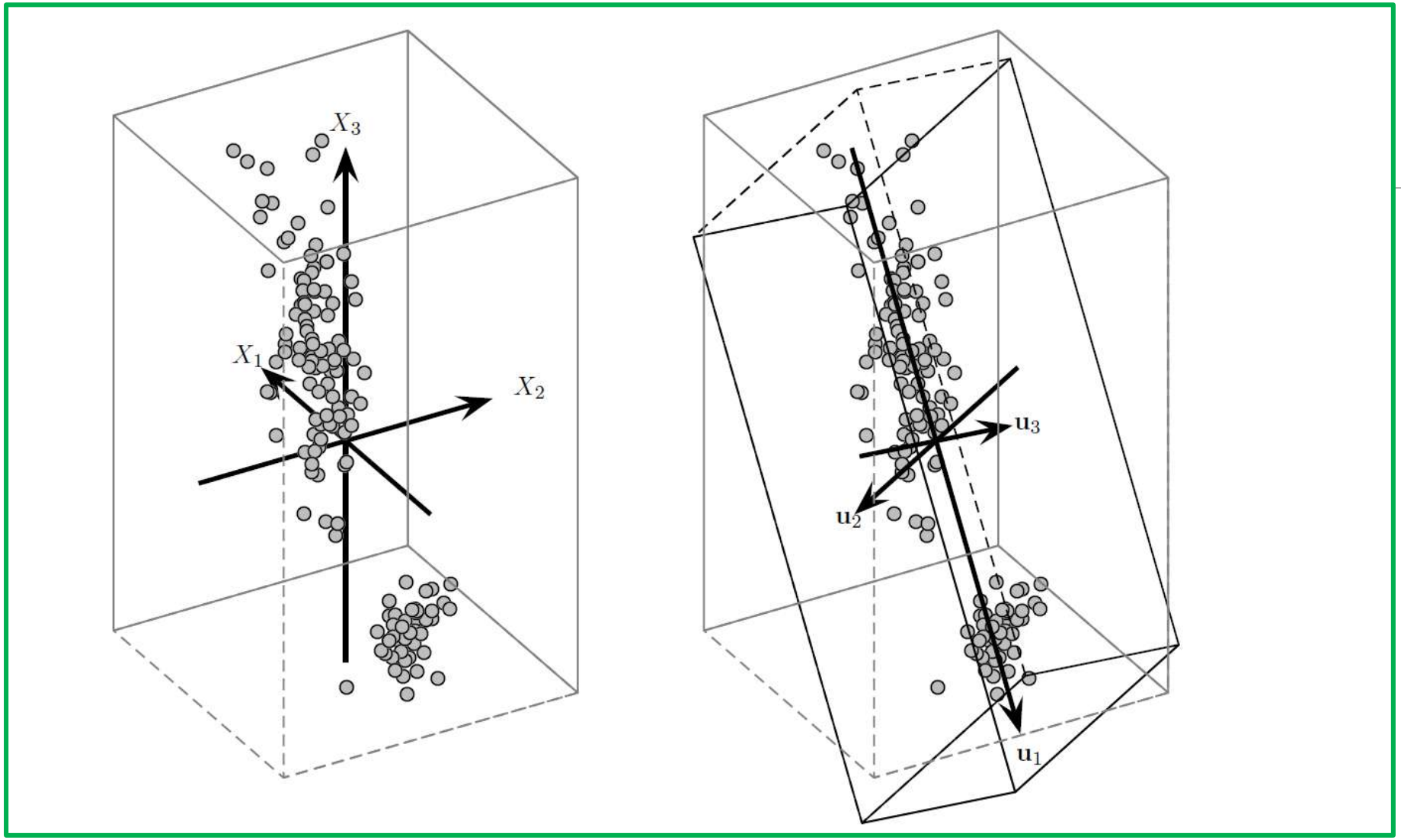


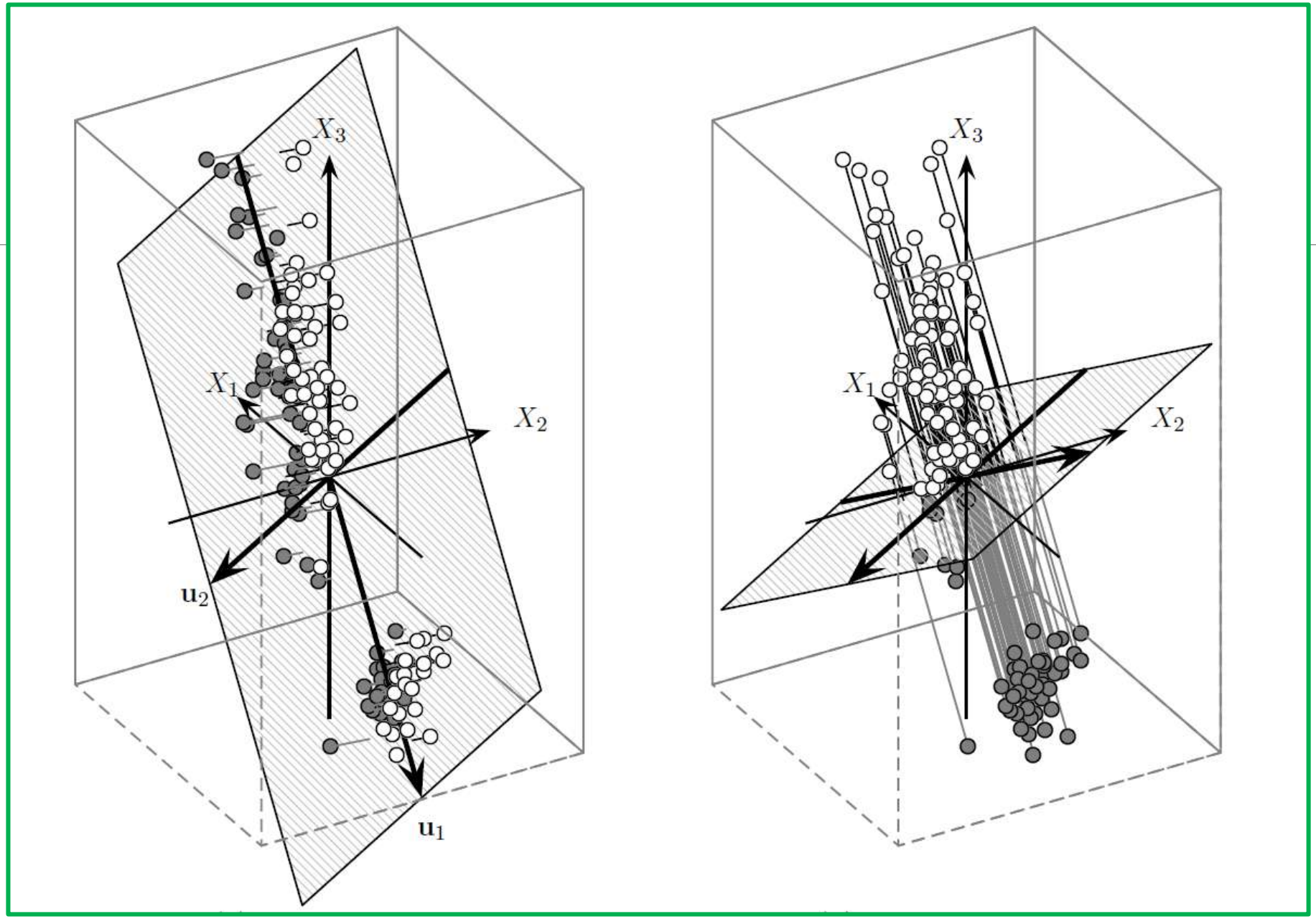
ОТЛИЧИЕ ОТ РЕГРЕССИИ:
В регрессионном анализе переменные играют разные роли (предикторы и предиктант), и минимизируется

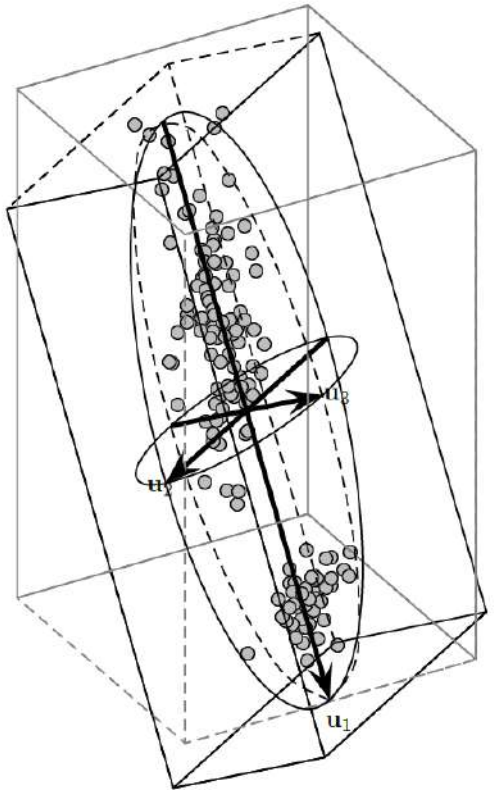
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

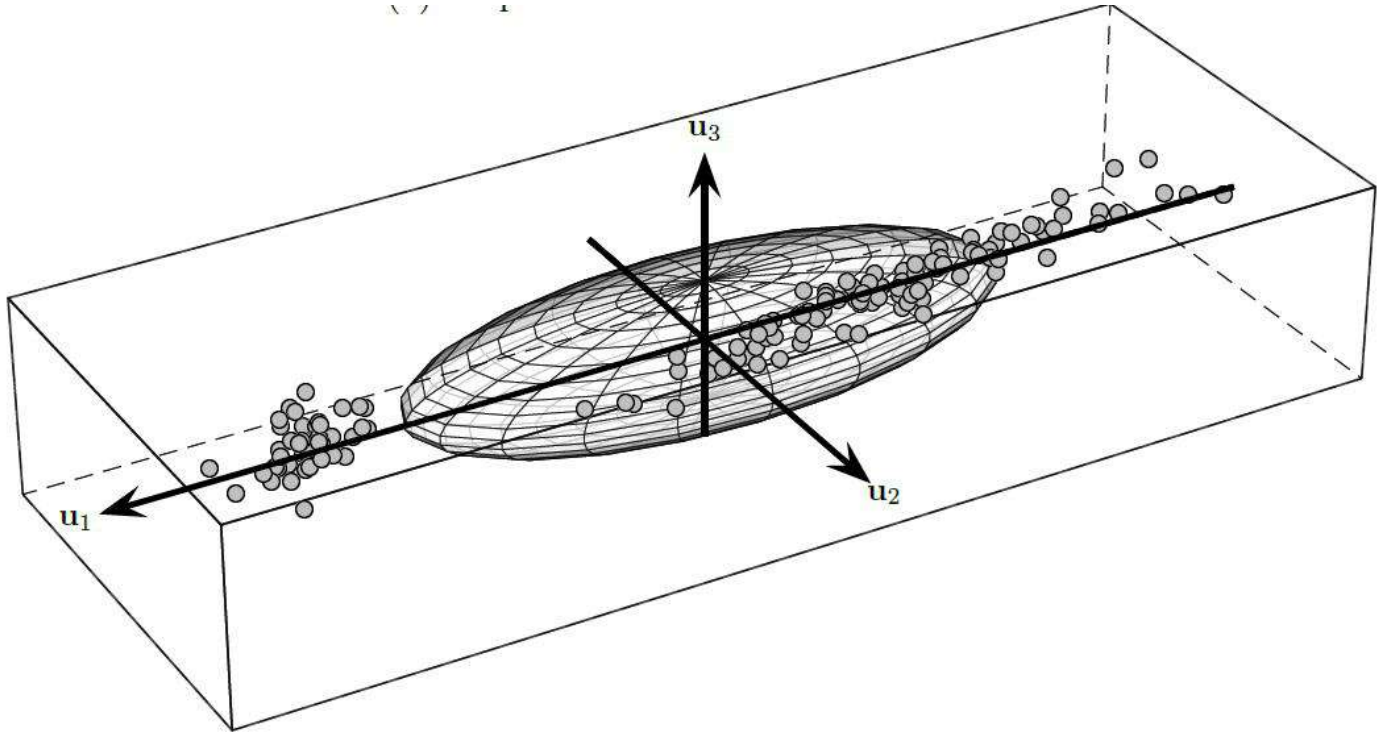
В методе главных компонент роли исходных переменных одинаковы, надо найти направление, для которого сумма квадратов опущенных из исходных точек перпендикуляров **минимальна, Такое направление «объясняет» максимум исходной дисперсии**




X_3
X_1
X_2
$\mathbf{u}_3$
$\mathbf{u}_2$
$\mathbf{u}_1$


X_3
X_1
X_2
$\mathbf{u}_2$
$\mathbf{u}_1$
X_3
X_1
X_2


u3
u2
u1


u3
u1
u2

Метод главных компонент – пример применения
(Principal Component Analysis)

- Пример: анализ криминальной обстановки в 50 штатах США по 7 переменным
- 7 переменных – число преступлений на 100 тысяч населения штата:
  - Murder - убийство
  - Rape - изнасилование
  - Robbery – ограбление (с применением насилия), разбой
  - Assault – разбойное нападение
  - Burglary – квартирная кража со взломом
  - Larceny – похищение имущества, кража
  - AutoTheft – кража авто

Метод главных компонент
(Principal Component Analysis)

- Пример: анализ криминальной обстановки в 50 штатах США (50 наблюдений) по 7 переменным
- 7 переменных – число преступлений на 100 тысяч населения штата
- По корреляционной матрице (размерностью 7 на 7) трудно делать выводы

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0							
	Murder	**Rape**	**Robbery**	**Assault**	**Burglary**	**Larceny**	**Auto_Theft**
Murder	1.00000	0.60122 <.0001	0.48371 0.0004	0.64855 <.0001	0.38582 0.0057	0.10192 0.4813	0.06881 0.6349
Rape	0.60122 <.0001	1.00000	0.59188 <.0001	0.74026 <.0001	0.71213 <.0001	0.61399 <.0001	0.34890 0.0130
Robbery	0.48371 0.0004	0.59188 <.0001	1.00000	0.55708 <.0001	0.63724 <.0001	0.44674 0.0011	0.59068 <.0001
Assault	0.64855 <.0001	0.74026 <.0001	0.55708 <.0001	1.00000	0.62291 <.0001	0.40436 0.0036	0.27584 0.0525
Burglary	0.38582 0.0057	0.71213 <.0001	0.63724 <.0001	0.62291 <.0001	1.00000	0.79212 <.0001	0.55795 <.0001
Larceny	0.10192 0.4813	0.61399 <.0001	0.44674 0.0011	0.40436 0.0036	0.79212 <.0001	1.00000	0.44418 0.0012
Auto_Theft	0.06881 0.6349	0.34890 0.0130	0.59068 <.0001	0.27584 0.0525	0.55795 <.0001	0.44418 0.0012	1.00000

Метод главных компонент
(Principal Component Analysis)

•Пример: анализ криминальной обстановки в 50 штатах США по 7 переменным

•7 переменных – число преступлений на 100 тысяч населения штата:

Собственные значения корреляционной матрицы (в сумме равно 7!)

Eigenvalues of the Correlation Matrix

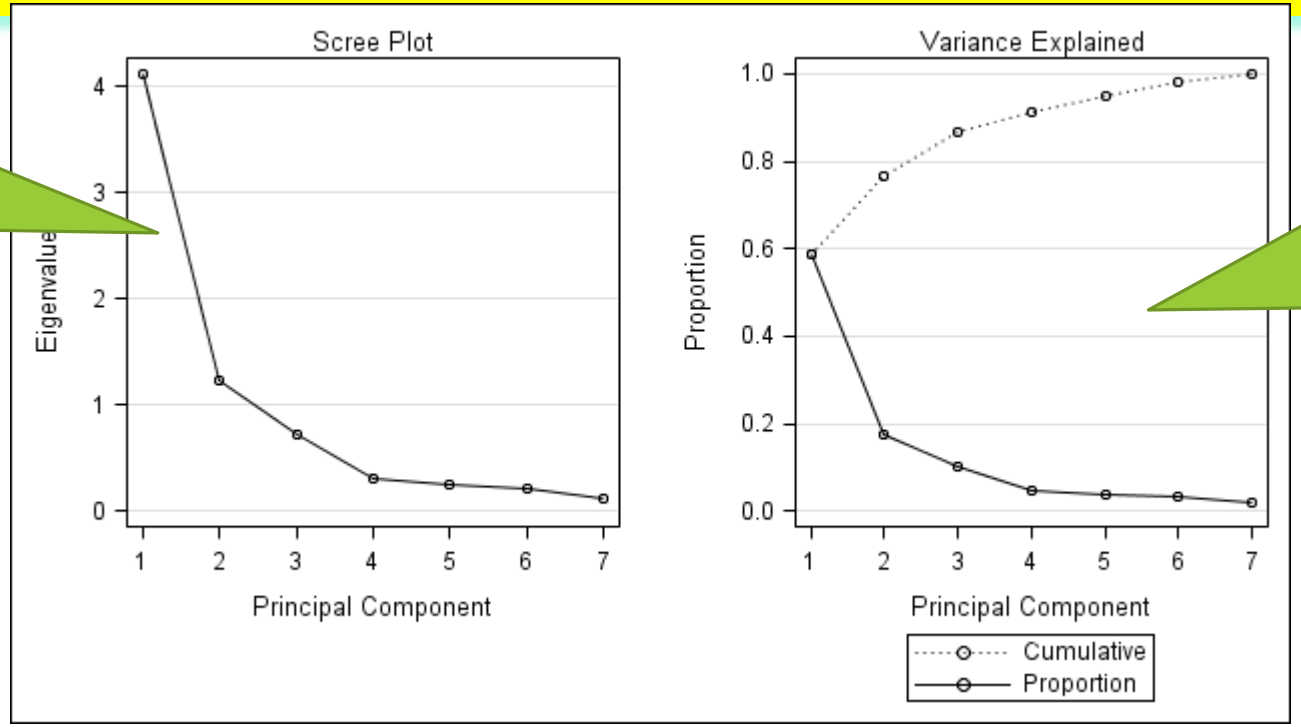
	Eigenvalue	Difference	Proportion	Cumulative
	4.11495951	2.87623768	0.5879	0.5879
2	1.23872183	0.51290521	0.1770	0.7648
3	0.72581663	0.40938458	0.1037	0.8685
4	0.31643205	0.05845759	0.0452	0.9137
5	0.25797446	0.03593499	0.0369	0.9506
6	0.22203947	0.09798342	0.0317	0.9823
7	0.12405606		0.0177	1.0000

Метод главных компонент
(Principal Component Analysis)

•Пример: анализ криминальной обстановки в 50 штатах США по 7 переменным

•7 переменных – число преступлений на 100 тысяч населения штата:

ГРАФИК: Собственные значения корреляционной матрицы в зависимости от номера PC (Scree plot)



ГРАФИКИ: Объясненная дисперсия (доля от общей дисперсии) в зависимости от номера PC (пропорция) и кумулятивная

Метод главных компонент
(Principal Component Analysis)

Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Murder	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
Rape	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
[illegible]	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
Assault	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
Burglary	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
Larceny	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
Auto_Theft	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

Нагрузки (loadings) главных компонент на исходные переменные

$$\begin{aligned}\text{Prin1} = {} & 0.300279 \times (\text{Murder}) \\ & + 0.431759 \times (\text{Rape}) \\ & + 0.396875 \times (\text{Robbery}) \\ & \cdot \\ & \cdot \\ & \cdot \\ & + 0.295177 \times (\text{Auto\_Theft})\end{aligned}$$

$$\begin{aligned}\text{Prin2} = {} & -0.629174 \times (\text{Murder}) \\ & - 0.169435 \times (\text{Rape}) \\ & + 0.042247 \times (\text{Robbery}) \\ & \cdot \\ & \cdot \\ & \cdot \\ & - 0.502421 \times (\text{Auto\_Theft})\end{aligned}$$

ИНТЕРПРЕТАЦИЯ:
Первая главная компонента – общий уровень преступности
Вторая главная компонента – противоположный характер преступлений против личности и преступлений против собственности
Последующие главные компоненты – трудно интерпретируемы

Метод главных компонент –что выдается в выходной набор
(Principal Component Analysis)

Дополнительно рассчитываются 7 новых переменных PRIN1 – PRIN7
Это 7 главных компонент, но из них первые 2 объясняют около 80% общей дисперсии!
Последними несколькими можно пренебречь, они «объясняют» очень малую долю исходной дисперсии…

Obs	State	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft	Prin1	Obs	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
1	Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7	-0.04988	1	-2.09610	0.50164	0.25099	0.49849	0.43362	0.11807
2	Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3	2.42151	2	0.16652	-0.06973	1.16047	1.47005	-1.49781	0.46481
3	Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5	3.01414	3	0.84495	-1.75195	-0.11621	0.28021	1.07044	0.05751
4	Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4	-1.05441	4	-1.34544	-0.01834	0.02154	0.02269	-0.38604	-0.31067
5	California	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5	4.28380	5	0.14319	0.27615	0.02512	0.05793	-0.37708	-0.46401
6	Colorado	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1	2.50929	6	0.91660	-1.15158	0.11260	-0.16923	-0.33103	-0.24066
7	Connecticut	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2	-0.54133	7	1.50123	0.78389	0.08619	0.18489	0.28140	-0.08962
8	Delaware	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0	0.96458	8	1.29674	-0.52586	-0.41732	-0.01872	0.40356	0.20616
9	Florida	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4	3.11175	9	-0.60392	-1.21541	0.49508	-0.81967	0.28958	0.35866
10	Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9	0.49041	10	-1.38079	0.244	0.06248	0.20210	0.02578	-0.33221

