МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский ядерный университет «МИФИ»
Обнинский институт атомной энергетики –
филиал федерального государственного автономного образовательного учреждения высшего образования
«Национальный исследовательский ядерный университет «МИФИ»
(ИАТЭ НИЯУ МИФИ)

Одобрено на заседании УМС
ИАТЭ НИЯУ МИФИ Протокол
от 30.08.2022 № 2-8/2022

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

МАШИННОЕ ОБУЧЕНИЕ

название дисциплины

для студентов специальности/направления подготовки

09.04.01 Информатика и вычислительная техника

специализация/профиль:
Большие данные и машинное обучение в задачах атомной энергетики

Форма обучения: очная

г. Обнинск 2022 г.

Программа составлена в соответствии с требованиями Федерального государственного образовательного стандарта высшего профессионального образования по направлению подготовки (специальности) 09.03.01 «Информатика и вычислительная техника».

Программу составил: ассистент Отделения Института ИКС(О) Д.А. Распопов

Рецензент: _____ __вице-президент Обнинской Торгово-промышленной палаты, канд. физ.-мат. наук В.Т. Радюхин

Рабочая программа одобрена на заседании Отделения ИКС

утверждена «_____» _____ 201__ года, протокол № _____.

Начальник Отделения ИКС _____ С.О. Старков
(подпись) *(И.О. Фамилия)*

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения ООП магистратуры обучающийся должен овладеть следующими результатами обучения по дисциплине:

Коды компетенций	**Результаты освоения ООП. Содержание компетенций**	**Перечень планируемых результатов обучения по дисциплине**
СПК-1	Способен использовать и развивать методы научных исследований и инструментарий в области интеллектуального анализа данных	Знать: • Специфику машинного обучения, связанную с проблемами вычислительной эффективности и переобучения • типологию задач обучения по прецедентам • основные задачи обучения по прецедентам: классификация, кластеризация, регрессия, понижение размерности, и методы их решения Уметь: • Применять технологии, методы и инструментальные средства обработки больших данных • Применять на практике основные математические модели в области специализации применять перспективные методы индуктивного обучения, анализировать достоинства, недостатки и границы применимости используемых методов Владеть:

		• Языком программирования Python • Инструментами data science – jupyter notebook, jupyter lab, PyCharm. • Python–фреймворками и библиотеками анализа данных, их визуализации и машинного обучения – Pandas, Numpy, Sklearn.

2. Место дисциплины в структуре ООП магистратуры

Дисциплина реализуется в рамках вариативной части.

Для освоения дисциплины необходимы знания из области математических дисциплин и программирования на базовом уровне. Дисциплина «Машинное обучение» реализуется во втором и третьем семестре в рамках вариативной части дисциплин (модулей) Блока 1 и является базовой для освоения последующей по учебному плану практики.

Дисциплина изучается на 1 курсе магистратуры во 2 семестре.

3. Объем дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам занятий) и на самостоятельную работу обучающихся

Общая трудоемкость (объем) дисциплины составляет 5 зачетных единиц (з.е.), 180 академических часов.

3.1. Объём дисциплины по видам учебных занятий (в часах)

Вид учебной работы	Всего часов	Семестры			
		2			
Аудиторные занятия (всего)	84	84			
В том числе:	-	-	-	-	-
Практические занятия	16	16			
Семинары	нет	нет			
Лабораторные работы	16	16			
В том числе:	-	-	-	-	-
интерактивные формы обучения (лекции)	16	16			
интерактивные формы обучения (практические занятия/семинары)	нет	нет			
Самостоятельная работа (всего)	96	96			
В том числе:	-	-	-	-	-
Учебный проект (работа)	нет	нет			
Расчетно-графические работы	нет	нет			
Реферат	нет	нет			
Вид промежуточной аттестации (зачет, экзамен)	Экзамен	Экзамен			
ОБЩАЯ ТРУДОЕМКОСТЬ					
час	180	180			
зач.ед.	5	5			

