МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ГЕОДЕЗИИ И КАРТОГРАФИИ
(МИИГАиК)

Факультет геоинформатики и информационной безопасности
Кафедра информатики и геоинформационных технологий

**ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ**

Сертификат: 5BFEE1D286E9C95A0C0D2C611E434BD5
Владелец: Степанченко Алексей Леонидович
Действителен: с 23.09.2026 по 17.12.2027

«УТВЕРЖДАЮ»
Проректор
А. Л. Степанченко
18 апреля 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ) «Б1.О.11 ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

Уровень высшего образования: магистратура

Направление подготовки: 09.04.03 Прикладная информатика

Направленность (профиль) подготовки: Анализ больших пространственных данных

Квалификация (степень) выпускника: магистр

Форма обучения: очная

Год набора (приема на обучение): 2025

Срок получения образования: 2 года

Объем: в зачетных единицах: 4 з.е.
в академических часах: 144 ак.ч.

Москва, 2025

Разработчики:
Старший преподаватель кафедры Информатики и
геоинформационных технологий Стрельцов А. С.

Рабочая программа дисциплины (модуля) составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.04.03 Прикладная информатика, утвержденного приказом Минобрнауки от 19.09.2017 № 916

Согласование и утверждение

№	Подразделение	Ответственное лицо	ФИО	Виза	Дата, протокол (при наличии)
1	факультет Геоинформатики и информационной безопасности	Председатель ученого совета факультета	Матерухин А. В.	Рассмотрено	14.10.2024, № 10-24

Содержание

1. Цель и задачи освоения дисциплины (модуля)

2. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с планируемыми результатами освоения образовательной программы

3. Место дисциплины в структуре ОП

4. Объем дисциплины и виды учебной работы

5. Содержание дисциплины
   5.1. Разделы, темы дисциплины и виды занятий
   5.2. Содержание разделов, тем дисциплины и формы текущего контроля

6. Оценочные материалы текущего контроля

7. Оценочные материалы промежуточной аттестации

8. Порядок проведения промежуточной аттестации

9. Материально-техническое и учебно-методическое обеспечение дисциплины
   9.1. Перечень основной и дополнительной учебной литературы
   9.2. Профессиональные базы данных и ресурсы «Интернет», к которым обеспечивается доступ обучающихся
   9.3. Программное обеспечение и информационно-справочные системы, используемые при осуществлении образовательного процесса по дисциплине
   9.4. Оборудование и технические средства обучения

10. Методические указания по освоению дисциплины (модуля)

1. Цель и задачи освоения дисциплины (модуля)

Цель освоения дисциплины - формирование у обучающихся системных знаний и практических навыков в области интеллектуального анализа данных (Data Mining) и применения современных методов извлечения знаний из больших объемов структурированных и неструктурированных данных для решения аналитических и прогностических задач в профессиональной деятельности.

Задачи изучения дисциплины:

- Освоение теоретических основ и практических методов предобработки, визуализации и исследовательского анализа данных с использованием современных библиотек Python и инструментов Data Mining.;
- Изучение и применение алгоритмов машинного обучения для решения основных задач интеллектуального анализа данных: классификации, кластеризации, регрессии, ассоциативного анализа и выявления аномалий.;
- Формирование навыков разработки комплексных аналитических решений, построения прогностических моделей и интерпретации результатов анализа для принятия обоснованных решений в профессиональной сфере..

2. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с планируемыми результатами освоения образовательной программы

Компетенции, индикаторы и результаты обучения

ОПК-2 Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач

ОПК-2.3 Разрабатывает программные средства на основе методов интеллектуального анализа данных для решения профессиональных задач

Знать:

ОПК-2.3/Зн1 Теоретические основы и математический аппарат методов интеллектуального анализа данных (Data Mining), включая алгоритмы классификации, кластеризации, регрессионного анализа, ассоциативных правил, а также принципы их программной реализации с использованием библиотек Python (NumPy, Pandas, Scikit-learn).

ОПК-2.3/Зн2 Методы предобработки, визуализации и исследовательского анализа данных (EDA), метрики оценки качества моделей машинного обучения, техники валидации и оптимизации гиперпараметров для создания эффективных программных решений.

ОПК-2.3/Зн3 Архитектуру и принципы построения программных систем интеллектуального анализа данных, включая конвейеры обработки данных (ML pipelines), методы развертывания моделей и создания пользовательских интерфейсов.