Значения новых переменных – главных компонент для конкретных наблюдений - метки (Scores) главных компонент на наблюдения

Метод главных компонент: статистики главных компонент

(Principal Component Analysis: statistics)

Можно (из любопытства!) потрудиться пересчитать статистики для 7 главных компонент

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Prin1	50	0	2.02854	0	-3.96408	5.26699
Prin2	50	0	1.11298	0	-2.54671	2.63105
Prin3	50	0	0.85195	0	-1.75195	2.73661
Prin4	50	0	0.56252	0	-1.18025	1.81310
Prin5	50	0	0.50791	0	-1.35485	1.47005
Prin6	50	0	0.47121	0	-1.49781	1.07044
Prin7	50	0	0.35222	0	-0.91333	0.82538

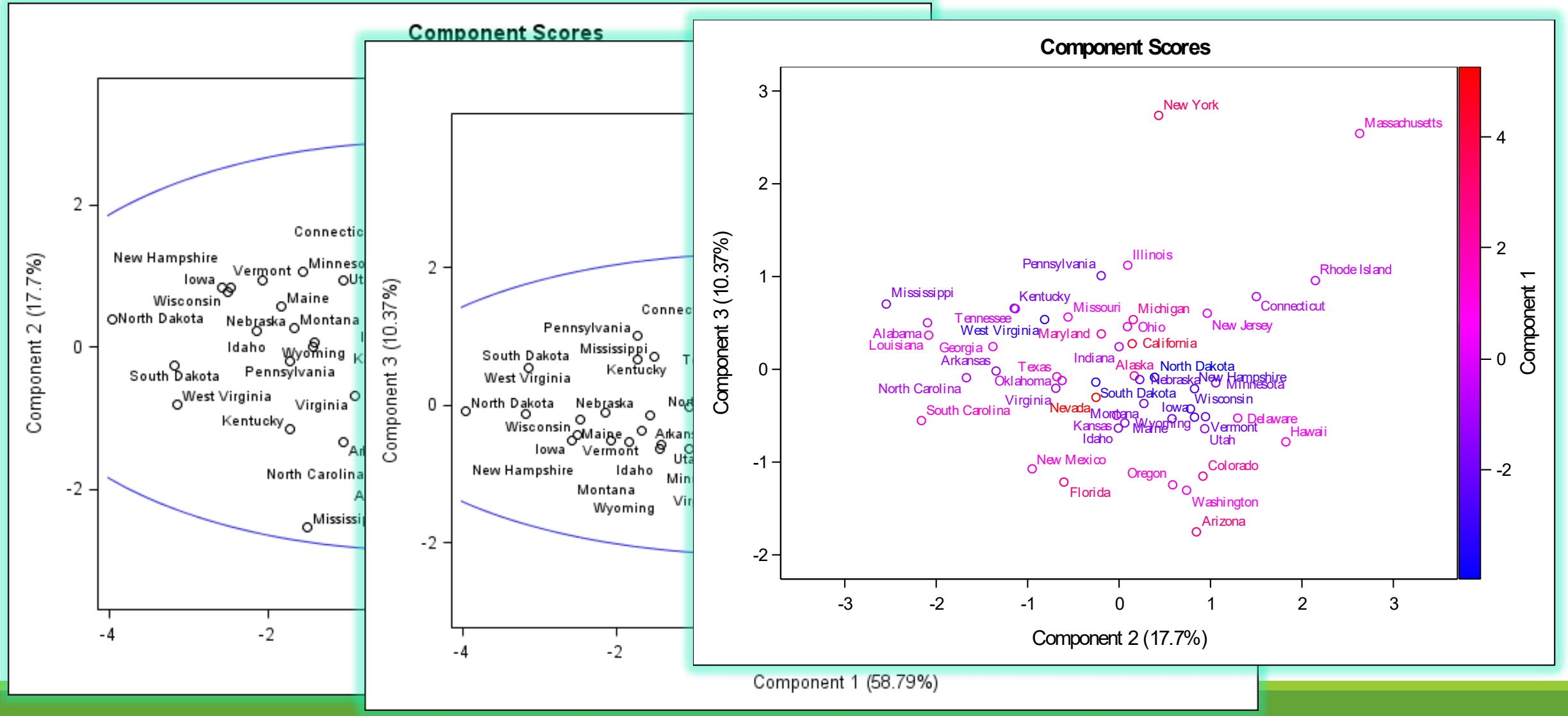
Все главные компоненты приведены к нулевому среднему!

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Prin1	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin2	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin3	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin4	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin5	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000
Prin6	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000
Prin7	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000

УБЕЖДАЕМСЯ:
Все главные компоненты взаимно не коррелированы!

Метод главных компонент: как выглядят различные наблюдения – различные штаты США, -в новых координатах – графики меток главных компонент

(Principal Component Analysis – component scores)



Анализ главных компонент: реализация

Реализация метода главных компонент – строгий математический аппарат, но допускается неоднозначность:
Важно найти взаимно ортогональные направления, а длины соответствующих векторов будут зависеть от того, какую матрицу мы используем (корреляционную, скорректированную или нескорректированную ковариационную)
Метод ресурсоемкий, могут возникнуть трудности при работе с БОЛЬШИМИ ДАННЫМИ
Есть неопределенность в выборе того, что считаем положительным или отрицательным направлением ортогональных векторов: можно нагрузки и метки одновременно умножить на -1, результат не изменится
Это надо учитывать, проводя анализ с различными программными средствами

Реализация метода главных компонент в SAS:

Основная процедура: **PROC PRINCOMP**

PROC FACTOR

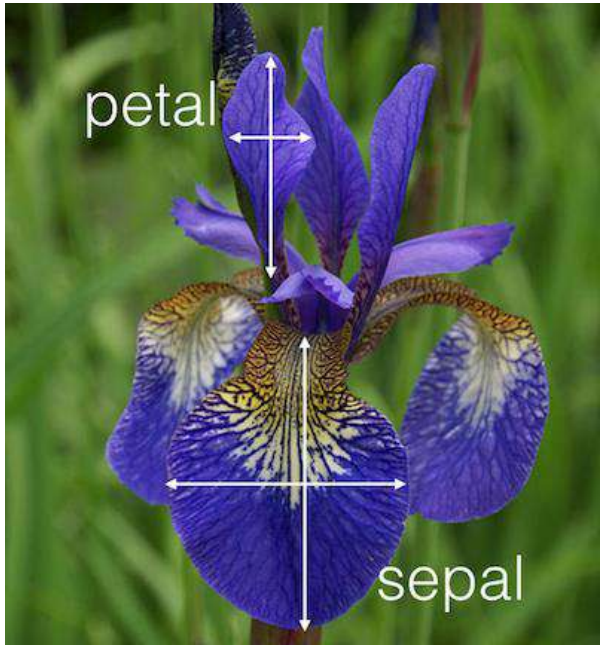
Реализация регрессии на главные компоненты в SAS:

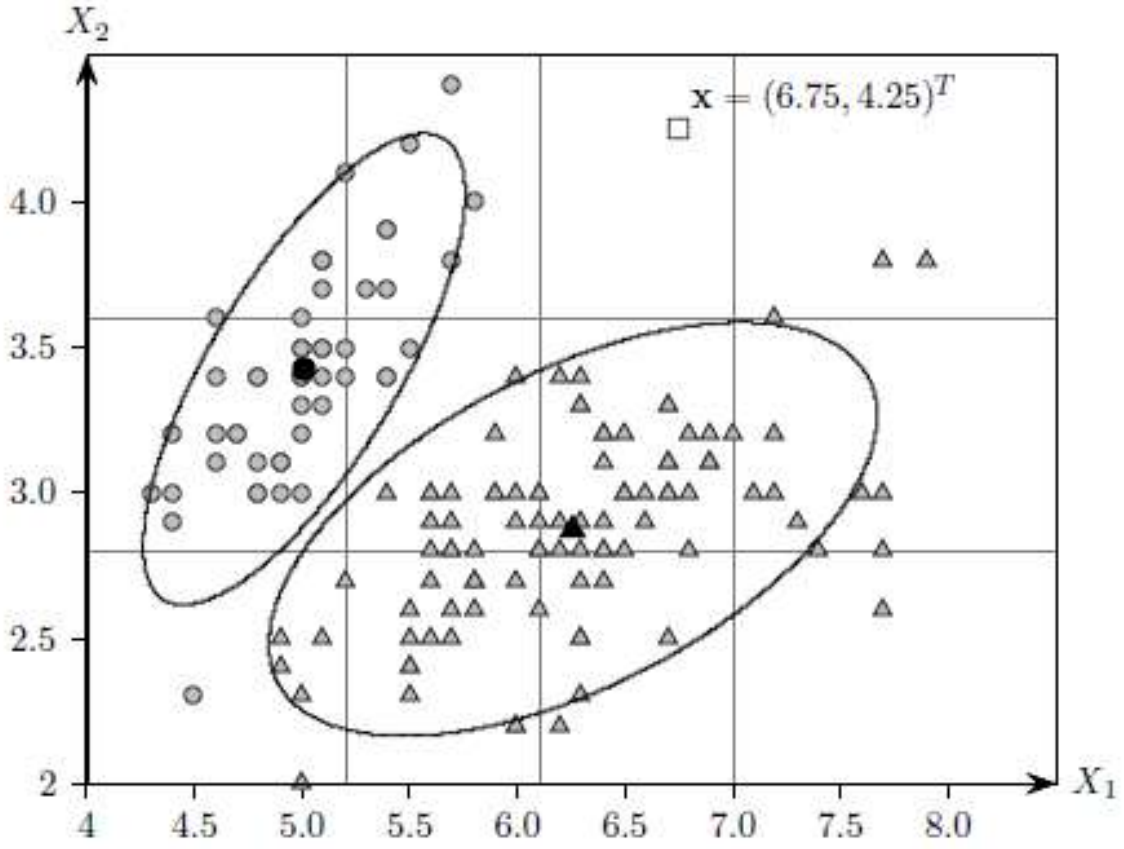
PROC PRINCOMP + PROC REG
Или :
PROC PLS (Partial Least Squares, METHOD=PCR)
SAS STUDIO: *Регрессия частных наименьших квадратов* – так называется задача

Но самое важное – интерпретация результатов!!!!

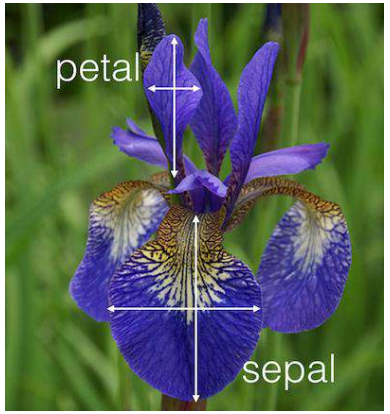
Задачи :
Анализа кандидатур, предпочтений
Анализ «выдающихся исторических личностей»

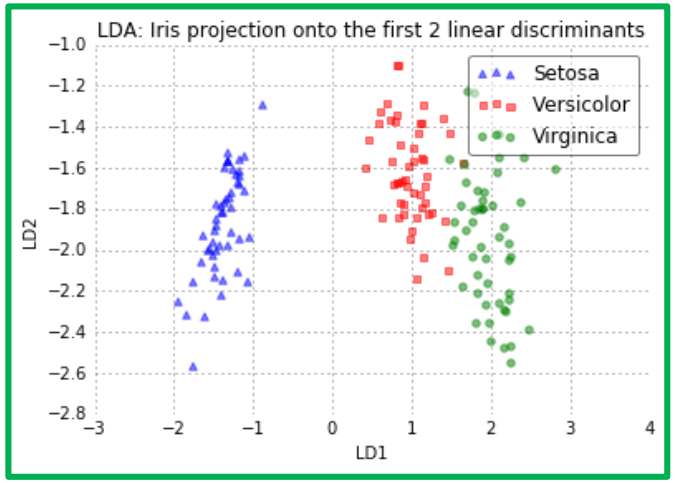
Классификация – дискриминантный анализ
Ирис Фишера – эталон массива данных для дискриминантного анализа

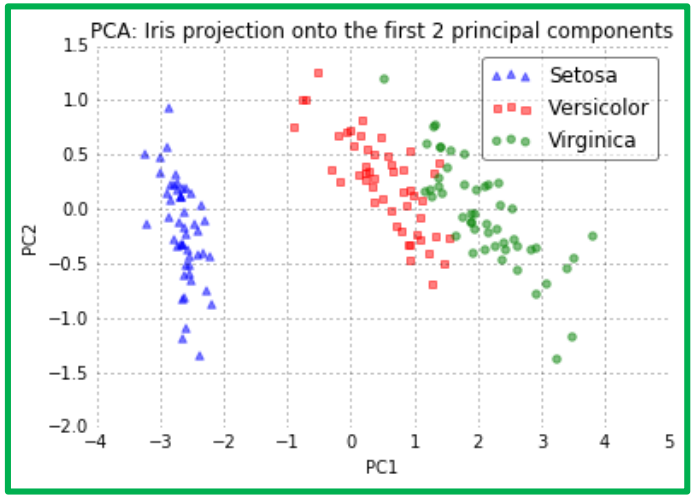


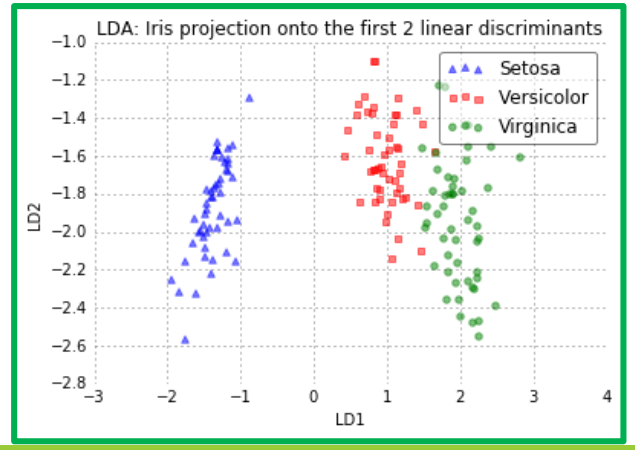


Классификация – дискриминантный анализ



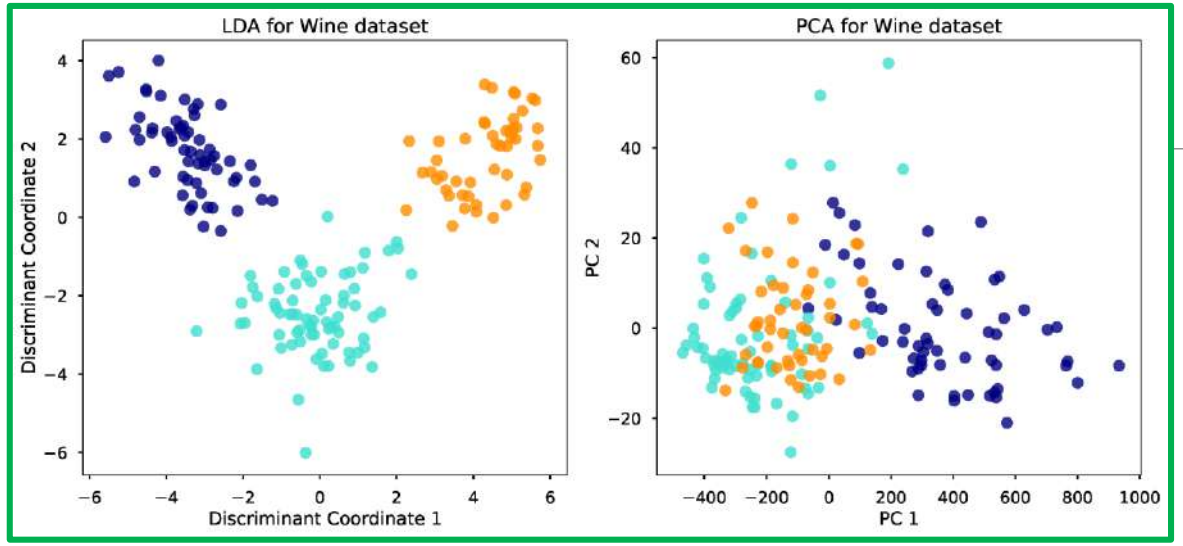






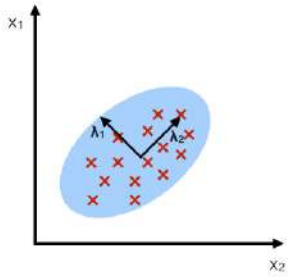
Поворот системы координат – линейные преобразование переменных
Может быть преобразование – методом главных компонент – наилучшим образом объяснить ОБЩУЮ дисперсию
Может быть преобразование – методом канонического дискриминантного анализа – наилучшим образом объяснить МЕЖГРУППОВУЮ дисперсию

Классификация – дискриминантный анализ

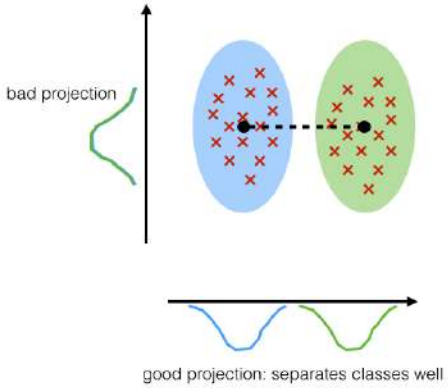


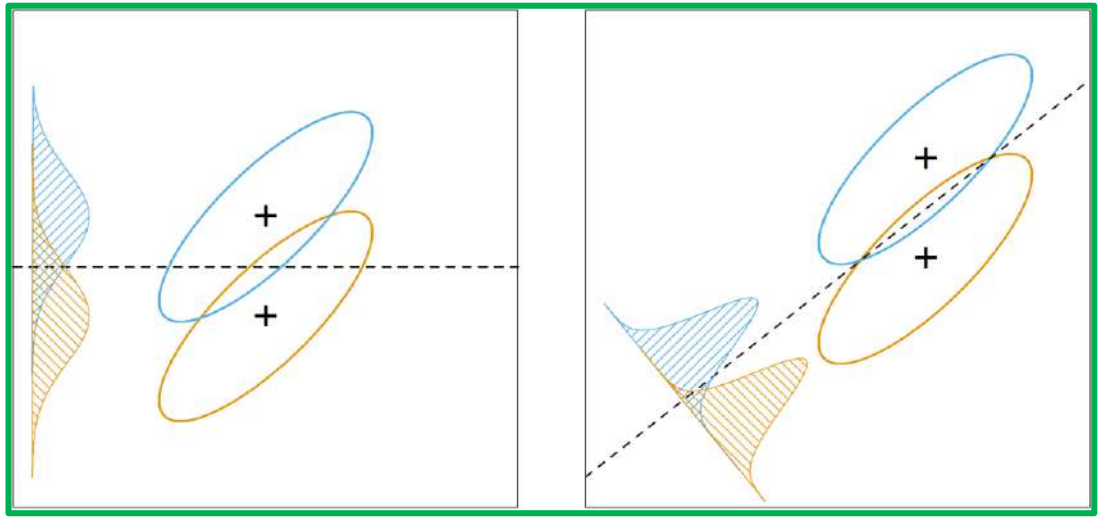
Слева – плотности в исходных («старых») переменных сильно перекрываются – если используем традиционный дискриминантный анализ
Справа – в результате канонического дискриминантного анализа получаются новые переменные, в пространстве которых плотности максимально НЕ ПЕРЕКРЫВАЮТСЯ!