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины и трудоемкость по видам учебных занятий (в академических часах)

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
Раздел 1. Введение Представление и разбор теоретической темы, решение задач по темам: 1. Постановка задач обучения по прецедентам. Объекты и признаки. 2. Типы шкал: бинарные, номинальные, порядковые, количественные. 3. Типы задач: классификация, регрессия, прогнозирование, кластеризация. Примеры прикладных задач. Основные понятия.	2	2		
Раздел 2. Обучение без учителя Представление и разбор теоретической темы, решение задач по темам: 1. Кластеризация Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связных компонент. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ. Функционалы качества кластеризации. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних	6	2	2	2

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
(k-means). 2. Сети Кохонена. Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена. Сети встречного распространения, их применение для кусочнопостоянной и гладкой аппроксимации функций. 3. Таксономия. Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи. Алгоритм построения дендрограммы. Определение числа кластеров. Свойства сжатия/растяжения, монотонности и редуктивности.				
Раздел 3. Метрические методы классификации Представление и разбор теоретической темы, решение задач по темам: 1. Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. 2. Метод окна Парзена. 3. Метрические методы классификации в задаче восстановления регрессии. 4. Обнаружение выбросов.	6	2	2	2

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
Раздел 4. Логические методы классификации и решающие деревья Представление и разбор теоретической темы, решение задач по темам: 1. Понятия закономерности и информативности Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности. 2. Асимптотическая эквивалентность статистического и энтропийного определения. Сравнение областей эвристических и статистических закономерностей. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости. Градиентный алгоритм синтеза конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Бинаризация признаков. 3. Решающие деревья для задач классификации и регрессии.	6	2	2	2
Раздел 5. Градиентные линейные методы классификации Представление и разбор теоретической темы, решение задач по темам:	6	2	2	2

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
1. Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия. 2. Метод стохастического градиента и частные случаи: адаптивный линейный элемент ADALINE, персептрон Розенблатта, правило Хэбба. Теорема Новикова о сходимости. Доказательство теоремы Новикова. Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, "выбивание" из локальных минимумов. 3. Метод стохастического среднего градиента SAG. Проблема мультиколлинеарности и переобучения, редукция весов (weight decay). Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели. Квадратичный (гауссовский) и лапласовский регуляризаторы. 4. Настройка порога решающего правила по критерию числа ошибок I и II рода. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации AUC.				

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
Раздел 6. Метод опорных векторов Представление и разбор теоретической темы, решение задач по темам: 1. Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin). Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно - линейная функция потерь. 2. Понятие опорных векторов. Рекомендации по выбору константы C. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Способы конструктивного построения ядер. Примеры ядер. Обучение SVM методом активных ограничений. SVM - регрессия. Метод релевантных векторов RVM. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM	6	2	2	2
Раздел 7. Многомерная линейная регрессия Представление и разбор теоретической темы, решение задач по темам: Задача регрессии, многомерная линейная регрессия. 1. Метод наименьших квадратов, его вероятностный смысл и геометрический смысл. 2. Сингулярное разложение.	6	2	2	2

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
Проблемы мультиколлинеарности и переобучения. Регуляризация. 3. Гребневая регрессия. Лассо Тибширани, сравнение с гребневой регрессией. 4. Метод главных компонент и декоррелирующее преобразование Карунена - Лоэва, его связь с сингулярным разложением.				
Раздел 8. Композиционные методы классификации и регрессии Представление и разбор теоретической темы, решение задач по темам: 1. Линейные композиции, бустинг Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция. Взвешенное голосование. Алгоритм AdaBoost. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга. Базовые алгоритмы в бустинге. Решающие пни. Градиентный бустинг. 2. Стохастические методы. Стохастические методы: бэггинг и метод случайных подпространств. Случайные леса.	6	2	2	2

