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 профессионального образования «Национальный исследовательский ядерный университет «МИФИ»
(ИАТЭ НИЯУ МИФИ)
Отделение интеллектуальных кибернетических систем

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

Методические указания по дисциплине

«Машинное обучение»

для студентов направления подготовки

09.04.01 Информатика и вычислительная техника

программа:
Большие данные и машинное обучение в задачах атомной энергетики

Форма обучения: очная

г. Обнинск 2022г.

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.

1. 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

2. 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.

3. Иные сведения и (или) материалы

3.1. 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.

3.2. 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.

3.3. 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.