PCA:
component axes that maximize the variance



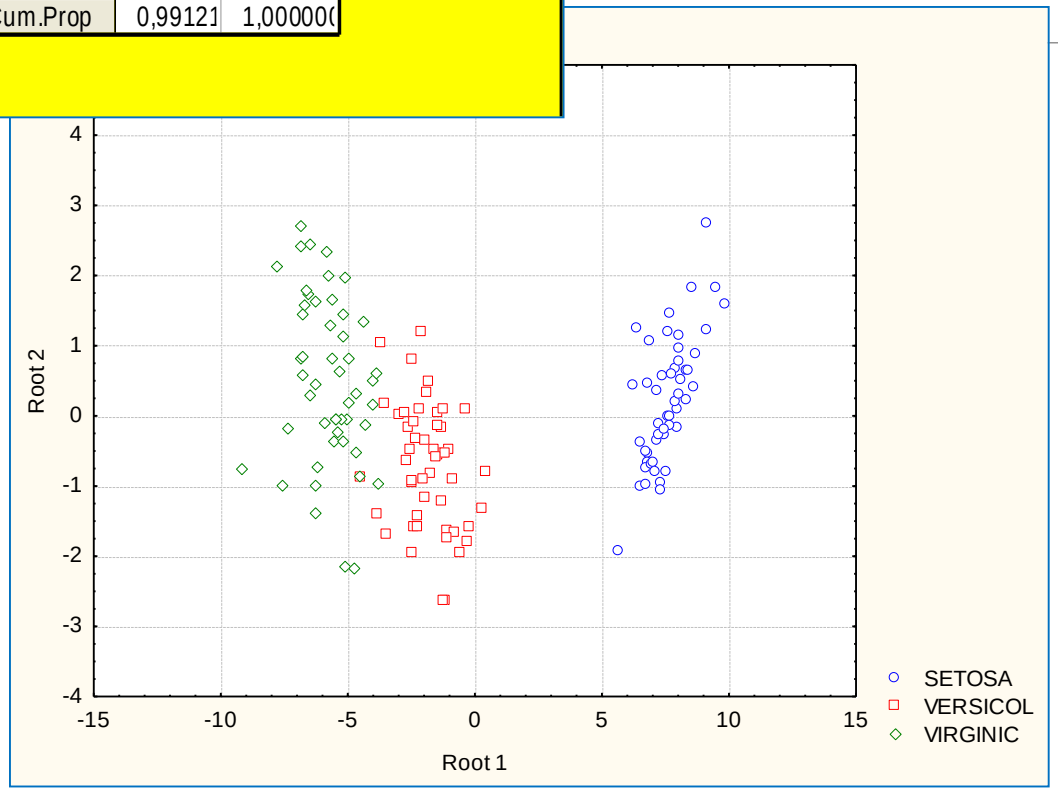
LDA:
maximizing the component axes for class-separation

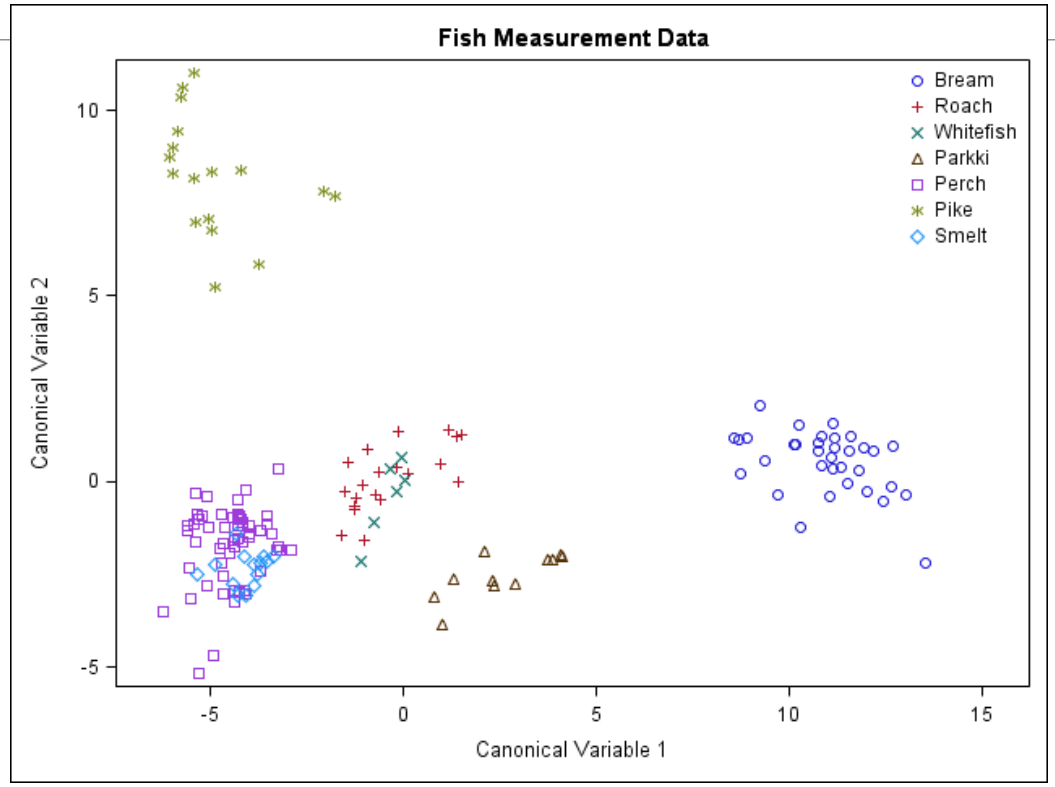




Variable	Standardized Coefficients (Irisdat) for Canonical Variables	
	Root 1	Root 2
SEPALLEN	0,42695	0,012408
SEPALWID	0,52124	0,735261
PETALLEN	-0,9472(	-0,401038
PETALWID	-0,5751(	0,581040
Eigenval	32,19193	0,285391
Cum.Prop	0,99121	1,000000

Дискриминантный анализ (Discriminant analysis)





Канонический дискриминантный анализ: уже первые несколько новых (канонических) переменных хорошо объясняют максимальную долю дисперсии – дисперсии МЕЖКЛАССОВОЙ!!! Всего же канонических переменных – минимум из: числа классов минус 1 и числа исходных переменных

Спасибо за внимание!

Лекция -окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

ТЕЛ. (484)3974658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

Кластерный анализ (Cluster analysis)

Есть ряд подходов, которые можно назвать АГРЕГАЦИЕЙ ДАННЫХ.

Можно группировать (агрегировать) переменные в факторы или компоненты, формировать обобщенные переменные, но при этом неизменным остается множество объектов (наблюдений)

- Метод (анализ) главных компонент (PCA)
- Факторный анализ (Factor analysis)

Можно группировать (агрегировать) объекты (наблюдения), формируя группы (кластеры) схожих в пространстве переменных объектов

- Cluster analysis

В анализе дисперсий (ANOVA) объекты заранее отнесены к известным группам

В кластерном анализе группы формируются из имеющихся объектов исходя из схожести объектов в пространстве переменных (группы не заданы априори)

Кластерный анализ (Cluster analysis)

Примеры практических применений кластеризации:

Сегментация пользователей услуги

Сегментация клиентов банка (не путать с задачей решения давать-не давать кредит)

Кластеризация источников сетевого траффика

Спам-фильтрация электронной почты (группировка сообщений в соответствии с результатами анализа их разных частей – отправитель, получатели, тема, особенности содержания, и т.д.)

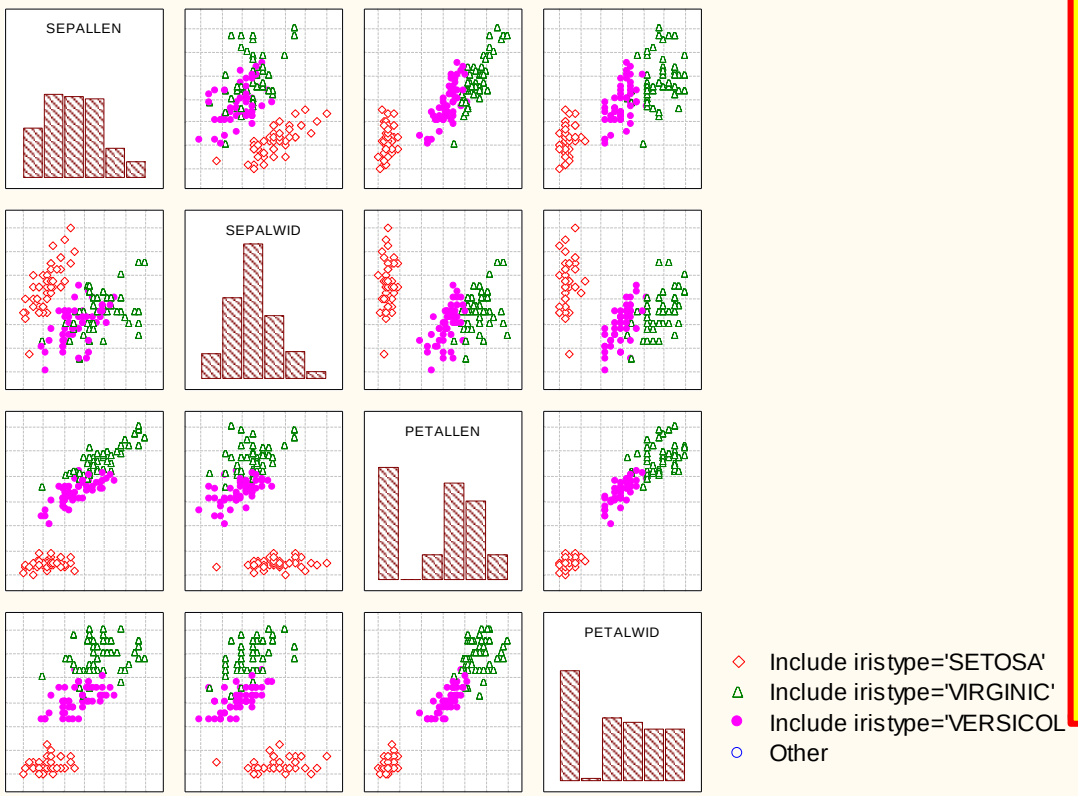
Городское планирование –кластеризация домов, квартир

Кластеризация документов

Анализ рынка недвижимости

В биологии – типизация объектов

Кластерный анализ (Cluster analysis)



Из википедии:
Кластерный анализ (англ. Data clustering) — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.
Кластерный анализ — это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы (кластеры).
Кластер — группа элементов, характеризуемых общим свойством, главная цель кластерного анализа — нахождение групп схожих объектов в выборке

Кластеризация наблюдений или кластеризация переменных? В SAS можно делать и то и другое!

Группа – кластер, таксон
Какое количество групп задать – это нетривиальная задача!!

Кластерный анализ (Cluster analysis)

Общие положения:

Есть выборка из L многомерных объектов – обучающая выборка:

$X_1, X_2, \ldots, X_l, (i=1,\ldots,L)$

Y_i –отклик, зависимая переменная, - отсутствует!

Надо разбить объекты на группы так, чтобы внутри каждой группы были максимально схожие между собой объекты, и при этом максимально не схожие с объектами других групп

Т.е., восстановить для каждого объекта обучающей выборки значение Y_i

Задача обучения БЕЗ УЧИТЕЛЯ **(UNSUPERVISED LEARNING)** – более общая, но обратное верно - всякая кластеризация – это **UNSUPERVISED LEARNING**

Имеются разновидности: **ЧАСТИЧНОЕ ОБУЧЕНИЕ (SEMI-SUPERVISED LEARNING)** – оставляем на будущее

Кластеризация: постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка;

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров,

$a\colon X \to Y$ — алгоритм кластеризации,

такие, что:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

Это задача *обучения без учителя* (unsupervised learning).

Кластерный анализ (Cluster analysis)

Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров Y j, как правило, неизвестно заранее;
- результат кластеризации сильно зависит от метрики , выбор которой также является эвристикой.

Пример: сколько здесь кластеров?

Кластерный анализ (Cluster analysis)

Цели и сферы использования кластеризации

- **Упростить дальнейшую обработку данных, разбить множество X` на группы схожих объектов, чтобы работать с каждой группой в отдельности (задачи классификации, регрессии, прогнозирования).**
- **Сократить объем хранимых данных, оставив по одному представителю от каждого кластера (задачи сжатия данных).**
- **Выделить нетипичные объекты, которые не подходят ни к одному из кластеров (задачи одноклассовой классификации).**
- **Построить иерархию множества объектов, пример - классификация животных и растений К.Линнея (задачи таксономии).**

Кластерный анализ (Cluster analysis)

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены – у нас НЕТ УЧИТЕЛЯ!!!!!

Синонимами термина "кластеризация" являются "автоматическая к

Кластеризация предназначена для разбиения совокупности объект

выборки представить как точки в признаковом пространстве, то зад

точек".

Здесь не нужно искать **правило отнесения новых объектов** к

Само понятие "кластер" определено неоднозначно: в каждом иссл

er)

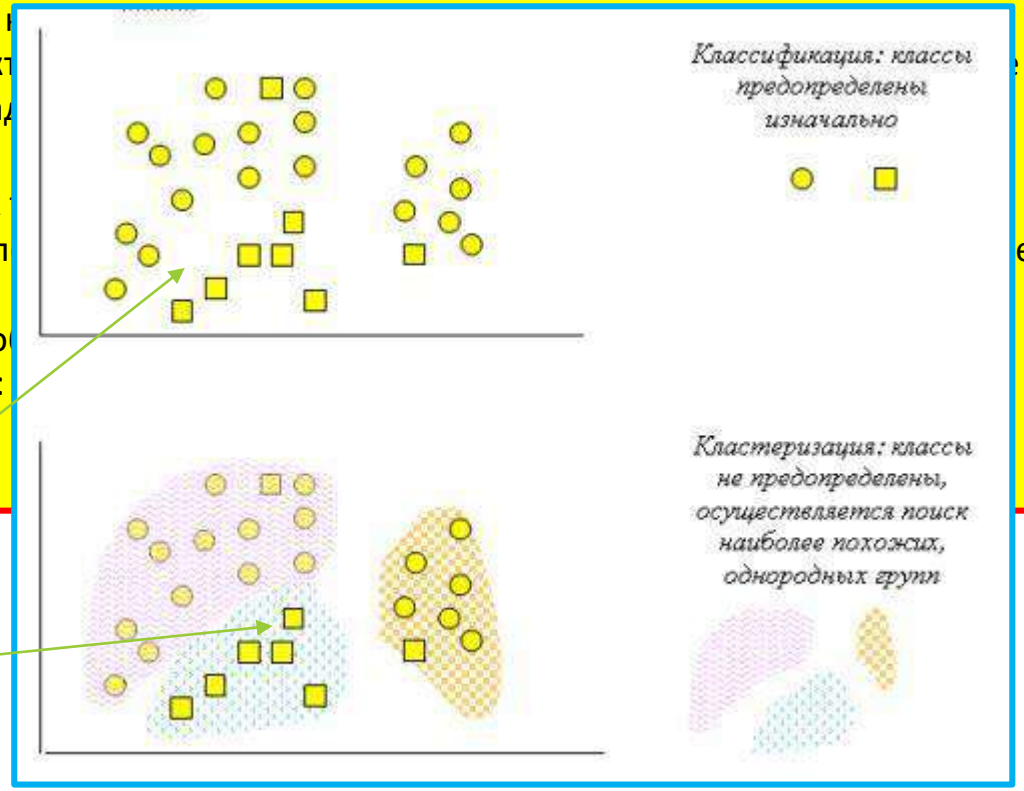
как "скопление", "гроздь".

Кластер можно охарактеризовать как группу объектов, имеющих о

Характеристиками каждого кластера можно назвать два признака:

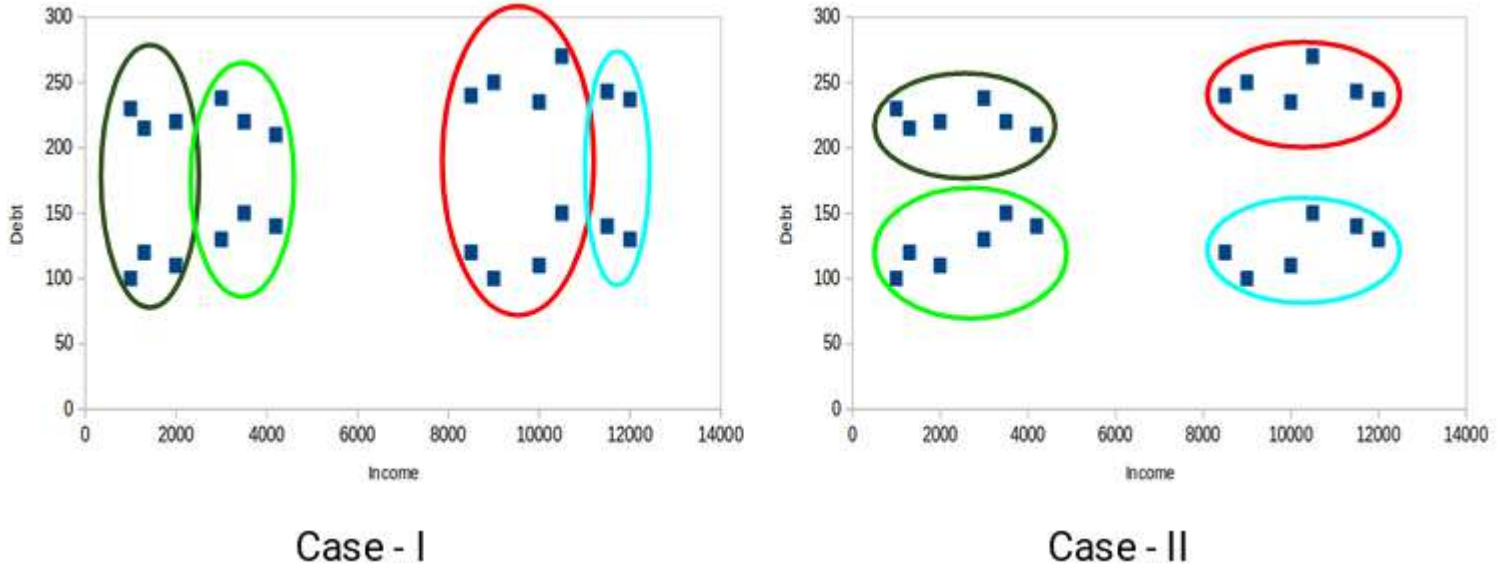
· внутренняя однородность;

· внешняя изолированность.



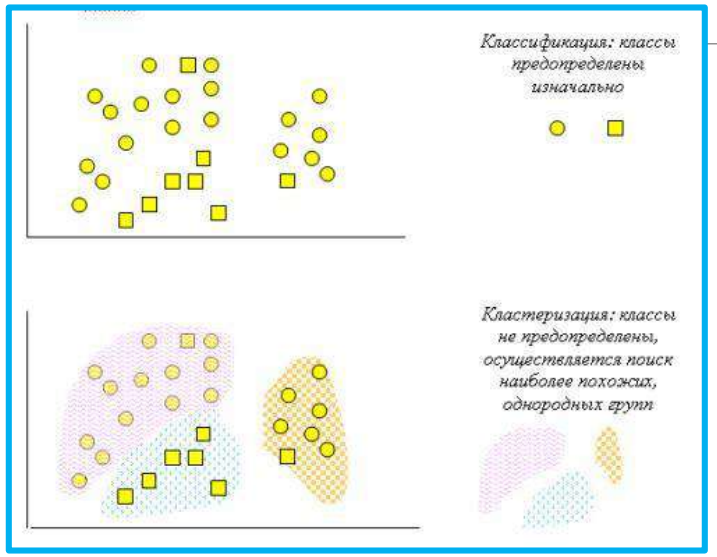
СОВСЕМ РАЗНЫЕ КЛАССИФИКАЦИИ!

Кластерный анализ (Cluster analysis)

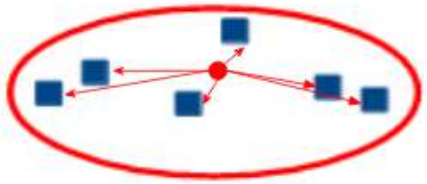


СОВСЕМ РАЗНЫЕ КЛАССИФИКАЦИИ! ПРИ ЭТОМ ЧИСЛО КЛАСТЕРОВ РАВНО 4 В ОБОИХ СЛУЧАЯХ!