Раздел, тема программы учебной дисциплины	Трудоемкость (час)			
	Всего	в том числе по видам учебных занятий		
		лекции	семинары, практические занятия	лабораторные занятия
1	**2**	**3**	**4**	**5**
Раздел 9. Байесовские методы классификации Представление и разбор теоретической темы, решение задач по темам: 1. Оптимальный байесовский классификатор Принцип максимума апостериорной вероятности. Функционал среднего риска. 2. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода. 3. Наивный байесовский классификатор.	4		2	2
Итого часов	**48**	**16**	**16**	**16**
Аудиторных часов	**48**	Формы рубежного (итогового) контроля знаний очной /заочной формы обучения – защита учебного проекта, экзамен		
Внеаудиторная самостоятельная работа	**96**			
Количество часов на выполнение учебного проекта	**48**			
Количество часов на подготовку к зачету/экзамену	**36**			
Всего часов на освоение учебного материала	**180**			

4.2. Содержание дисциплины, структурированное по разделам (темам)

№ п/п	Наименование раздела	Содержание раздела
1.	Раздел 1. Введение	Представление и разбор теоретической темы, решение задач по темам: 1. Постановка задач обучения по прецедентам. Объекты и признаки. 2. Типы шкал: бинарные, номинальные, порядковые, количественные. 3. Типы задач: классификация, регрессия, прогнозирование, кластеризация. Примеры прикладных задач. Основные понятия.
2.	Раздел 2. Обучение без учителя	Представление и разбор теоретической темы, решение задач по темам: 1. Кластеризация Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связных компонент. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ. Функционалы качества кластеризации. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k-means). 2. Сети Кохонена. Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена. Сети встречного распространения, их применение для кусочнопостоянной и гладкой аппроксимации функций. 3. Таксономия. Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи. Алгоритм построения дендрограммы. Определение числа кластеров. Свойства сжатия/растяжения, монотонности и редуктивности.

№ п/п	Наименование раздела	Содержание раздела
3.	Раздел 3. Метрические методы классификации	Представление и разбор теоретической темы, решение задач по темам: 1. Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. 2. Метод окна Парзена 3. Метрические методы классификации в задаче восстановления регрессии. 4. Обнаружение выбросов.
4.	Раздел 4. Логические методы классификации и решающие деревья	Представление и разбор теоретической темы, решение задач по темам: 1. Понятия закономерности и информативности Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности. Асимптотическая эквивалентность статистического и энтропийного определения. Сравнение областей эвристических и статистических закономерностей. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости. 2. Градиентный алгоритм синтеза конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Бинаризация признаков. 3. Решающие деревья для задач классификации и регрессии.

№ п/п	Наименование раздела	Содержание раздела
5.	Раздел 5. Градиентные линейные методы классификации	Представление и разбор теоретической темы, решение задач по темам: 1. Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия. 2. Метод стохастического градиента и частные случаи: адаптивный линейный элемент ADALINE, персептрон Розенблатта, правило Хэбба. Теорема Новикова о сходимости. Доказательство теоремы Новикова. Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, "выбивание" из локальных минимумов. 3. Метод стохастического среднего градиента SAG. Проблема мультиколлинеарности и переобучения, редукция весов (weight decay). Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели. Квадратичный (гауссовский) и лапласовский регуляризаторы. 4. Настройка порога решающего правила по критерию числа ошибок I и II рода. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации AUC.

№ п/п	Наименование раздела	Содержание раздела
6.	Раздел 6. Метод опорных векторов	Представление и разбор теоретической темы, решение задач по темам: 1. Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin). Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно - линейная функция потерь. Задача квадратичного программирования и двойственная задача. 2. Понятие опорных векторов. Рекомендации по выбору константы C. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Способы конструктивного построения ядер. Примеры ядер. Обучение SVM методом активных ограничений. SVM - регрессия. Метод релевантных векторов RVM. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.
7.	Раздел 7. Многомерная линейная регрессия	Представление и разбор теоретической темы, решение задач по темам: Задача регрессии, многомерная линейная регрессия. 1. Метод наименьших квадратов, его вероятностный смысл и геометрический смысл. 2. Сингулярное разложение. Проблемы мультиколлинеарности и переобучения. Регуляризация. 3. Гребневая регрессия. Лассо Тибширани, сравнение с гребневой регрессией. 4. Метод главных компонент и декоррелирующее преобразование Карунена - Лоэва, его связь с сингулярным разложением.