Уметь:

ОПК-2.3/Ум1 Разрабатывать программные модули для решения задач интеллектуального анализа данных с использованием алгоритмов машинного обучения (KNN, деревья решений, случайный лес, SVM, наивный Байес, ансамблевые методы).

ОПК-2.3/Ум2 Проектировать и реализовывать комплексные программные решения для анализа структурированных и неструктурированных данных, включая обработку текстовой информации, временных рядов и пространственных данных.

ОПК-2.3/Ум3 Интегрировать разработанные программные средства интеллектуального анализа данных в существующие информационные системы, создавать API и веб-интерфейсы для взаимодействия с аналитическими моделями.

Владеть:

ОПК-2.3/Нв1 Навыками программирования на языке Python и использования специализированных библиотек (NumPy, Pandas, Matplotlib, Plotly, Scikit-learn) для разработки программных средств интеллектуального анализа данных.

ОПК-2.3/Нв2 Методами построения, обучения, валидации и оптимизации моделей машинного обучения, а также навыками оценки их эффективности и интерпретации результатов для решения профессиональных задач.

ОПК-2.3/Нв3 Технологиями развертывания программных решений интеллектуального анализа данных, включая сериализацию моделей, создание REST API с FastAPI, разработку интерактивных веб-приложений с Streamlit.

3. Место дисциплины в структуре ОП

Дисциплина (модуль) Б1.О.11 «Интеллектуальный анализ данных» относится к обязательной части образовательной программы и изучается в семестре(ах): 3.

В процессе изучения дисциплины студент готовится к решению типов задач профессиональной деятельности, предусмотренных ФГОС ВО и образовательной программой.

4. Объем дисциплины (модуля) и виды учебной работы

Период обучения	Общая трудоемкость (часы)	Общая трудоемкость (ЗЕТ)	Лекционные занятия (часы)	Практические занятия (часы)	Самостоятельная работа (часы)	Промежуточная аттестация (часы)
3 семестр	144	4	34	34	49	Экзамен (27)
Всего	144	4	34	34	49	27

5. Содержание дисциплины (модуля)

5.1. Разделы, темы дисциплины и виды занятий

(часы промежуточной аттестации не указываются)

Наименование раздела, темы	Всего	Лекционные занятия	Практические занятия	Самостоятельная работа	Планируемые результаты обучения, соотнесенные с результатами освоения программы
Раздел 1. Математические основы анализа данных	**6**	**2**	**2**	**2**	ОПК-2.3

Тема 1.1. Библиотека NumPy для численных вычислений	3	1	1	1	
Тема 1.2. Статистические вычисления в NumPy	3	1	1	1	
Раздел 2. Визуализация и анализ данных	**15**	**4**	**4**	**7**	ОПК-2.3
Тема 2.1. Визуализация данных с Matplotlib	4	1	1	2	
Тема 2.2. Интерактивная визуализация с Plotly	3	1	1	1	
Тема 2.3. Работа с табличными данными в Pandas	4	1	1	2	
Тема 2.4. Предобработка данных с Pandas	4	1	1	2	
Раздел 3. Основы теории машинного обучения	**12**	**3**	**3**	**6**	ОПК-2.3
Тема 3.1. Введение в машинное обучение	4	1	1	2	
Тема 3.2. Метрики оценки качества моделей	4	1	1	2	
Тема 3.3. Валидация моделей и кросс-валидация	4	1	1	2	
Раздел 4. Методы обучения на основе расстояний	**8**	**2**	**2**	**4**	ОПК-2.3
Тема 4.1. Метод k-ближайших соседей (KNN)	4	1	1	2	
Тема 4.2. Оптимизация гиперпараметров	4	1	1	2	
Раздел 5. Линейные модели	**12**	**2**	**2**	**8**	ОПК-2.3
Тема 5.1. Линейная регрессия	6	1	1	4	
Тема 5.2. Логистическая регрессия	6	1	1	4	
Раздел 6. Древовидные модели	**16**	**5**	**5**	**6**	ОПК-2.3
Тема 6.1. Решающие деревья для классификации	6	2	2	2	
Тема 6.2. Решающие деревья для регрессии	6	2	2	2	
Тема 6.3. Критерии останова и регуляризация деревьев	4	1	1	2	
Раздел 7. Ансамблевые методы	**18**	**6**	**6**	**6**	ОПК-2.3
Тема 7.1. Случайный лес (Random Forest)	6	2	2	2	
Тема 7.2. Градиентный бустинг	6	2	2	2	
Тема 7.3. Стекинг и блендинг моделей	6	2	2	2	
Раздел 8. Вероятностные и дистанционные методы	**12**	**4**	**4**	**4**	ОПК-2.3
Тема 8.1. Наивный байесовский классификатор	6	2	2	2	
Тема 8.2. Метод опорных векторов (SVM)	6	2	2	2	