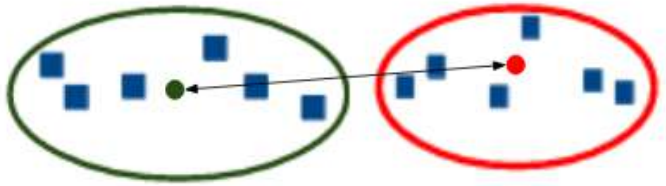
Кластерный анализ (Cluster analysis) - эвристики



Индекс Данна (Dunn Index):



Intra cluster distance



Inter cluster distance

$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

Clusters are far apart

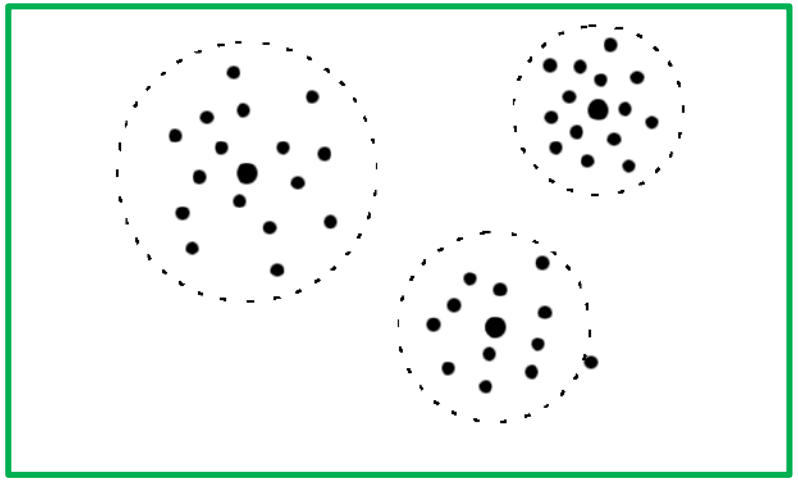
$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

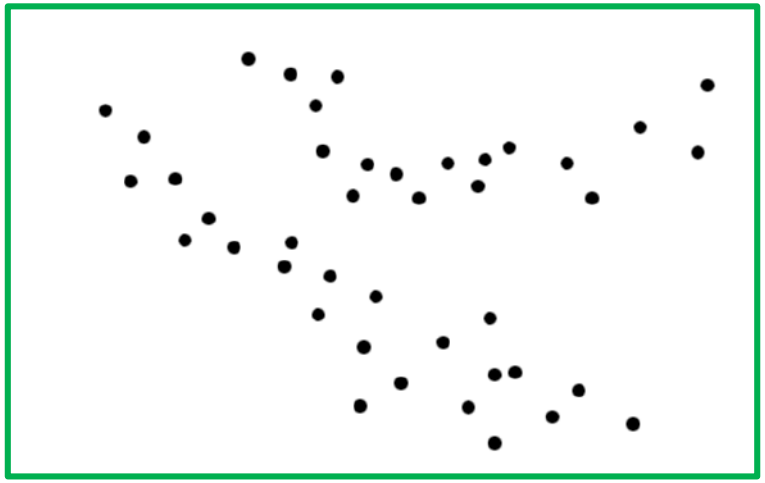
$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

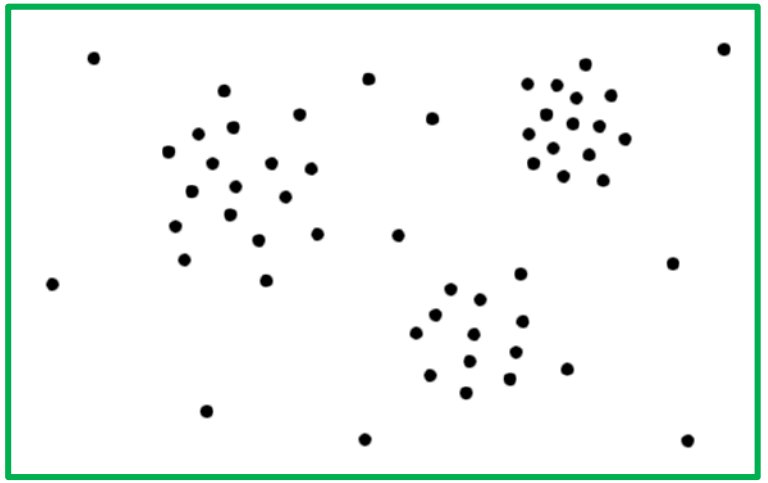
Clusters are compact

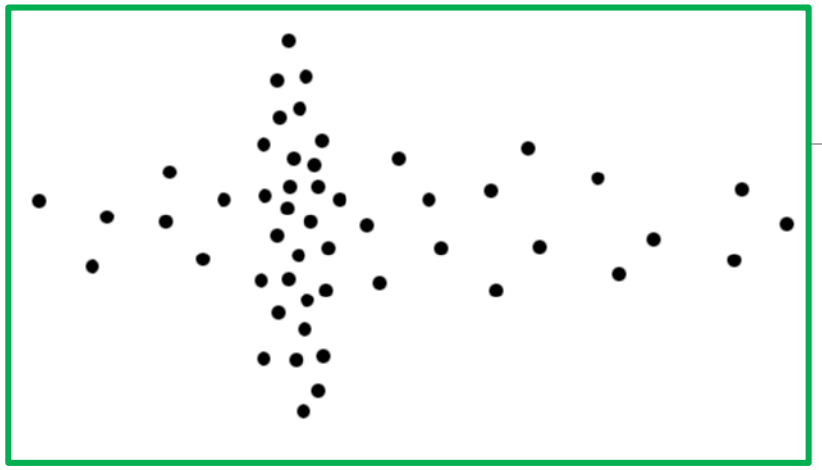
Инерция кластера – сумма расстояний всех точек внутри кластера от центра кластера. А почему не сумма квадратов? Или другая подобная характеристика? Это все-эвристика…

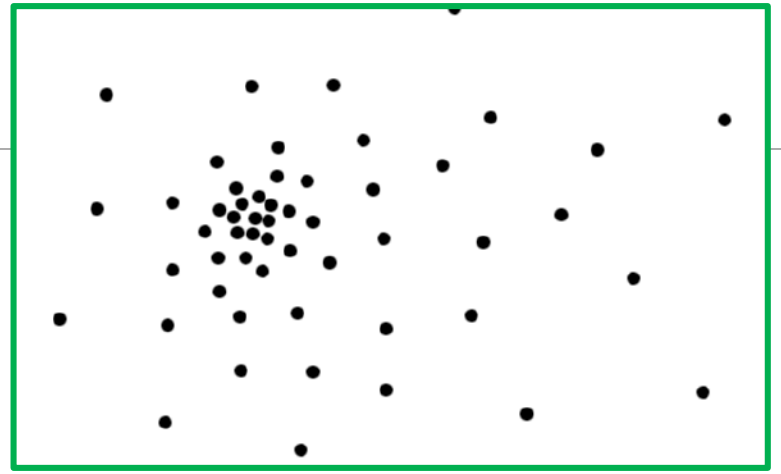
Типы кластерных структур

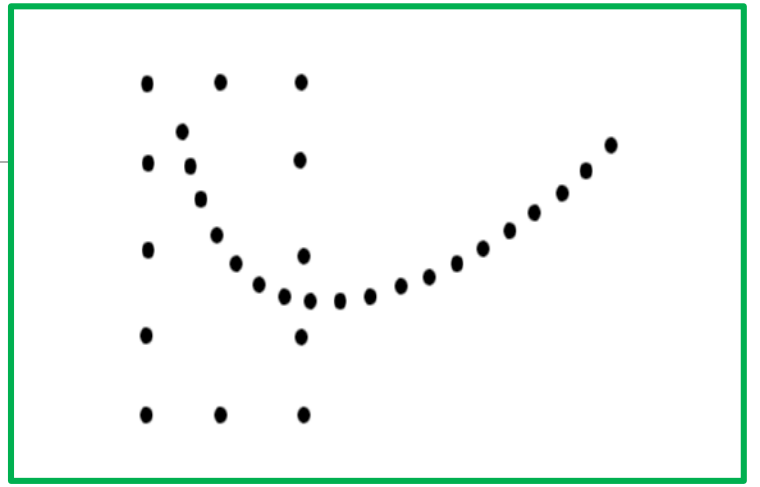






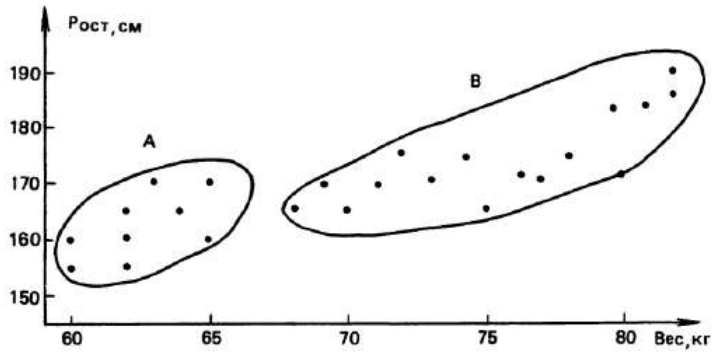




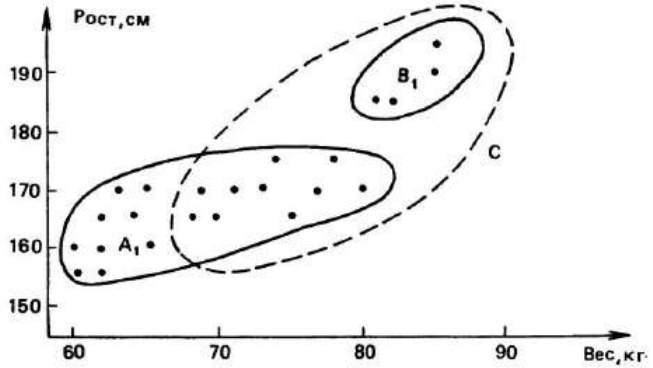


Типы кластерных структур

- **Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.**
- **Понятие «тип кластерной структуры» зависит от метода кластеризации и также не имеет формального определения**
- **Результат зависит от нормировки признаков:**



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

Качество кластеризации в метрическом пространстве

Пусть известны только попарные расстояния между объектами.

- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum\limits_{i<j} [a_i = a_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [a_i = a_j]} \to \min .$$

- Среднее межкластерное расстояние:

$$F_1 = \frac{\sum\limits_{i<j} [a_i \neq a_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [a_i \neq a_j]} \to \max .$$

- Отношение пары функционалов: $F_0/F_1 \to \min$.

Здесь нужно просчитывать все объекты попарно! Выбрать, во-первых, более не менее универсальные критерии, и, во-вторых, попытаться свести задачу к оптимизационной

Качество кластеризации в линейном векторном пространстве

Пусть объекты x_i задаются векторами $\big(f_1(x_i), \ldots, f_n(x_i)\big)$.

- Сумма средних внутрикластерных расстояний:

$$\Phi_0 = \sum_{a \in Y} \frac{1}{|X_a|} \sum_{i:\, a_i = a} \rho(x_i, \mu_a) \to \min,$$

$X_a = \{x_i \in X^\ell \mid a_i = a\}$ — кластер a,
μ_a — центр масс кластера a.

- Сумма межкластерных расстояний:

$$\Phi_1 = \sum_{a,b \in Y} \rho(\mu_a, \mu_b) \to \max.$$

- Отношение пары функционалов: $\Phi_0 / \Phi_1 \to \min$.

Если известны центры кластеров, и мы можем усреднять все координаты объектов кластеров, чтобы уточнять координаты центров кластеров (пересчитывать координаты центров при изменениях составов кластеров), - можно вычислять другие функционалы

Методы кластеризации

Три основные группы алгоритмов:

- **Алгоритмы к средних (k- means) – неиерархическая**
- **Алгоритмы С-средних с нечеткими множествами (C-Means Fuzzy Clustering)**
- **Иерархические алгоритмы (Hierarchial)**
- **Алгоритмы, основанные на плотности (Density Based Clustering)**

В способах и алгоритмах решения задач кластеризации –очень много эвристики!

- **На сколько кластеров проводить разбиение?**
- **Каков критерий качества кластеризации – целевая функция, которую нужно оптимизировать?**
- **В какой метрике проводить расчет расстояний между многомерными объектами – от этого зависит результат дальнейших обобщений!**

Кластерный анализ (Cluster analysis)

Неиерархическая кластеризация - Алгоритм k-средних (k-means)

Метод К средних – алгоритм (алгоритм для линейного векторного пространства:

Минимизация суммы квадратов внутрикластерных расстояний:

$$\sum_{i=1}^{\ell} \|x_i - \mu_{a_i}\|^2 \to \min_{\{a_i\},\{\mu_a\}}, \quad \|x_i - \mu_a\|^2 = \sum_{j=1}^{n} \big(f_j(x_i) - \mu_{aj}\big)^2$$

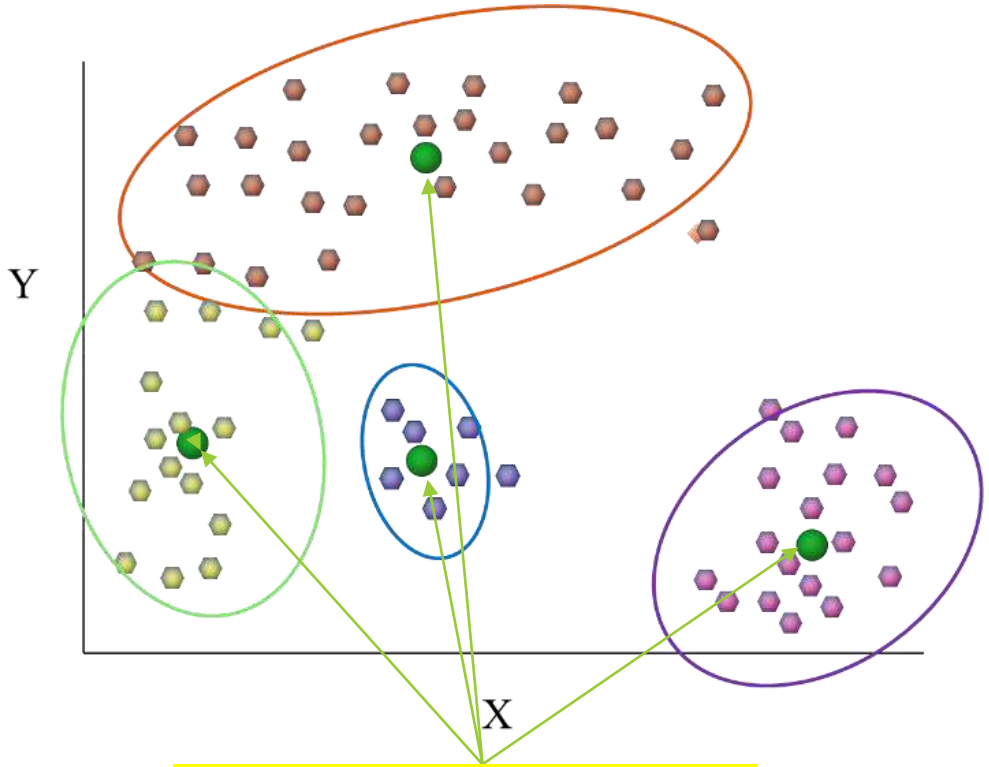
Алгоритм Ллойда (если число кластеров К – фиксировано, Y - метки):

вход: X^ℓ, $K = |Y|$; **выход:** центры кластеров μ_a, $a \in Y$;
$\mu_a :=$ начальное приближение центров, для всех $a \in Y$;
повторять
  отнести каждый x_i к ближайшему центру:
  $a_i := \arg\min_{a \in Y} \|x_i - \mu_a\|, \quad i = 1, \ldots, \ell;$
  вычислить новые положения центров:
  $\mu_a := \dfrac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y;$
пока a_i не перестанут изменяться;

Про межкластерные расстояния – забыли, используем только внутрикластерные Алгоритм сходится!!!

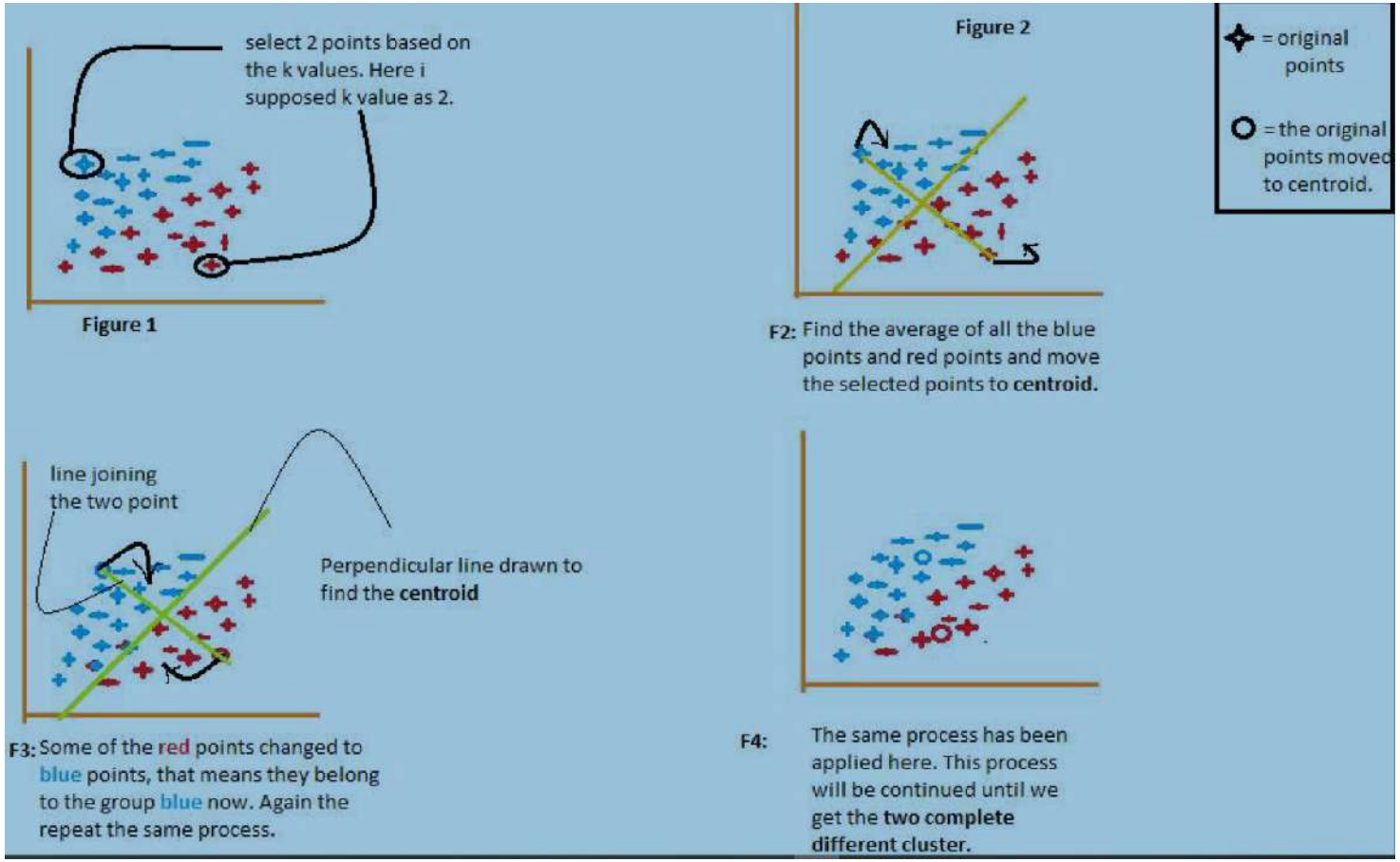
Кластерный анализ
(Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means)

Алгоритм k-средних (k-means) строит k кластеров, расположенных на возможно бОльших расстояниях друг от друга.
Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно.
Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.
Общая идея алгоритма: заданное фиксированное число k кластеров наблюдений сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.



Центроиды кластеров

Кластерный анализ (Cluster analysis) Неиерархическая кластеризация - Алгоритм k-средних (k-means)

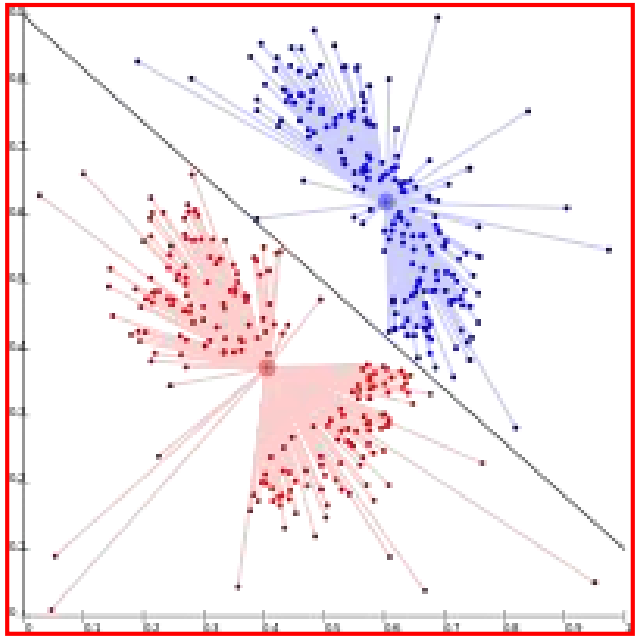


1. Предположили, что к=2 и будем этого придерживаться в дальнейшем. Выделили наугад две точки и предположили, что это центроиды. Отнесли к каждому кластеру те точки, которые ближе к его центроиду, чем к центроиду другого кластера.
2. Нашли средние всех координат. Пересчитали координаты центроидов.
3. Отнесли точки к уточненным центроидам
4. Повторяем процесс, пока координаты центроидов не перестанут изменяться, или пока число итераций не превысит заданный порог

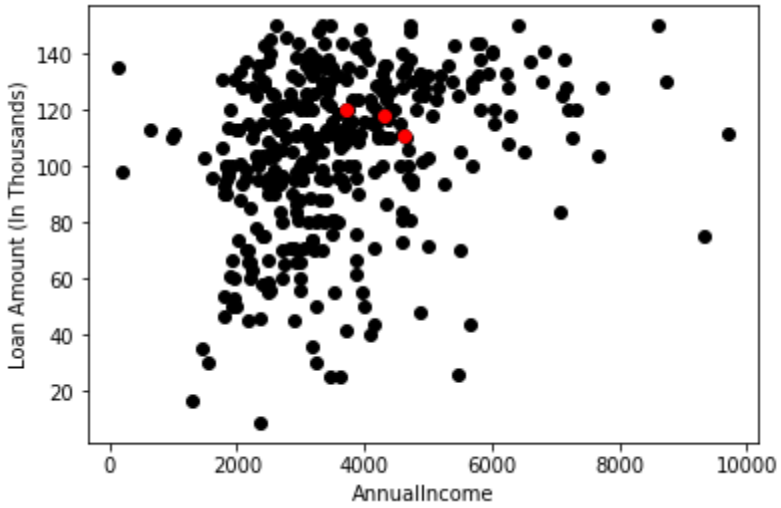
Кластерный анализ
(Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means)

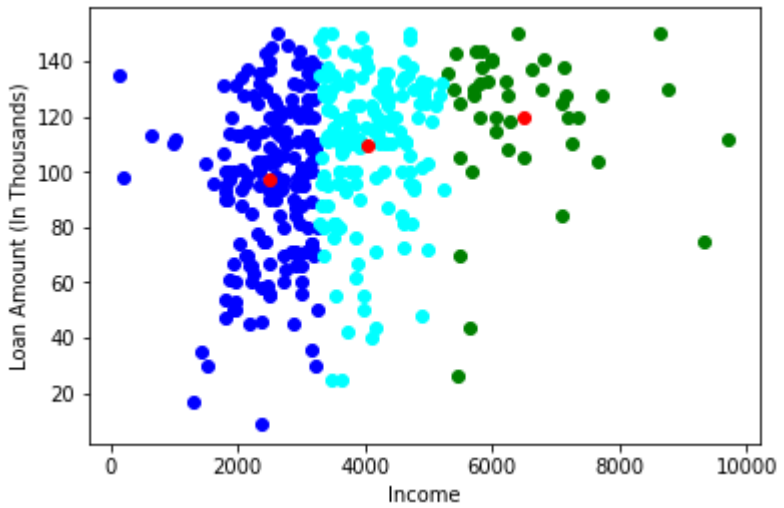
Метод К средних:

1. Выбрать К **произвольных** точек в качестве начальных центроидов
2. Отнести все наблюдения к одному или нескольким из К выбранных центроидов, основываясь на мере близости (способ оценки меры задается, но его выбор - эвристика)
3. Пересчитать центроиды каждой из К групп
4. Повторять шаги 2 и 3 до тех пор, пока центры не стабилизируются, или состав кластеров не перестанет меняться (или не исчерпается заданное число итераций)
5. Полученная на этот момент кластеризация на К групп и будет найденным решением!