№ п/п	Наименование раздела	Содержание раздела
8.	Раздел 8. Композиционные методы классификации и регрессии	Представление и разбор теоретической темы, решение задач по темам: 1. Линейные композиции, бустинг Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция. Взвешенное голосование. Алгоритм AdaBoost. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга. Базовые алгоритмы в бустинге. Решающие пни. Градиентный бустинг. 2. Стохастические методы. Стохастические методы: бэггинг и метод случайных подпространств. Случайные леса.
9.	Раздел 9. Байесовские методы классификации	Представление и разбор теоретической темы, решение задач по темам: 1. Оптимальный байесовский классификатор Принцип максимума апостериорной вероятности. Функционал среднего риска. 2. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода. 3. Наивный байесовский классификатор.

Практические занятия (семинары)

№ п/п	Наименование раздела	Содержание раздела
1.	Раздел 2. Обучение без учителя	Практическое занятие на тему: «Кластеризация Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Сети Кохонена. Алгоритмы кластеризации»
2.	Раздел 3. Метрические методы классификации	Практическое занятие на тему: «Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. Метод окна Парзена. Метрические методы классификации в задаче восстановления регрессии. Обнаружение выбросов»
3.	Раздел 4. Логические методы классификации и решающие деревья	Практическое занятие на тему: «Решающие деревья для задач классификации и регрессии»
4.	Раздел 5. Градиентные линейные методы классификации	Практическое занятие на тему: «Градиентные и линейные методы классификации»
5.	Раздел 6. Метод опорных векторов	Практическое занятие на тему: «Ядерные алгоритмы машинного обучения – SVM, SVC, SVR»
6.	Раздел 7. Многомерная линейная регрессия	Практическое занятие на тему: «Метод наименьших квадратов, его вероятностный смысл и геометрический смысл. Сингулярное разложение. Проблемы мультиколлинеарности и переобучения. Регуляризация. Гребневая регрессия. Метод главных компонент.»

№ п/п	Наименование раздела	Содержание раздела
7.	Раздел 8. Композиционные методы классификации и регрессии	Практическое занятие на тему: «Композиционные методы классификации и регрессии»
8.	Раздел 9. Байесовские методы классификации	Практическое занятие на тему: «Оптимальный байесовский классификатор Принцип максимума апостериорной вероятности. Функционал среднего риска. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода»

Лабораторные занятия (как этапы выполнения учебных проектов)

№ п/п	№ раздела дисциплины	Наименование лабораторных работ	Трудо-емкость (час.)
1.	Раздел 2. Обучение без учителя	Применение алгоритмов кластеризации для диагностики кризиса теплообмена в ЯЭУ.	2
2.	Раздел 3. Метрические методы классификации	Использование Метода ближайших соседей (kNN) для решения задачи классификации. Настройка гиперпараметров модели машинного обучения	2
3.	Раздел 4. Логические методы классификации и решающие деревья	Решающие деревья для задач классификации и регрессии	2
4.	Раздел 5. Градиентные линейные методы классификации	Применение линейных методов классификации для диагностики кризиса теплообмена в ЯЭУ.	2
5.	Раздел 6. Метод опорных векторов	Мультиклассификация с помощью SVM моделей	2