Кластерный анализ (Cluster analysis) Неиерархическая кластеризация - Алгоритм k-средних (k-means): примеры неудачной кластеризации

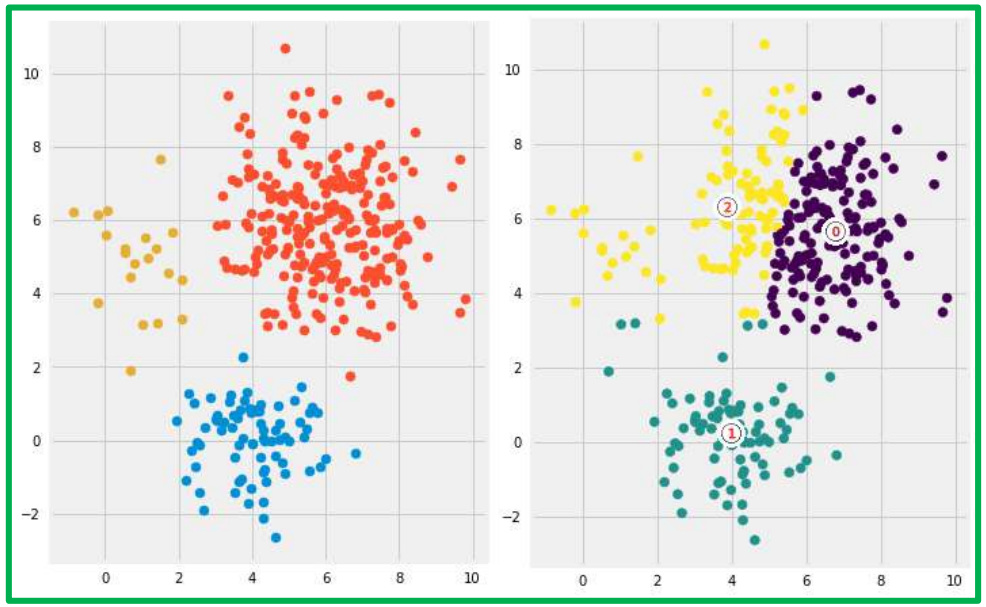


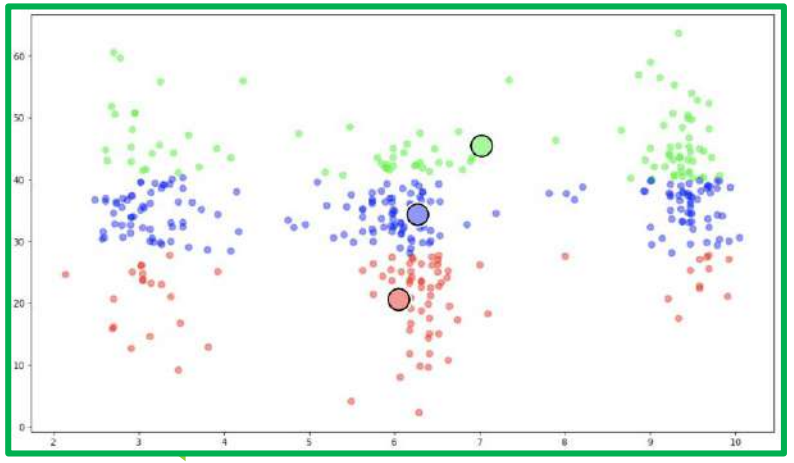


Берем три произвольные точки – исходные центры трех центроидов
Получаем три кластера с совершенно другими центроидами
Выбор трех произвольных точек - слева – остается один кластер, справа – получаем три кластера

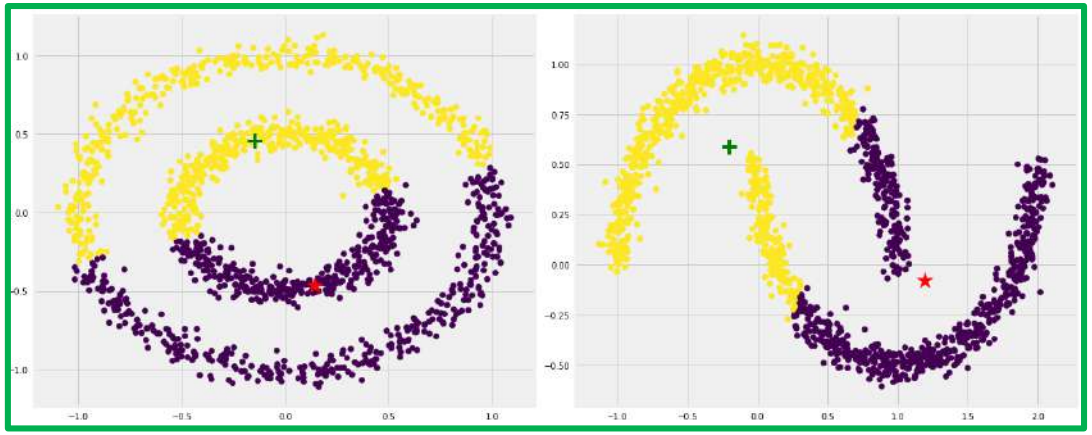
Кластерный анализ (Cluster analysis) Неиерархическая кластеризация - Алгоритм k-средних (k-means): примеры неудачной кластеризации

Причина: неудачное начальное приближение или существенная негауссовость кластеров





Три кластера разделены горизонтальными линиями!
А правильнее было разделить вертикальными!



Многих неудачных ситуаций (концентрические окружности, «два банана»), имеющих место при использовании метода к-средних, можно избежать, используя методы, основанные на анализе плотности, например, DBSCAN

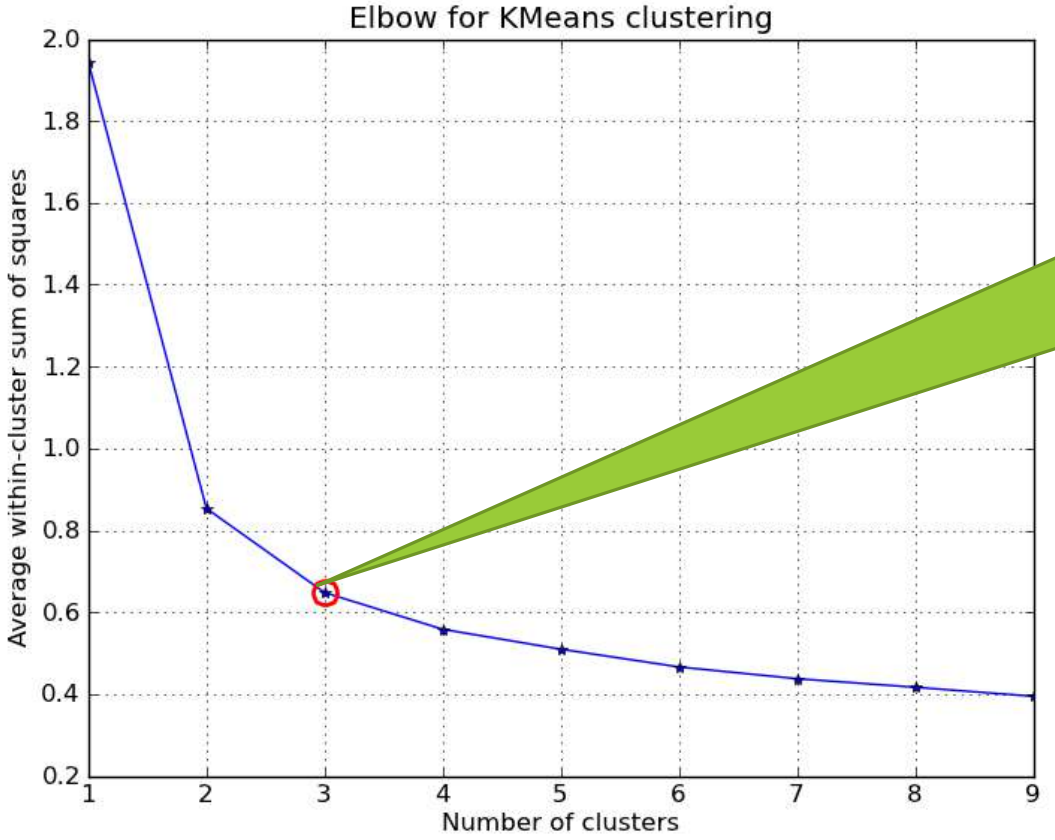
Кластерный анализ - сколько кластеров должно быть для кластеризации по К-средним? (Cluster analysis)

Идея. Выбрать критерий качества кластеризации и построить его значение для $K = 1, 2, \ldots$

- средняя сумма квадратов расстояния до центроида
- средний диаметр кластера

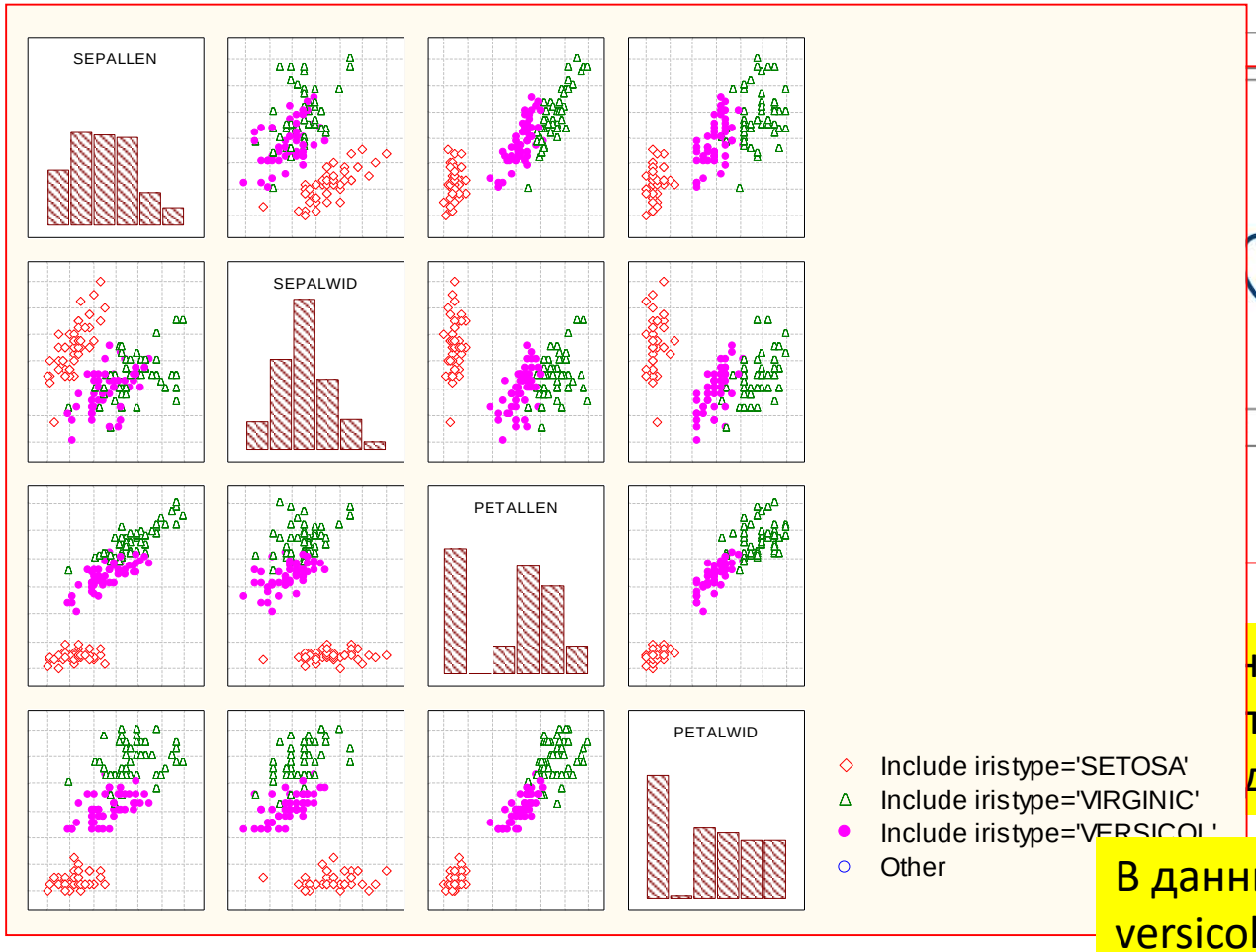
Метод «локтя» - Elbow method

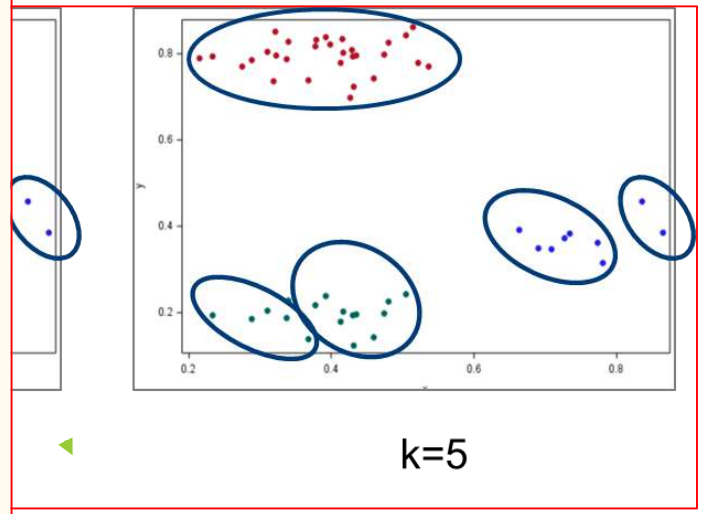
Но это – чистая эвристика!!!



Точка излома для кривой зависимости

Кластерный анализ
(Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means)





ного числа кластеров!!!
при двух переменных! При
дает!

В данных с Ирисом Фишера если задать maxclusters=2, versicol и virginic – объединятся!

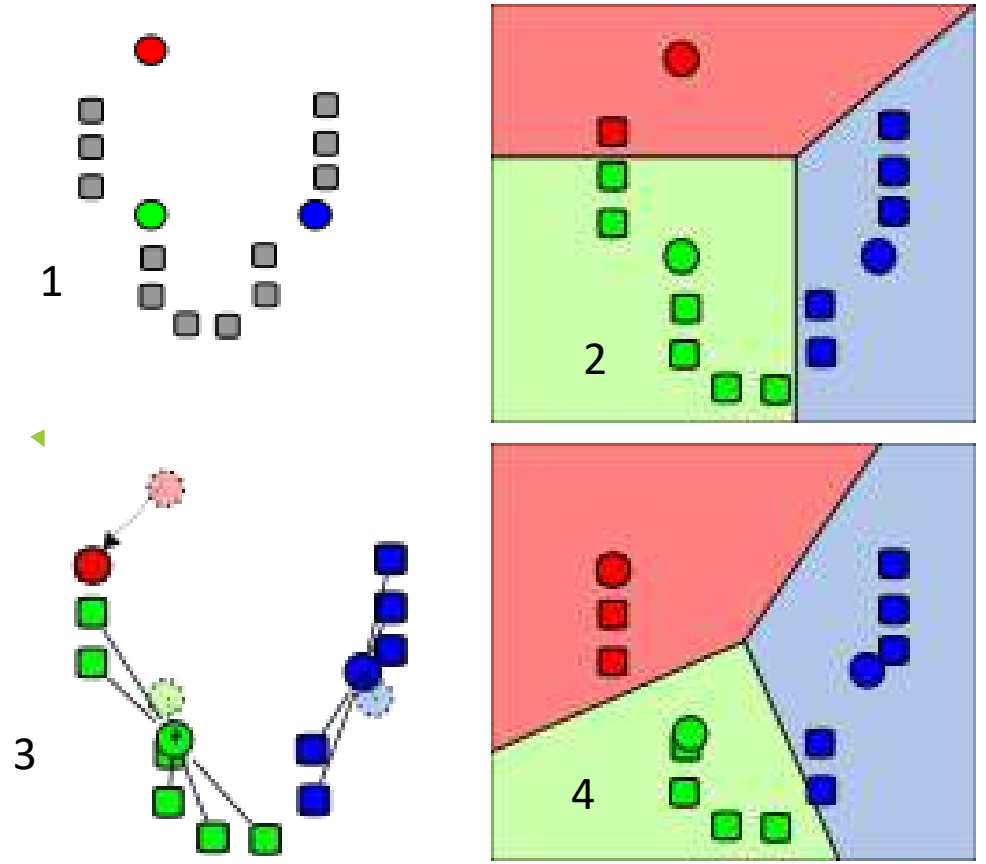
Кластерный анализ
(Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means)

Метод К средних – алгоритм (алгоритм для линейного векторного пространства:

Стандартный алгоритм был впервые предложен Стюартом Ллойдом из Bell Labs в 1957 году как метод импульсно-кодовой модуляции , хотя он не был опубликован в виде журнальной статьи до 1982 года. В 1965 году Эдвард В. Форги опубликовал, по сути, тот же метод, поэтому его иногда называют алгоритмом Ллойда – Форги

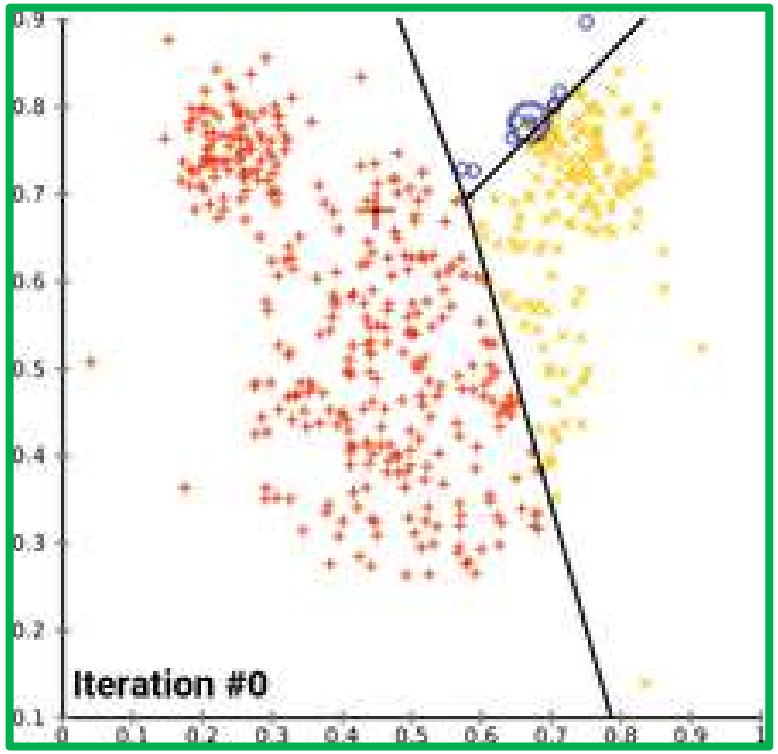
Алгоритм Ллойда (если число кластеров К – фиксировано, Y - метки) Наиболее распространенный алгоритм использует метод итеративного уточнения. Из-за его повсеместного распространения его часто называют «алгоритмом k-средних»; его также называют алгоритмом Ллойда , особенно в компьютерном сообществе. Иногда его также называют «наивными k-средними», потому что существуют гораздо более быстрые альтернативы

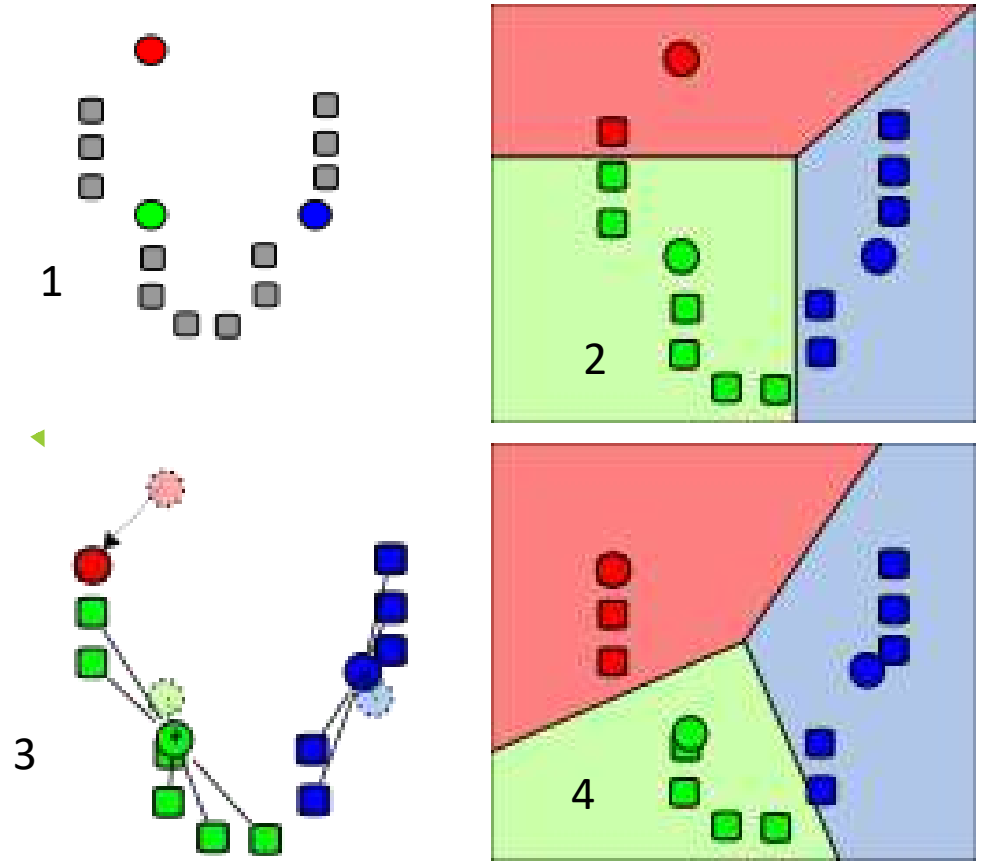
В SAS алгоритм к-средних реализует PROC FASTCLUS. Ей надо задать максимальное число кластеров и начальные точки – центроиды кластеров



Кластерный анализ
(Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means)

Метод К средних – алгоритм (алгоритм для линейного векторного пространства:

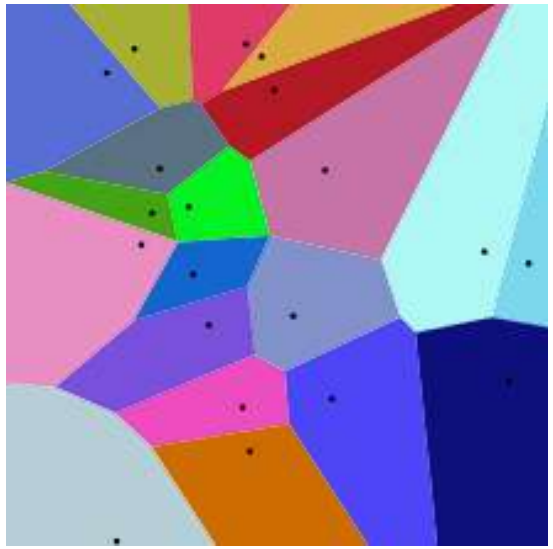




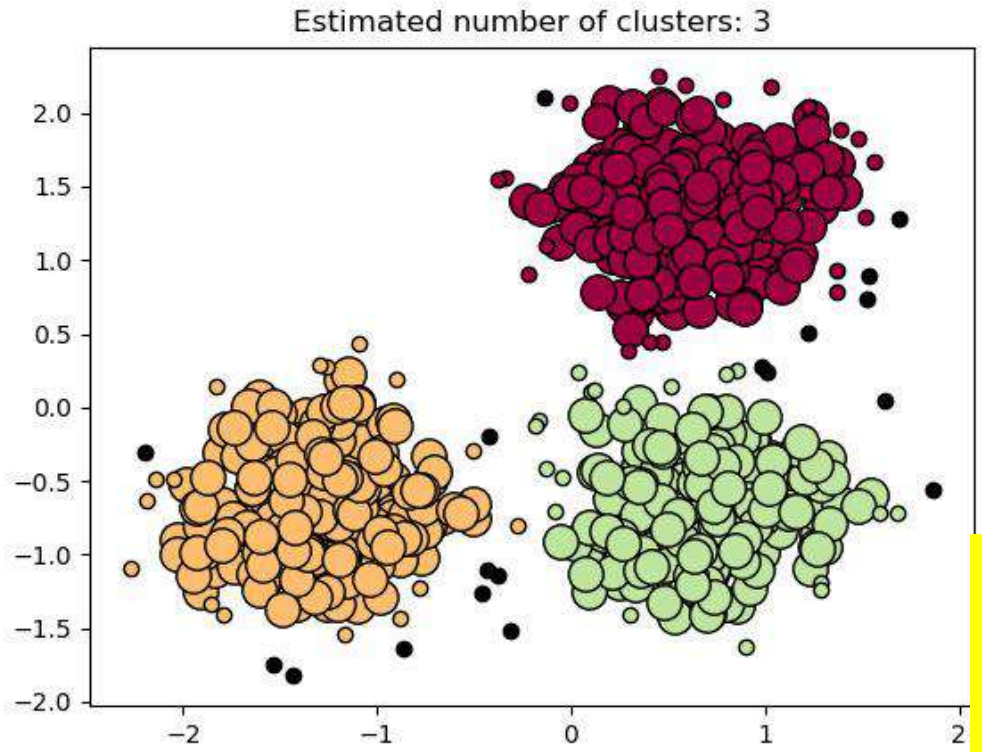
Кластерный анализ
(Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means)

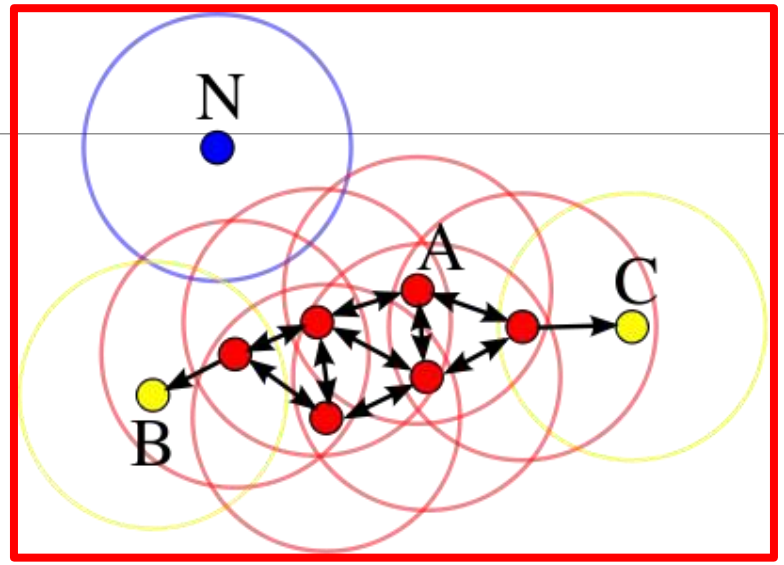
Метод К средних – алгоритм (алгоритм для линейного векторного пространства)

По сути, метод к –средних – разбиение пространства на области Вороного (построение диаграмм Вороного). В математике , диаграмма Вороного - это разбиение плоскости на области, близкие к каждому из заданного набора объектов. В простейшем случае это объекты - это просто конечное число точек на плоскости (называемых «семенами» (seeds), зародышами, узлами или генераторами). Для каждого «семени» существует соответствующая область, состоящая из всех точек плоскости, более близких к этому «семени», чем к любой другой. Эти области называются ячейками Вороного. Диаграмма Вороного для набора точек двойственна своей триангуляции Делоне



Кластерный анализ – суть метода DBSCAN (Cluster analysis)





DBSCAN – решает дополнительно задачу обнаружения выбросов, которые ни к одному кластеру не относятся

Определяем два параметра:

- радиус окрестности r
- Граничное число точек для окрестности M

Каждый элемент выборки может быть:

- корневой: имеющий плотную окрестность, т.е. в его окрестности заданного радиуса лежат не менее M точек
- Начинаем с точки A – в ее окрестности три точки, точка A – корневая
- Другие КРАСНЫЕ точки – тоже корневые
- граничный: не корневой, но в окрестности корневого, т.е. в его окрестности заданного радиуса лежат менее чем M точек, но они есть!
- Для точек B и C в окрестностях лежат всего по одной точке, поэтому они граничные
- шумовой (выброс): не корневой и не граничный - в его окрестности нет точек - это N – шумовая точка

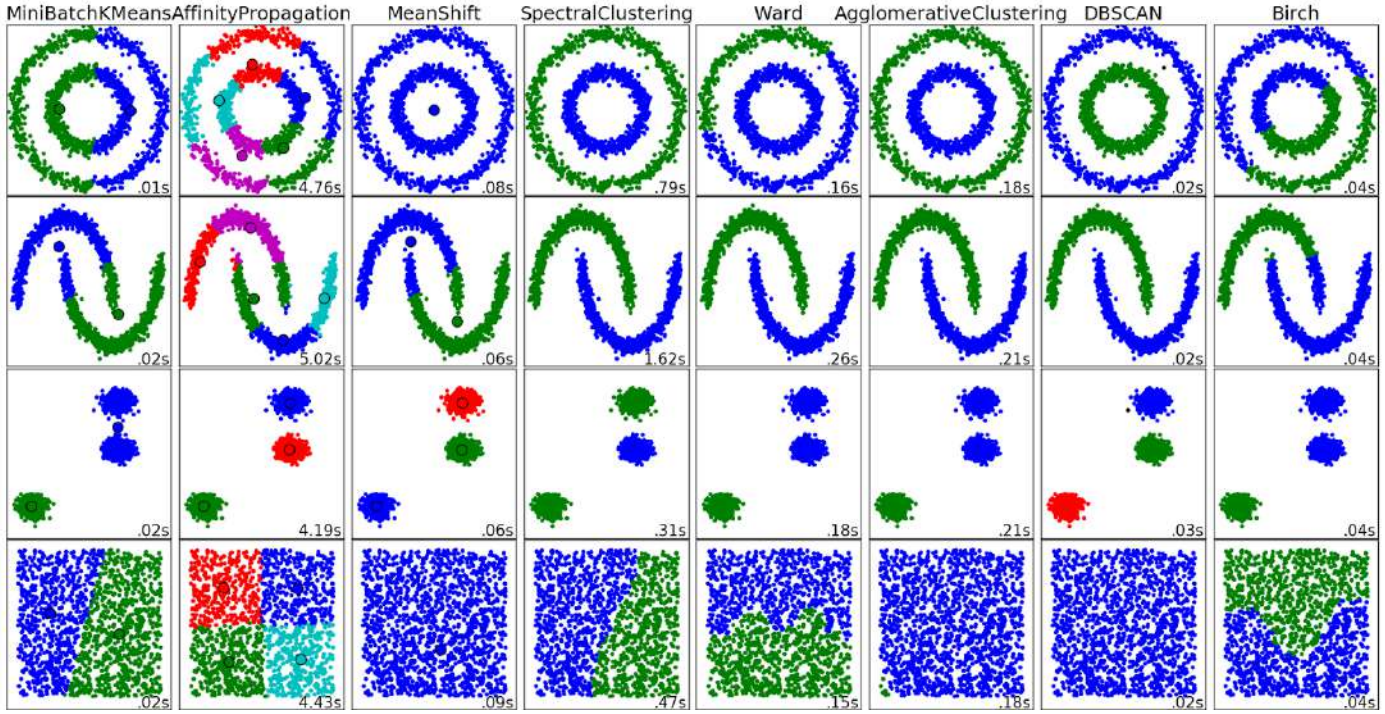
Кластерный анализ - (Cluster analysis)
что делать с экзотическими кластерами? DBSCAN!!!

вход: выборка $X^\ell = \{x_1, \ldots, x_\ell\}$; параметры ε и m;
выход: разбиение выборки на кластеры и шумовые выбросы;
$U := X^\ell$ — непомеченные; $a := 0$;
пока в выборке есть непомеченные точки, $U \neq \varnothing$:
  взять случайную точку $x \in U$;
  если $|U_\varepsilon(x)| < m$ **то**
    пометить x как, возможно, шумовой;
  иначе
    создать новый кластер: $K := U_\varepsilon(x)$; $a := a + 1$;
    для всех $x' \in K$, не помеченных или шумовых
      если $|U_\varepsilon(x')| \geqslant m$ **то** $K := K \cup U_\varepsilon(x')$;
      иначе пометить x' как граничный кластера K;
    $a_i := a$ для всех $x_i \in K$;
    $U := U \setminus K$;

Изначально все объекты выборки – неразмеченные
Внешний цикл – по всем объектам выборки
Берет точку случайным образом и проверяет для нее выполнение условия что это – корневая точка
Если это – не корневая точка, то сначала метим ее как шумовой объект (в итоге может оказаться шумовым или граничным для данного кластера)
Если же корневая – создаем новый кластер, к-во объектов увеличим на 1. Для всех объектов этого кластера помечаем, если они – корневыедля этого кластера
Затем для каждого кластера проверяем, не являются ли помеченные нами как шумовые объекты фактически граничными для каких-то из созданных кластеров. Внутренний цикл – разбухание выделенного кластера

В отличие от k-средних, DBSCAN определит количество кластеров!
DBSCAN работает, определяя, имеется ли минимальное количество достаточно близких точек, чтобы считаться частью одного кластера.
При этом определяются точки-выбросы!
DBSCAN очень чувствителен к масштабу
DBSCAN все же имеет два параметра, которые выбираются эвристически!

Кластерный анализ (Cluster analysis) Неиерархическая кластеризация - DBSCAN



Каждому из восьми алгоритмов кластеризации предлагаются четыре задачи кластеризации «экзотического» расположения объектов. Ни один из алгоритмов, кроме DBSCAN, не решает ВСЕ ЧЕТЫРЕ задачи правильно!

Density-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise (**DBSCAN**)

Метод, основанный на оценке плотности – DBSCAN – позволяет кластеризацию не только выпуклых кластеров (в отличие от метода К-средних, формирующих **выпуклые** кластеры вокруг центроидов). Это выявление участков наибольшей плотности по сравнению с другими вариантами

Метод DBSCAN позволяет определить и исключить из кластеризации шумы и выбросы

Спасибо за внимание!

Лекция -окончена

Методы интеллектуального анализа данных

(ДЛЯ МАГИСТРОВ ИВТ 1 года)

ПРЕПОДАВАТЕЛЬ:

СТЕРИН АЛЕКСАНДР МАРКОВИЧ,

ДОКТОР ФИЗ.-МАТ. НАУК, СТ. Н. СОТР., ПРОФЕССОР ОИКС (О)

ТЕЛ. (484)3974658,

ЭЛ. ПОЧТА: AMSTERIN@OIATE.RU; ALEX.STERIN@GMAIL.COM;

STERIN@METEO.RU

Кластерный анализ (Cluster analysis)

Есть ряд подходов, которые можно назвать АГРЕГАЦИЕЙ ДАННЫХ.

Можно группировать (агрегировать) переменные в факторы или компоненты, формировать обобщенные переменные, но при этом неизменным остается множество объектов (наблюдений)

- Метод (анализ) главных компонент (PCA)
- Факторный анализ (Factor analysis)

Можно группировать (агрегировать) объекты (наблюдения), формируя группы (кластеры) схожих в пространстве переменных объектов

- Cluster analysis

В анализе дисперсий (ANOVA) объекты заранее отнесены к известным группам

В кластерном анализе группы формируются из имеющихся объектов исходя из схожести объектов в пространстве переменных (группы не заданы априори)

Кластерный анализ (Cluster analysis)

Примеры практических применений кластеризации:

Сегментация пользователей услуги

Сегментация клиентов банка (не путать с задачей решения давать-не давать кредит)

Кластеризация источников сетевого траффика

Спам-фильтрация электронной почты (группировка сообщений в соответствии с результатами анализа их разных частей – отправитель, получатели, тема, особенности содержания, и т.д.)

Городское планирование –кластеризация домов, квартир

Кластеризация документов

Анализ рынка недвижимости

В биологии – типизация объектов

Кластерный анализ (Cluster analysis)

Общие положения:

Есть выборка из L многомерных объектов – обучающая выборка:

$X_1, X_2, \ldots, X_l, (i=1,\ldots,L)$

Y_i –отклик, зависимая переменная, - отсутствует!

Надо разбить объекты на группы так, чтобы внутри каждой группы были максимально схожие между собой объекты, и при этом максимально не схожие с объектами других групп

Т.е., восстановить для каждого объекта обучающей выборки значение Y_i

Задача обучения БЕЗ УЧИТЕЛЯ **(UNSUPERVISED LEARNING)** – более общая, но обратное верно - всякая кластеризация – это **UNSUPERVISED LEARNING**

Имеются разновидности: **ЧАСТИЧНОЕ ОБУЧЕНИЕ (SEMI-SUPERVISED LEARNING)** – оставляем на будущее

Кластеризация: постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка;

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров,

$a\colon X \to Y$ — алгоритм кластеризации,

такие, что:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

Это задача *обучения без учителя* (unsupervised learning).

Методы кластеризации

Три основные группы алгоритмов:

- **Алгоритмы к средних (k- means) – неиерархическая**
- **Алгоритмы С-средних с нечеткими множествами (C-Means Fuzzy Clustering) – рассматриваются редко**
- **Иерархические алгоритмы (Hierarchial)!!!)**
- **Алгоритмы, основанные на плотности (Density Based Clustering)**

В способах и алгоритмах решения задач кластеризации –очень много эвристики!

- **На сколько кластеров проводить разбиение?**
- **Каков критерий качества кластеризации – целевая функция, которую нужно оптимизировать?**
- **В какой метрике проводить расчет расстояний между многомерными объектами – от этого зависит результат дальнейших обобщений!**

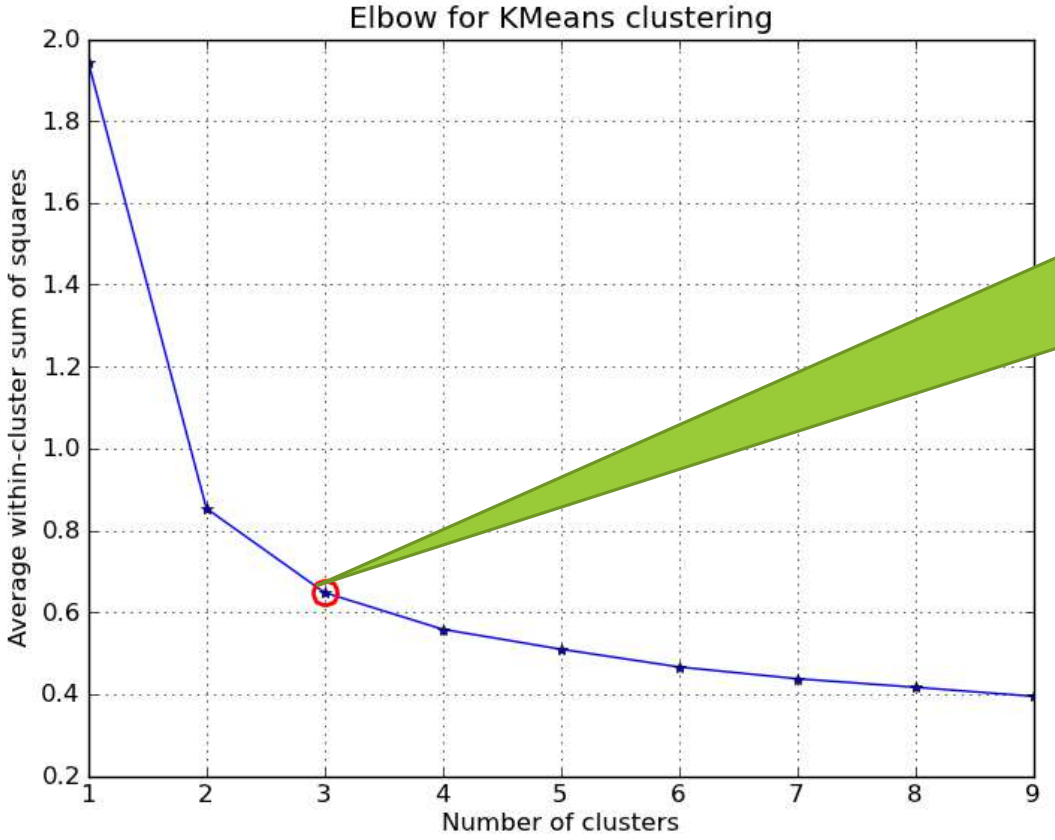
Кластерный анализ - сколько кластеров должно быть для кластеризации по К-средним? (Cluster analysis)

Идея. Выбрать критерий качества кластеризации и построить его значение для $K = 1, 2, \ldots$

- средняя сумма квадратов расстояния до центроида
- средний диаметр кластера

Метод «локтя» - Elbow method

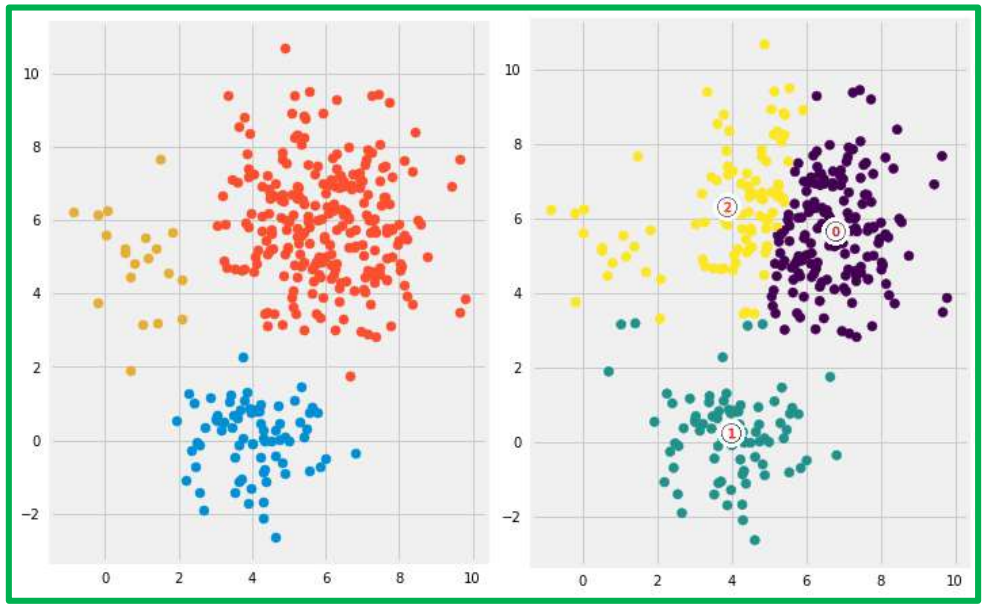
Но это – чистая эвристика!!!