№ п/п	№ раздела дисциплины	Наименование лабораторных работ	Трудо-емкость (час.)
6.	Раздел 7. Многомерная линейная регрессия	Понижение размерности данных с помощью метода главных компонент и сингулярного разложения. Предсказание высоты дефекта в сварных швах трубопроводов АЭС с помощью линейных регрессионных моделей.	2
7.	Раздел 8. Композиционные методы классификации и регрессии	Техники градиентного бустинга, бэггинга и стэкинга для решения задач классификации и регрессии. Решение задачи классификации типа дефекта в сварных швах трубопроводов АЭС.	2
8.	Раздел 9. Байесовские методы классификации	Наивный байесовский классификатор для определения стороны дефекта в сварных швах трубопроводов АЭС.	2

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

5.1. Основная литература

1. Рашка, С. Python и машинное обучение [Электронный ресурс] : руководство / С. Рашка ; пер. с англ. Логунова А.В.. — Электрон. дан. — Москва : ДМК Пресс, 2017. — 418 с. — Режим доступа: https://e.lanbook.com/book/100905. — Загл. с экрана.

2. Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — 2nd ed. — Springer-Verlag, - 2009. https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Дополнительная литература

1. . Северенс, Ч. Введение в программирование на Python / Ч. Северенс. - 2-е изд., испр. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 231 с. : схем., ил. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=429184

5.3. Программное обеспечение

1. Образовательный портал «Кафедра онлайн» http://x.obninsk.ru
2. Инструмент Python–разработчика JetBrains PyCharm
3. Инструмент для анализа данных Jupyter Notebook
4. Дистрибутив языков программирования Python и R – Anaconda
5. Python IDE – интерпретатор Python.

6. 5.4. Электронные ресурсы:

1. Образовательный портал «Кафедра онлайн»: http://x.obninsk.ru, http://vt.obninsk.ru.

2. Официальный сайт Python: https://www.python.org/

3. Библиотеки Python: http://www.numpy.org/, http://matplotlib.org/, http://scikit-learn.org/stable/, http://pandas.pydata.org/

Материалы учебно-методического комплекса дисциплины в полном объеме доступны на образовательном портале «Кафедра онлайн».

6. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

6.1. Паспорт фонда оценочных средств по дисциплине

№ п/п	Контролируемые модули, разделы (темы) дисциплины	Код контролируемой компетенции (или ее части)	Наименование оценочного средства
1		СПК-1	Самостоятельная работа, экзамен. Письменно.
2	Раздел 2. Обучение без учителя	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.
3	Раздел 3. Метрические методы классификации	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.
4	Раздел 4. Логические методы классификации и решающие деревья	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.
5	Раздел 5. Градиентные линейные методы классификации	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.
6	Раздел 6. Метод опорных векторов	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.

7	Раздел 7. Многомерная линейная регрессия	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.
8	Раздел 8. Композиционные методы классификации и регрессии	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.
9	Раздел 9. Байесовские методы классификации	СПК-1	Отчет по лабораторной работе, экзамен. Письменно.

6.2. Типовые контрольные задания или иные материалы

6.2.1. Вопросы для подготовки к экзамену

1. Постановка задач обучения по прецедентам.
2. Типы шкал: бинарные, номинальные, порядковые, количественные.
3. Типы задач: классификация, регрессия, прогнозирование, кластеризация. Примеры прикладных задач.
4. Постановка задачи кластеризации. Примеры прикладных задач.
5. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связных компонент.
6. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ.
7. Функционалы качества кластеризации
8. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k-means).
9. Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM
10. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных.
11. Искусство интерпретации карт Кохонена. Сети встречного распространения, их применение для кусочнопостоянной и гладкой аппроксимации функций.
12. Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи.
13. Алгоритм построения дендрограммы. Определение числа кластеров. Свойства сжатия/растяжения, монотонности и редуктивности
14. Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля.
15. Метод окна Парзена.
16. Метрические методы классификации в задаче восстановления регрессии. Обнаружение выбросов.
17. Понятия закономерности и информативности. Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности.
18. Асимптотическая эквивалентность статистического и энтропийного определения. Сравнение областей эвристических и статистических закономерностей.
19. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости.