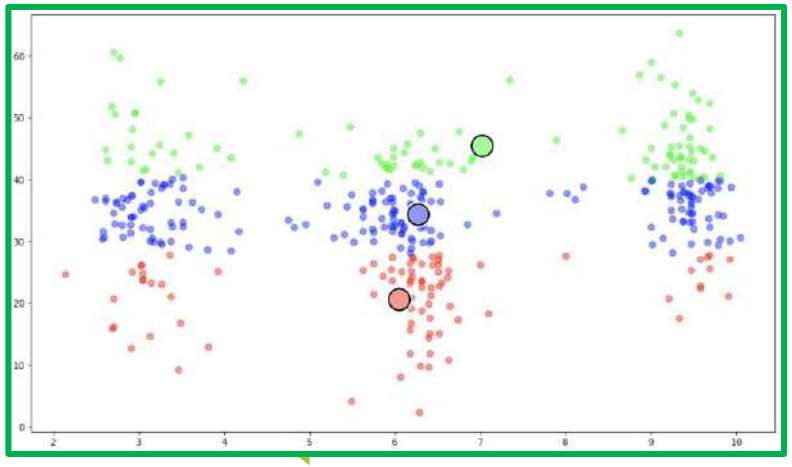


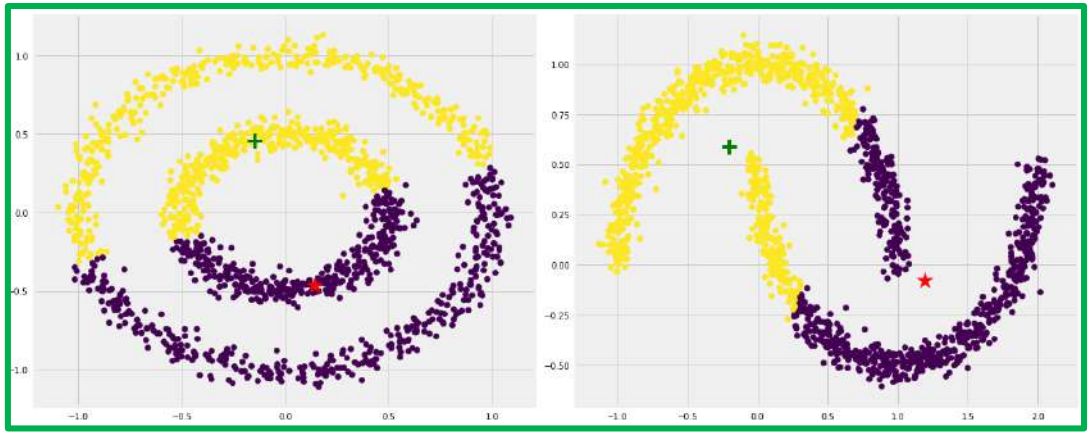
Точка излома для кривой зависимости

Кластерный анализ (Cluster analysis)
Неиерархическая кластеризация - Алгоритм k-средних (k-means): примеры неудачной кластеризации

Причина: неудачное начальное приближение или существенная негауссовость кластеров

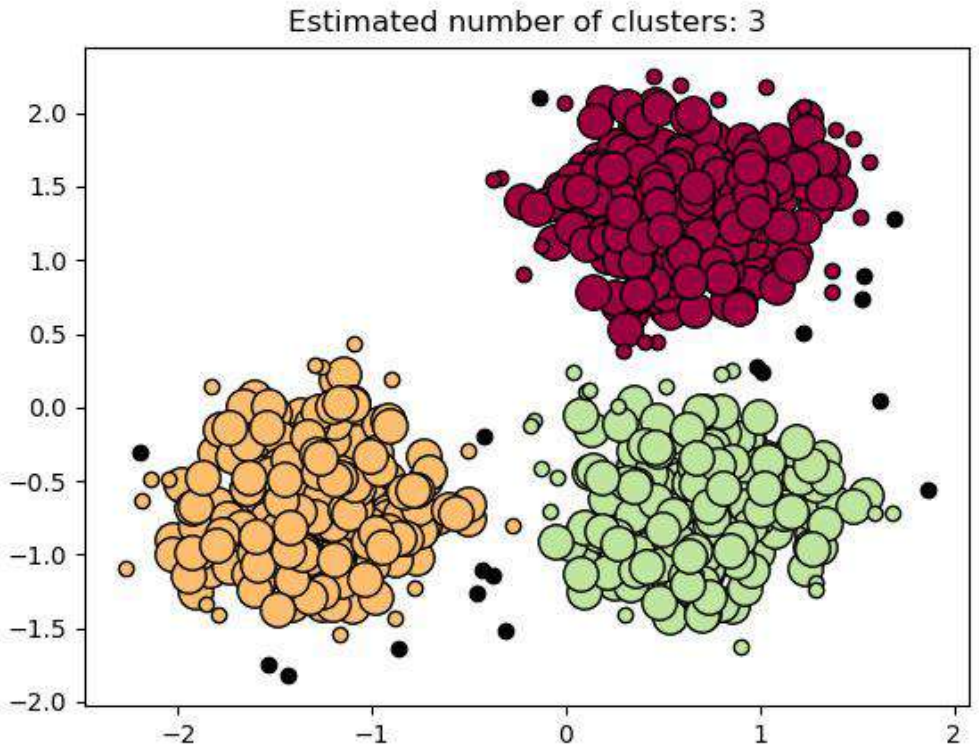




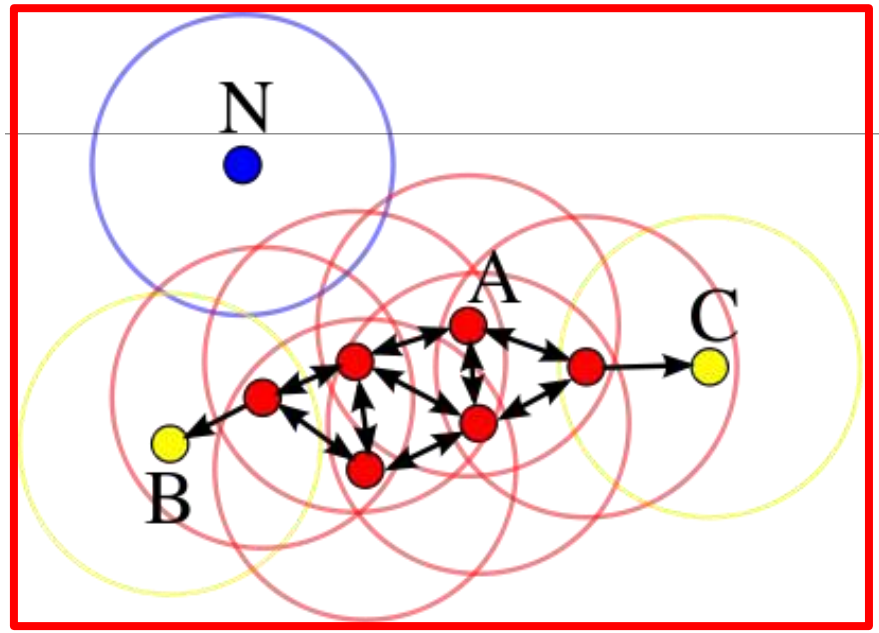


Многих неудачных ситуаций, имеющих место при использовании метода к-средних, можно избежать, используя методы, основанные на анализе плотности, например, DBSCAN

Кластерный анализ – суть метода DBSCAN (Cluster analysis)



DBSCAN – решает дополнительно задачу обнаружения выбросов, которые ни к одному кластеру не относятся



Определяем два параметра:

- радиус окрестности r
- Граничное число точек для окрестности M

Каждый элемент выборки может быть:

- корневой: имеющий плотную окрестность, т.е. в его окрестности заданного радиуса лежат не менее M точек
- граничный: не корневой, но в окрестности корневого, т.е. в его окрестности заданного радиуса лежат менее чем M точек, но они есть!
- Для точек B и C в окрестностях лежат всего по одной точке, поэтому они граничные
- шумовой (выброс): не корневой и не граничный - в его окрестности нет точек - это N

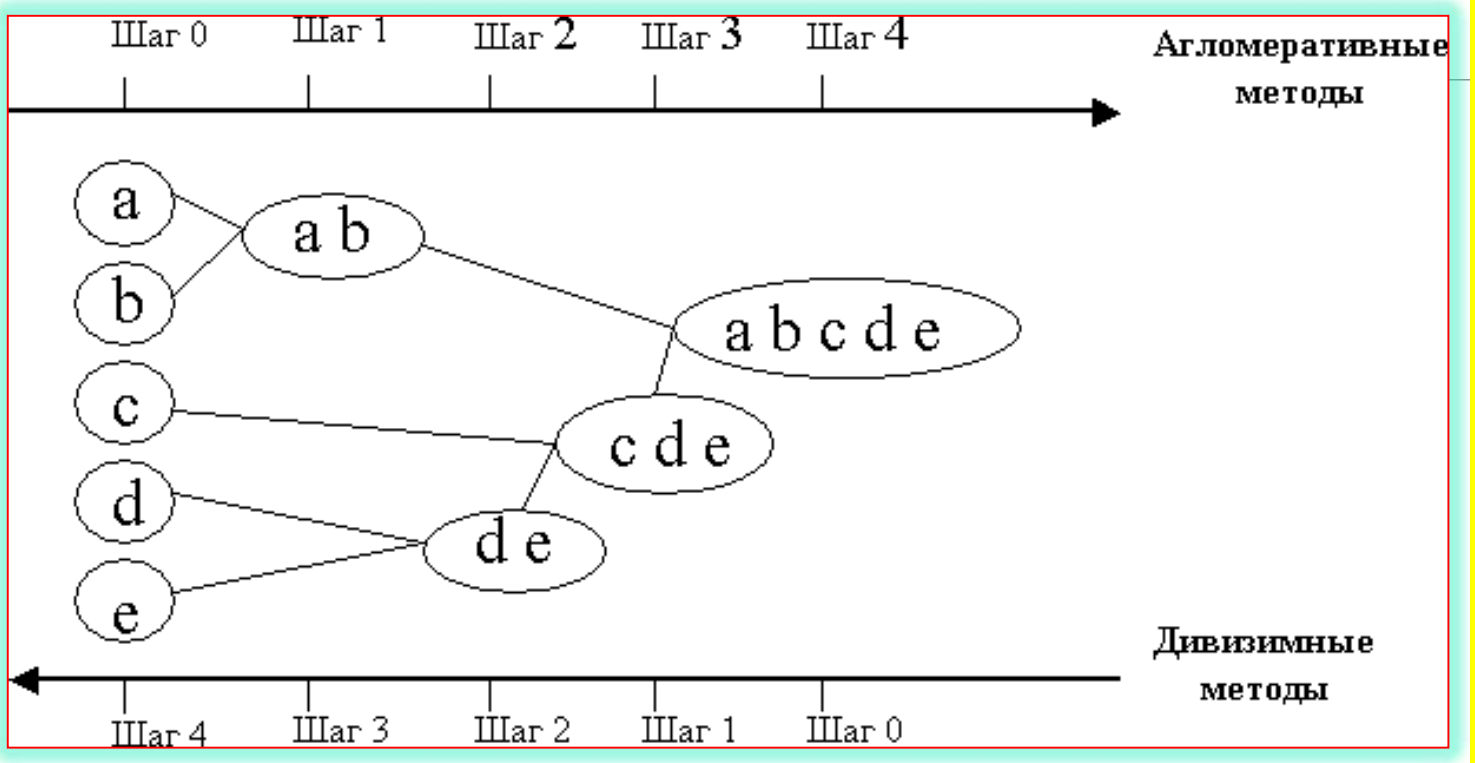
Кластерный анализ - (Cluster analysis)
что делать с экзотическими кластерами? DBSCAN!!!

вход: выборка $X^\ell = \{x_1, \ldots, x_\ell\}$; параметры ε и m;
выход: разбиение выборки на кластеры и шумовые выбросы;
$U := X^\ell$ — непомеченные; $a := 0$;
пока в выборке есть непомеченные точки, $U \neq \varnothing$:
  взять случайную точку $x \in U$;
  если $|U_\varepsilon(x)| < m$ **то**
    пометить x как, возможно, шумовой;
  иначе
    создать новый кластер: $K := U_\varepsilon(x)$; $a := a + 1$;
    для всех $x' \in K$, не помеченных или шумовых
      если $|U_\varepsilon(x')| \geqslant m$ **то** $K := K \cup U_\varepsilon(x')$;
      иначе пометить x' как граничный кластера K;
    $a_i := a$ для всех $x_i \in K$;
    $U := U \setminus K$;

Изначально все объекты выборки – неразмеченные
Внешний цикл – по всем объектам выборки
Берет точку случайным образом и проверяет для нее выполнение условия что это – корневая точка
Если это – не корневая точка, то сначала метим ее как шумовой объект (в итоге может оказаться шумовым или граничным для данного кластера)
Если же корневая – создаем новый кластер, к-во объектов увеличим на 1. Для всех объектов этого кластера помечаем, если они – корневыедля этого кластера
Затем для каждого кластера проверяем, не являются ли помеченные нами как шумовые объекты фактически граничными для каких-то из созданных кластеров. Внутренний цикл – разбухание выделенного кластера

В отличие от k-средних, DBSCAN определит количество кластеров!
DBSCAN работает, определяя, имеется ли минимальное количество достаточно близких точек, чтобы считаться частью одного кластера.
При этом определяются точки-выбросы!
DBSCAN очень чувствителен к масштабу
DBSCAN все же имеет два параметра, которые выбираются эвристически!

Кластерный анализ (Cluster Analysis)
Иерархическая кластеризация



ДВЕ ПРОТИВОПОЛОЖНЫЕ СТРАТЕГИИ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

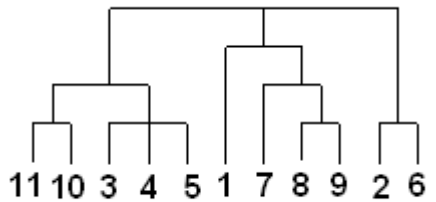
Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизивные (делимые) методы (DIvisive ANAlysis, DIANA)

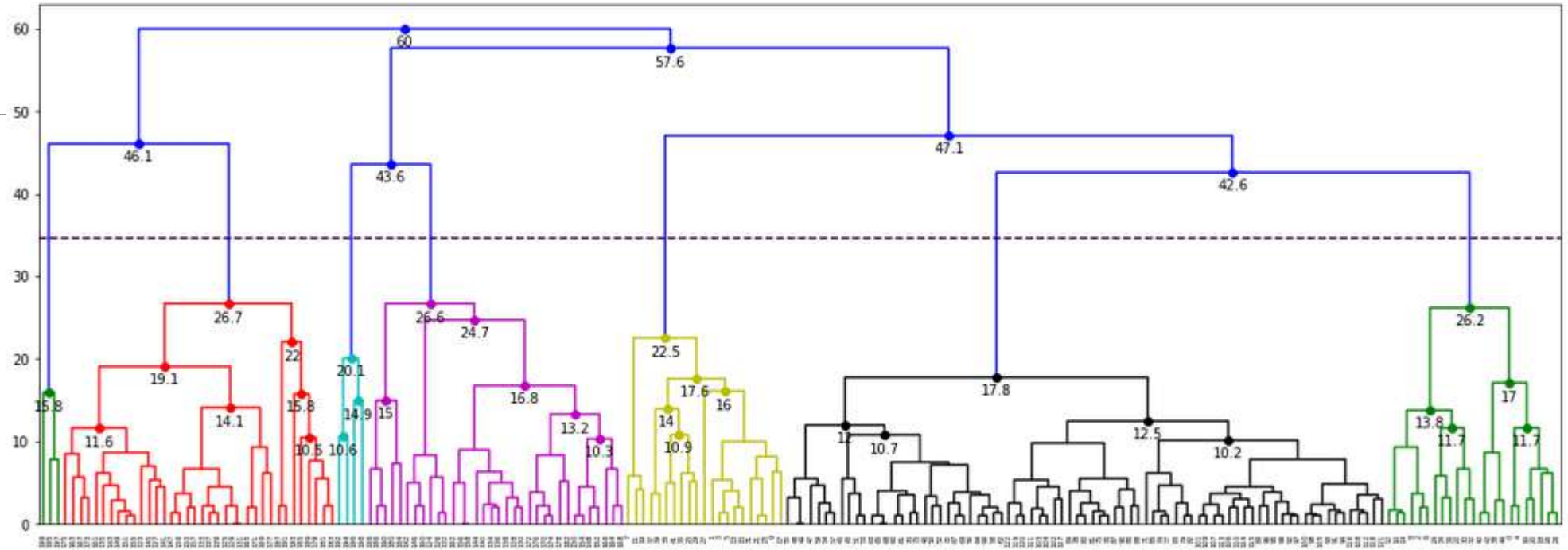
Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров



ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

дендрограмма – неискаженное представление кластерной структуры данных



- Кластеры группируются вдоль горизонтальной оси
- По вертикальной оси откладываются расстояния – значения критерия иерархической кластеризации, например R_t
- Расстояния возрастают, линии нигде не пересекаются
- Верхние уровни различимы лучше, чем нижние
- Уровень отсечения определяет число кластеров

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ:
Алгоритм иерархической кластеризации (причем, это **агломеративная** кластеризация)

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967):

итеративный пересчет расстояний R_{UV} между кластерами при слиянии кластеров U;V.

$C_1 := \{\{x_1\}, \ldots, \{x_\ell\}\}$ — все кластеры 1-элементные;
$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;
для всех $t = 2, \ldots, \ell$ (t — номер итерации):
  найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV};
  слить их в один кластер:
  $W := U \cup V$;
  $C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;
  для всех $S \in C_t$
    вычислить R_{WS} по формуле Ланса-Уильямса:
    $R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma |R_{US} - R_{VS}|$;

Кластеры U,V сливаются в один кластер W
Как при этом меняется расстояние до произвольного кластера S?
При этом известны расстояния кластера S до кластеров U и V перед их слиянием
Тогда может быть использована формула Ланса-Уильямса

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Кластеры U,V сливаются в один кластер W
Как при этом меняется расстояние от W до произвольного кластера S (исходные расстояния до слияния - R_{US}, R_{VS} – известны)?

$$R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma |R_{US} - R_{VS}|$$

Формула Ланса-Уильямса:
Для большинства метрик, в которых вычисляются расстояния между кластерами, слияние двух кластеров позволяет пересчитать расстояния после слияния, используя 4 коэффициента!

Два момента:
Первое. Как считать расстояние между кластерами?
Второе: Как, на основе расстояния, строить иерархическое дерево?
Существует много вариантов расчета расстояний между кластерами.
Уорда (WARD) критерий минимальной дисперсии сводит к минимуму общую внутрикластерную дисперсию. **Чтобы реализовать этот метод, на каждом шаге найдите пару кластеров, которая приводит к минимальному увеличению общей внутрикластерной дисперсии после слияния.** Это увеличение представляет собой взвешенный квадрат расстояния между центрами кластеров. На начальном этапе все кластеры являются синглетонами (кластерами, содержащими одну точку). Чтобы применить рекурсивный алгоритм к этой целевой функции , начальное расстояние между отдельными объектами должно быть (пропорционально) квадрату евклидова расстояния.
Т.е. метод Уорда еще дает указание, как строить иерархическую кластеризацию!

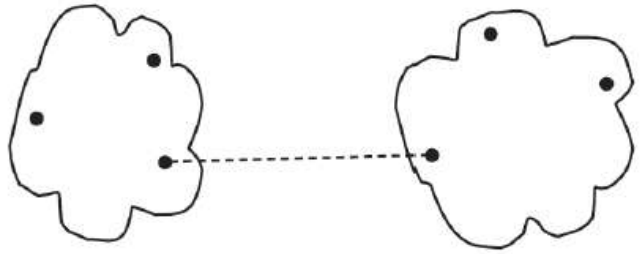
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

частные случаи формулы Ланса-Уильямса: как по-разному можно считать расстояния между кластерами

1. Расстояние ближнего соседа:

$$R^{б}_{WS} = \min_{w \in W, s \in S} \rho(w, s);$$

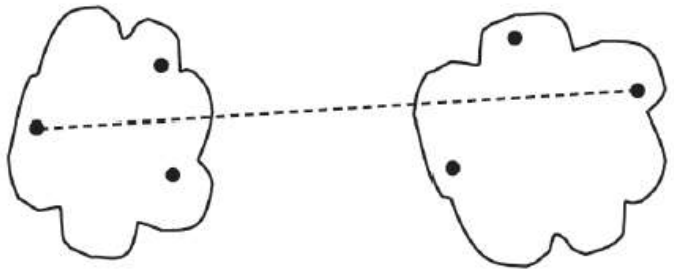
$$\alpha_U = \alpha_V = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{д}_{WS} = \max_{w \in W, s \in S} \rho(w, s);$$

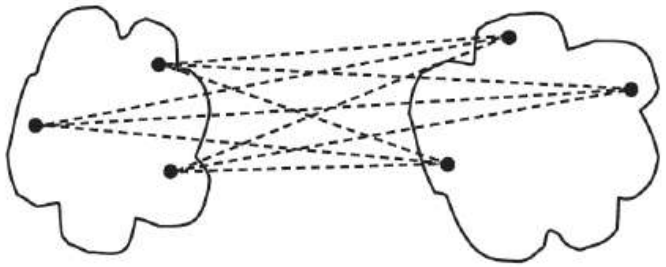
$$\alpha_U = \alpha_V = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}.$$



3. Групповое среднее расстояние:

$$R^{г}_{WS} = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s);$$

$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|}, \quad \beta = \gamma = 0.$$



ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

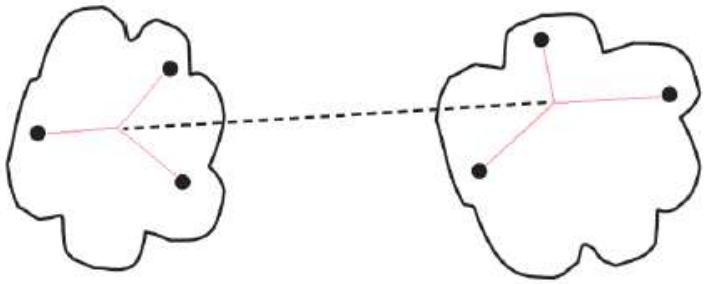
частные случаи формулы Ланса-Уильямса: как по-разному можно считать расстояния между кластерами

4. Расстояние между центрами:

$$R^{\text{ц}}_{WS} = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V, \quad \gamma = 0.$$



5. Расстояние Уорда:

$$R^{\text{у}}_{WS} = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}, \quad \alpha_V = \frac{|S|+|V|}{|S|+|W|}, \quad \beta = \frac{-|S|}{|S|+|W|}, \quad \gamma = 0.$$

Формула Ланса-Уильямса НЕ РАБОТАЕТ с расстояниями между отдельными объектами, она работает с расстояниями между кластерами, слегка их подправляя при слияниях!

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Диаграмма вложения – полезна только на маленьких данных, дендрограмма – полезна на любых объемах данных!

Всегда объединяем два самых близких кластера

На начальном этапе объединяем два самых близких единичных объекта

Смотрим, как растет целевая функция, и как идет процесс агломерации (для конкретного способа расчета расстояния) – видно на дендрограмме, т.е. возможна визуализация кластерной структуры!

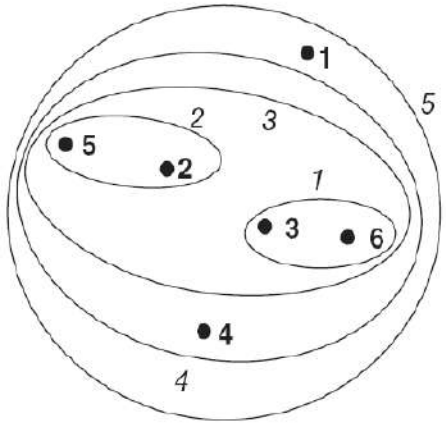
Разные расстояния – получаются разные диаграммы и разные дендрограммы

Находят пару, на которой большой разрыв, и ставят отсечение

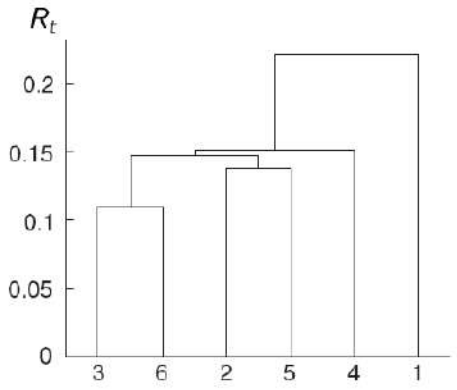
R_t –расстояние для номера t итерации при объединениях R_{uv}

1. Расстояние ближнего соседа:

Диаграмма вложения

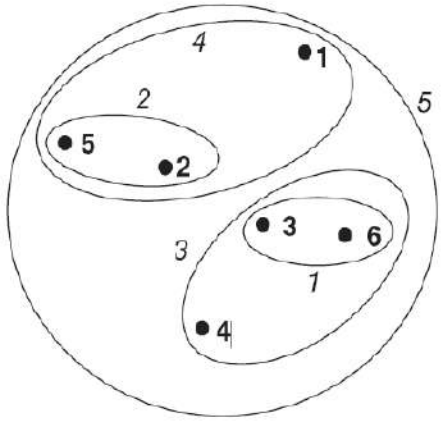


Дендрограмма

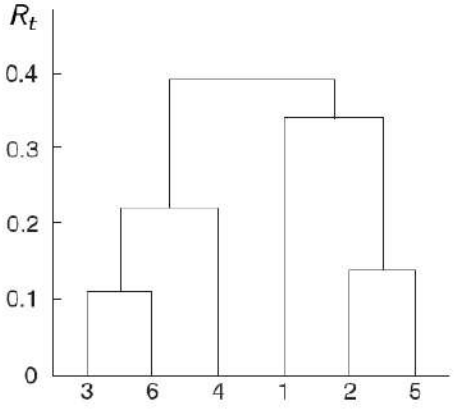


2. Расстояние дальнего соседа:

Диаграмма вложения



Дендрограмма



ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Диаграмма вложения – полезна только на маленьких данных, дендрограмма – полезна на любых объемах данных!

Всегда объединяем два самых близких кластера

На начальном этапе объединяем два самых близких единичных объекта

Смотрим, как растет целевая функция, и как идет процесс агломерации (для конкретного способа расчета расстояния)– видно на дендрограмме, т.е. возможна визуализация кластерной структуры!

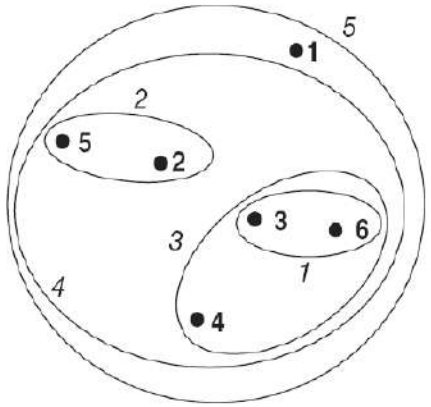
Разные расстояния – получаются разные диаграммы и разные дендрограммы

Находят пару, на которой большой разрыв, и ставят отсечение

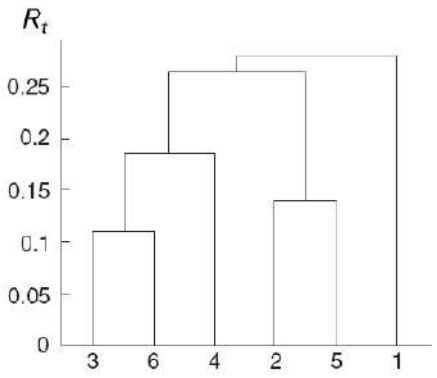
R_t –расстояние для номера t итерации при объединениях R_{uv}

3. Групповое среднее расстояние:

Диаграмма вложения

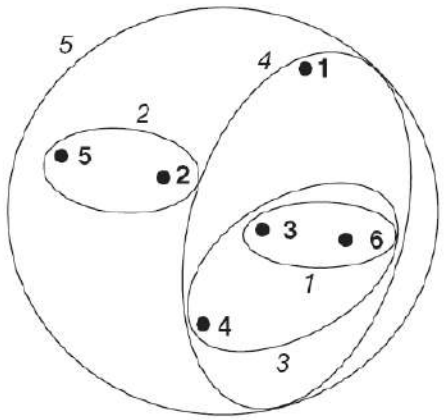


Дендрограмма

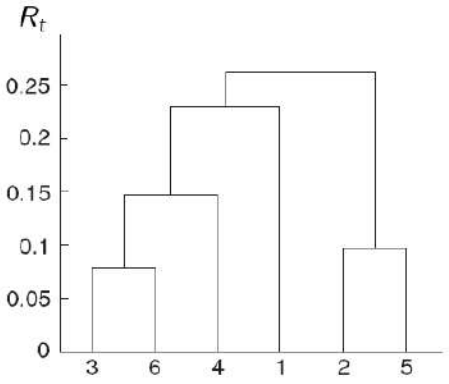


5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ
Основные свойства иерархической кластеризации

Монотонность: дендрограмма не имеет самопересечений, при каждом слиянии - итерации расстояние между объединяемыми кластерами только увеличивается:

$$R_2 \leqslant R_3 \leqslant \ldots \leqslant R_\ell.$$

- *Сжимающее расстояние*: $R_t \leqslant \rho(\mu_U, \mu_V), \ \forall t.$
- *Растягивающее расстояние*: $R_t \geqslant \rho(\mu_U, \mu_V), \ \forall t$

Сравнивается Rt с расстояниями между ЦЕНТРАМИ кластеров
Растягивающее расстояние – предпочтительнее, в том числе для определения небольшого числа кластеров

Теорема (Миллиган, 1979)

Кластеризация монотонна, если выполняются условия

$$\alpha_U \geqslant 0, \ \ \alpha_V \geqslant 0, \ \ \alpha_U + \alpha_V + \beta \geqslant 1, \ \ \min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0.$$

$R^{ч}$ не монотонно; $R^{б}$, $R^{д}$, $R^{г}$, $R^{у}$ — монотонны.

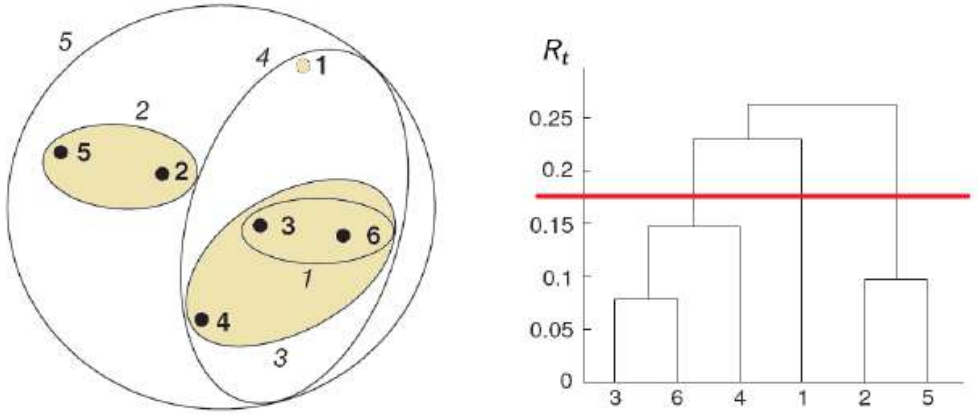
$R^{б}$ — сжимающее; $R^{д}$, $R^{у}$ — растягивающие;

Это вполне логично!!!!

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Монотонность: дендрограмма не имеет самопересечений, при каждом слиянии расстояние между объединяемыми кластерами только увеличивается

- рекомендуется пользоваться расстоянием Уорда $R^{У}$;
- обычно строят несколько вариантов и выбирают лучший визуально по дендрограмме;
- определение числа кластеров — по максимуму $|R_{t+1} - R_t|$, тогда результирующее множество кластеров $:= C_t$.



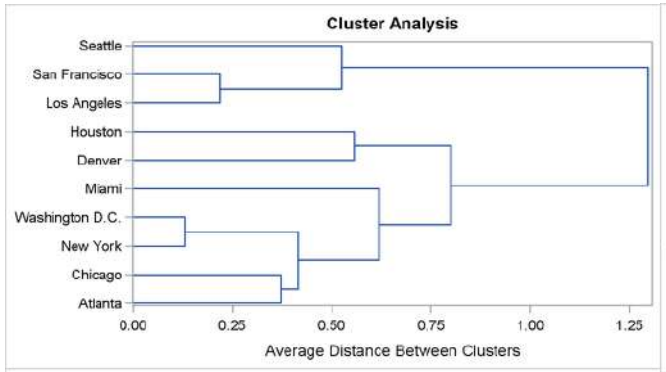
Расстояние Уорда – всегда работает хорошо!

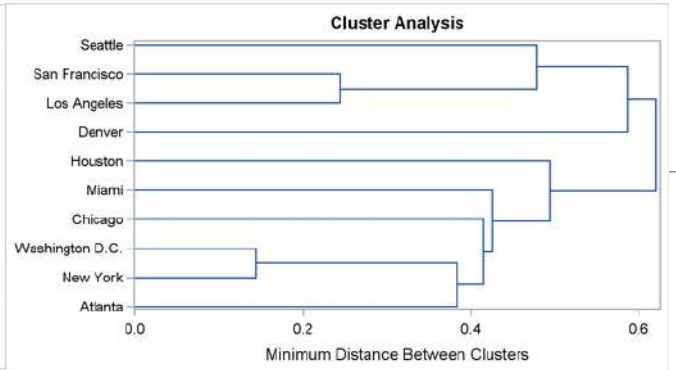
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ: реализация в SAS

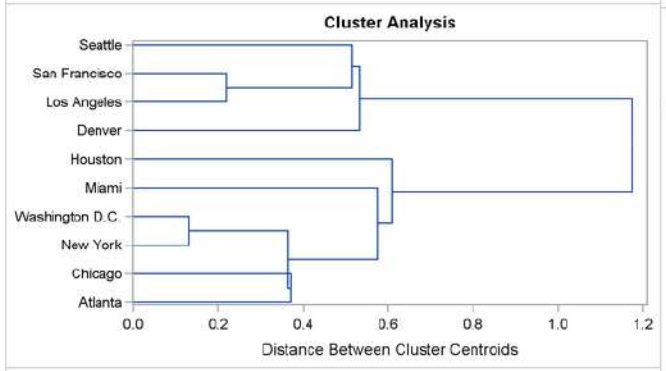
```
title2 'By Ward''s Method’;
ods graphics on;
proc cluster data=iris method=ward print=15 ccc pseudo;
 var petal: sepal:;
copy species;
 run;
 proc tree noprint ncl=3 out=out;
copy petal: sepal: species;
run;
```

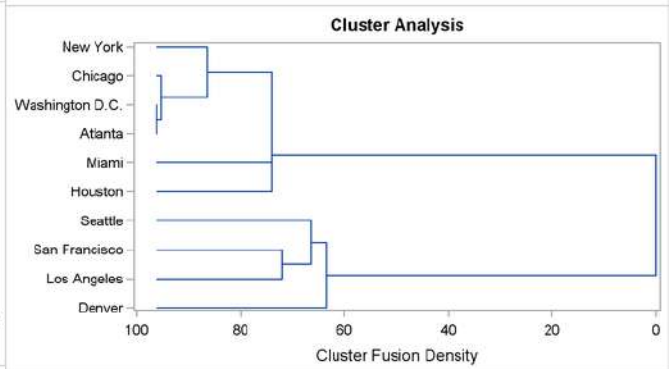
Кластерный анализ (Cluster Analysis)

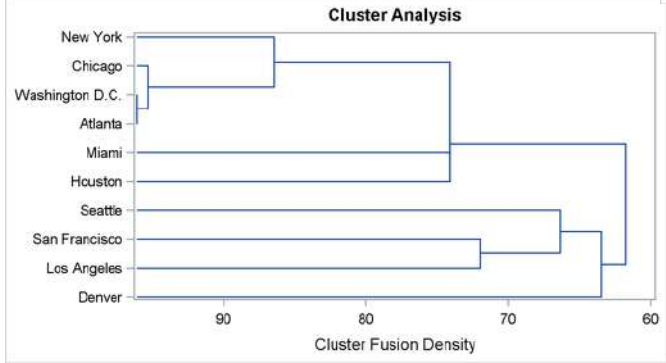
Иерархическая кластеризация –дендрограммы с различными критериями для объединения и различными метриками

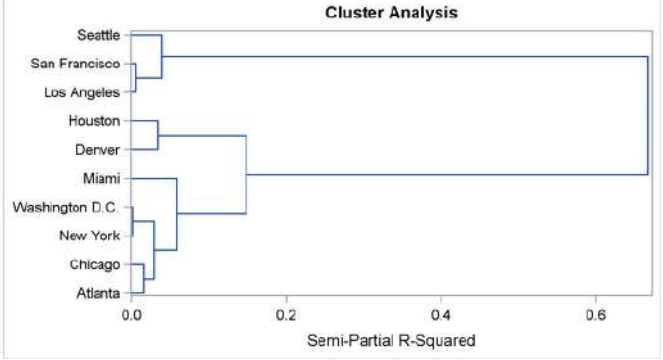


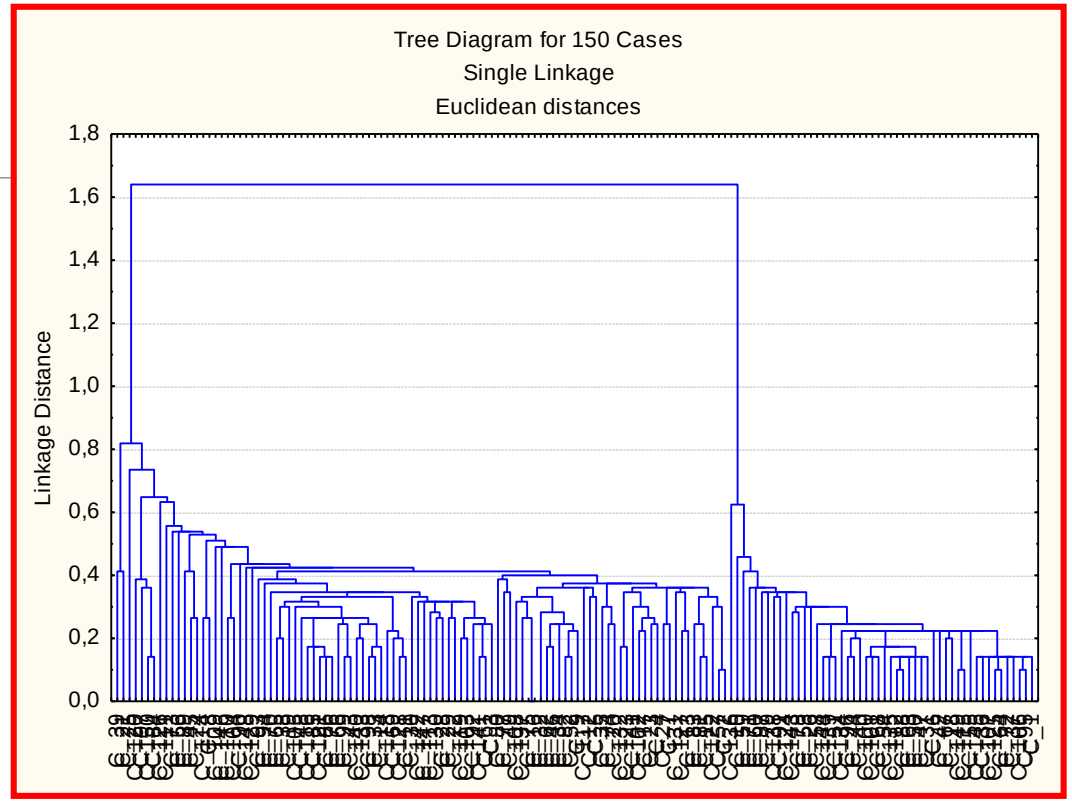












Могут быть очень насыщенные дендрограммы
Это дендрограмма для Ириса Фишера – кластеризация 150 наблюдений

Спасибо за внимание!

Лекция -окончена