20. Градиентный алгоритм синтеза конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Бинаризация признаков.

21. Решающие деревья для задач классификации и регрессии.

22. Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия.

23. Метод стохастического градиента и частные случаи: адаптивный линейный элемент ADALINE, персептрон Розенблатта, правило Хэбба.

24. Теорема Новикова о сходимости. Доказательство теоремы Новикова.

25. Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, "выбивание" из локальных минимумов.

26. Метод стохастического среднего градиента SAG.

27. Проблема мультиколлинеарности и переобучения, редукция весов (weight decay).

28. Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели. Квадратичный (гауссовский) и лапласовский регуляризаторы.

29. Настройка порога решающего правила по критерию числа ошибок I и II рода.

30. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации AUC.

31. Понятие опорных векторов. Рекомендации по выбору константы C. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера.

32. Способы конструктивного построения ядер. Примеры ядер. Обучение SVM методом активных ограничений.

33. SVM - регрессия. Метод релевантных векторов RVM. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

34. Метод наименьших квадратов, его вероятностный смысл и геометрический смысл.

35. Сингулярное разложение. Проблемы мультиколлинеарности и переобучения. Регуляризация.

36. Гребневая регрессия. Лассо Тибширани, сравнение с гребневой регрессией.

37. Метод главных компонент и декоррелирующее преобразование Карунена - Лоэва, его связь с сингулярным разложением.

38. Линейные композиции, бустинг Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция. Взвешенное голосование.

39. Алгоритм AdaBoost. Процесс последовательного обучения базовых алгоритмов.

40. Теорема о сходимости бустинга. Базовые алгоритмы в бустинге. Решающие пни. Градиентный бустинг.
41. Стохастические методы: бэггинг и метод случайных подпространств. Случайные леса.
42. Оптимальный байесовский классификатор. Принцип максимума апостериорной вероятности. Функционал среднего риска.
43. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора.
44. Оценивание плотности распределения: три основных подхода.
45. Наивный байесовский классификатор.

6.2.2. Перечень предметных областей для учебных проектов (циклов лабораторных работ)

Контрольные точки учебного проекта:

Точка 1: Лабораторная работа №1. Применение алгоритмов кластеризации для диагностики кризиса теплообмена в ЯЭУ.

Точка 2: Лабораторная работа №2. Использование Метода ближайших соседей (kNN) для решения задачи классификации. Настройка гиперпараметров модели машинного обучения.

Точка 3: Лабораторная работа №3. Решающие деревья для задач классификации и регрессии.

Точка 4: Лабораторная работа №4. Применение линейных методов классификации для диагностики кризиса теплообмена в ЯЭУ.

Точка 5: Лабораторная работа №5. Мультиклассификация с помощью SVM моделей.

Точка 6: Лабораторная работа №6. Понижение размерности данных с помощью метода главных компонент и сингулярного разложения. Предсказание высоты дефекта в сварных швах трубопроводов АЭС с помощью линейных регрессионных моделей

Точка 7: Лабораторная работа №7 Техники градиентного бустинга, бэггинга и стэкинга для решения задач классификации и регрессии. Решение задачи классификации типа дефекта в сварных швах трубопроводов АЭС.

Точка 8: Лабораторная работа №8. Наивный байесовский классификатор для определения стороны дефекта в сварных швах трубопроводов АЭС.

6.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Н.Ф. Ефремова, В.Г. Казанович. Оценка качества подготовки обучающихся в рамках ФГОС ВПО.

7. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

1. Рашка, С. Python и машинное обучение [Электронный ресурс] : руководство / С. Рашка ; пер. с англ. Логунова А.В.. — Электрон. дан. —

Москва: ДМК Пресс, 2017. — 418 с. — Режим доступа: https://e.lanbook.com/book/100905. — Загл. с экрана.

2. Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — 2nd ed. — Springer-Verlag, - 2009. https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Дополнительная литература

1. . Северенс, Ч. Введение в программирование на Python / Ч. Северенс. - 2-е изд., испр. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 231 с. : схем., ил. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=429184

8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» (далее - сеть «Интернет»), необходимых для освоения дисциплины

1. Образовательный портал «Кафедра онлайн»: http://x.obninsk.ru, http://vt.obninsk.ru.

2. Официальный сайт Python: https://www.python.org/

3. Библиотеки Python: http://www.numpy.org/, http://matplotlib.org/, http://scikit-learn.org/stable/, http://pandas.pydata.org/

Материалы учебно-методического комплекса дисциплины в полном объеме доступны на образовательном портале «Кафедра онлайн».

9. Методические указания для обучающихся по освоению дисциплины

В ходе лабораторного практикума каждый студент выполняет лабораторные работы. В течение семестра последовательно создаются артефакты и компоненты программного обеспечения. Защита лабораторных происходит во время контрольных точек каждые две недели семестра.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем (при необходимости)

Использование видео- и аудиоматериалов (учебный видео канал), дистанционного консультирования и вебинаров с применением интернет-портала «Кафедра онлайн», расположенного по сетевому адресу http://x.obninsk.ru, см. компонент «Учебный форум студентов».

1. Полнотекстовые журналы Springer Journals за 1997-2015 г., электронные книги (20052016 гг.), коллекция научных биомедицинских и биологических протоколов SpringerProtocols, коллекция научных материалов

в области физических наук и инжиниринга SpringerMaterials, реферативная БД по чистой и прикладной математике zbMATH.

2. Электронная библиотека диссертаций Российской государственной библиотеки (ЭБД РГБ)
3. Электронные ресурсы Web of Science Core Collection (Thomson Reuters Scientific LLC.), Journal Citation Reports + ESI
4. БД Scopus (Elsevier)

№	Наименование	Назначение
1	Презентационное оборудование (мультимедиа-проектор, экран, компьютер для управления)	Для проведения семинарских занятий
2	Компьютерный класс (с выходом в Internet)	Для организации самостоятельной работы обучающихся

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Специализированные компьютерные классы, ауд. 2-510, 2-521, 2-604 аудиторного фонда ИАТЭ НИЯУ МИФИ.

20 компьютеризованных рабочих мест в ауд. 2-521 ИАТЭ НИЯУ МИФИ и 24 компьютеризованных рабочих мест в ауд. 2-604 ИАТЭ НИЯУ МИФИ.

12. Иные сведения и (или) материалы

12.1. Перечень образовательных технологий, используемых при осуществлении образовательного процесса по дисциплине

В ходе преподавания дисциплины применяются следующие методы интерактивного обучения:

1. Методы проектного управления для малых проектных групп как форма выполнения студентами лабораторных работ (проектных заданий).
2. Круглый стол, дискуссия, дебаты как форма консультирования студентов.
3. Деловые и ролевые игры как форма организации работы в проектных группах и форма защиты выполненных проектных заданий.
4. Мозговой штурм, case-study (коллективный анализ конкретных ситуаций, ситуационный анализ) при поиске вариантов решения задач, сформулированных в проектных заданиях.
5. Мастер классы, тренинги и симуляции, которые организуют студенты-магистранты.

12.2. Формы организации самостоятельной работы обучающихся (темы, выносимые для самостоятельного изучения; вопросы для самоконтроля; типовые задания для самопроверки

Список вопросов для самостоятельной работы

1. Постановка задач обучения по прецедентам.
2. Типы шкал: бинарные, номинальные, порядковые, количественные.
3. Типы задач: классификация, регрессия, прогнозирование, кластеризация. Примеры прикладных задач.
4. Постановка задачи кластеризации. Примеры прикладных задач.
5. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связных компонент.
6. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ.
7. Функционалы качества кластеризации
8. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k-means).
9. Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM
10. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных.
11. Искусство интерпретации карт Кохонена. Сети встречного распространения, их применение для кусочнопостоянной и гладкой аппроксимации функций
12. Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи.
13. Алгоритм построения дендрограммы. Определение числа кластеров. Свойства сжатия/растяжения, монотонности и редуктивности
14. Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля.
15. Метод окна Парзена.
16. Метрические методы классификации в задаче восстановления регрессии. Обнаружение выбросов.
17. Понятия закономерности и информативности. Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности.
18. Асимптотическая эквивалентность статистического и энтропийного определения. Сравнение областей эвристических и статистических закономерностей.

19. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости.

20. Градиентный алгоритм синтеза конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Бинаризация признаков.

21. Решающие деревья для задач классификации и регрессии.

22. Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия.

23. Метод стохастического градиента и частные случаи: адаптивный линейный элемент ADALINE, персептрон Розенблатта, правило Хэбба.

24. Теорема Новикова о сходимости. Доказательство теоремы Новикова.

25. Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, "выбивание" из локальных минимумов.

26. Метод стохастического среднего градиента SAG.

27. Проблема мультиколлинеарности и переобучения, редукция весов (weight decay).

28. Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели. Квадратичный (гауссовский) и лапласовский регуляризаторы.

29. Настройка порога решающего правила по критерию числа ошибок I и II рода.

30. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации AUC.

31. . Понятие опорных векторов. Рекомендации по выбору константы C. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера.

32. Способы конструктивного построения ядер. Примеры ядер. Обучение SVM методом активных ограничений.

33. SVM - регрессия. Метод релевантных векторов RVM. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

34. Метод наименьших квадратов, его вероятностный смысл и геометрический смысл.

35. Сингулярное разложение. Проблемы мультиколлинеарности и переобучения. Регуляризация.

36. . Гребневая регрессия. Лассо Тибширани, сравнение с гребневой регрессией.

37. Метод главных компонент и декоррелирующее преобразование Карунена - Лоэва, его связь с сингулярным разложением.

38. Линейные композиции, бустинг Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция. Взвешенное голосование.

39. Алгоритм AdaBoost. Процесс последовательного обучения базовых алгоритмов.
40. Теорема о сходимости бустинга. Базовые алгоритмы в бустинге. Решающие пни. Градиентный бустинг.
41. Стохастические методы: бэггинг и метод случайных подпространств. Случайные леса.
42. Оптимальный байесовский классификатор. Принцип максимума апостериорной вероятности. Функционал среднего риска.
43. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора.
44. Оценивание плотности распределения: три основных подхода.
45. Наивный байесовский классификатор.

12.3. Краткий терминологический словарь

Машинное обучение (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Различают два типа обучения. Обучение по прецедентам, или индуктивное обучение, основано на выявлении эмпирических закономерностей в данных. Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний. Дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.

Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации (англ. information extraction), интеллектуальным анализом данных (data mining).

Обучение с учителем — для каждого прецедента задаётся пара «ситуация, требуемое решение».

Обучение без учителя — для каждого прецедента задаётся только «ситуация», требуется сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов, и/или понизить размерность данных.

Обучение с подкреплением — для каждого прецедента имеется пара «ситуация, принятое решение».

Бустинг (англ. boosting — улучшение) — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов.

Big Data – термин подразумевает две трактовки. С одной стороны, это собственно большие объемы данных, с другой – совокупность технологий, которые имеют дело с незаурядными по интенсивности и объему потоками данных. В частности, это технологии работы с быстро поступающей

информацией, когда требуется обрабатывать параллельно и в реальном времени большие массивы данных, в том числе слабо структурированных. Когда говорят о Big Data, подразумевают три аспекта (три V):

- Volume – большой объем данных;
- Variety – разнообразие данных;
- Velocity – необходимость обрабатывать данные очень быстро.