Раздел 9. Практическое применение машинного обучения	**18**	**6**	**6**	**6**	ОПК-2.3
Тема 9.1. Классификация текстовых данных	6	2	2	2	
Тема 9.2. Построение комплексных ML-решений	6	2	2	2	
Тема 9.3. Развертывание ML-моделей	6	2	2	2	
Итого	**117**	**34**	**34**	**49**	

5.2. Содержание разделов, тем дисциплин и формы текущего контроля

Раздел 1. Математические основы анализа данных

Тема 1.1. Библиотека NumPy для численных вычислений

Основы работы с многомерными массивами ndarray. Создание массивов различными способами. Индексация и срезы. Математические операции над массивами: поэлементные операции, матричные операции, векторизация вычислений. Функции агрегации. Broadcasting - правила транслирования. Линейная алгебра в NumPy.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 1.2. Статистические вычисления в NumPy

Генерация случайных чисел и распределений . Описательная статистика: среднее, медиана, дисперсия, стандартное отклонение. Корреляционный анализ. Работа с пропущенными значениями. Применение NumPy для предобработки данных в машинном обучении.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Раздел 2. Визуализация и анализ данных

Тема 2.1. Визуализация данных с Matplotlib

Основы построения графиков с библиотекой Matplotlib . Типы визуализаций: линейные графики, гистограммы, диаграммы рассеяния, box-plot. Настройка осей, легенд, заголовков. Создание подграфиков. Визуализация распределений и зависимостей в данных для машинного обучения.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 2.2. Интерактивная визуализация с Plotly

Создание интерактивных графиков с библиотекой Plotly . 3D-визуализации. Анимированные графики. Dashboard-визуализации. Визуализация результатов кластеризации и классификации. Построение интерактивных отчетов по результатам анализа.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 2.3. Работа с табличными данными в Pandas

Структуры данных Series и DataFrame . Загрузка данных из различных источников (CSV, Excel, SQL). Индексация и выборка данных. Группировка и агрегация. Объединение датафреймов. Работа с временными рядами. Pivot-таблицы.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 2.4. Предобработка данных с Pandas

Обработка пропущенных значений . Удаление дубликатов. Преобразование типов данных. Кодирование категориальных признаков. Нормализация и стандартизация данных. Feature engineering - создание новых признаков. Подготовка данных для машинного обучения.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Раздел 3. Основы теории машинного обучения

Тема 3.1. Введение в машинное обучение

Основные понятия и терминология машинного обучения. Типы задач: классификация, регрессия, кластеризация. Обучение с учителем и без учителя. Этапы разработки ML-модели. Проблемы переобучения и недообучения. Bias-variance tradeoff.

Тема 3.2. Метрики оценки качества моделей

Метрики для задач регрессии: MSE, RMSE, MAE, R^2 . Метрики для задач классификации: accuracy, precision, recall, F1-score. Матрица ошибок. ROC-кривая и AUC. Выбор оптимальной метрики в зависимости от бизнес-задачи. Интерпретация метрик качества.

Тема 3.3. Валидация моделей и кросс-валидация

Разделение данных на обучающую, валидационную и тестовую выборки. Методы кросс-валидации: k-fold, stratified k-fold, leave-one-out. Временная валидация для временных рядов. Оценка стабильности модели. Выбор оптимальной модели по результатам валидации.

Раздел 4. Методы обучения на основе расстояний

Тема 4.1. Метод k-ближайших соседей (KNN)

Алгоритм KNN для задач классификации и регрессии . Выбор метрики расстояния: евклидово, манхэттенское, расстояние Минковского. Влияние параметра k на качество модели. Взвешенный KNN. Масштабирование признаков. Применение KNN для рекомендательных систем.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 4.2. Оптимизация гиперпараметров

Методы подбора оптимальных гиперпараметров . Grid Search и Random Search. Байесовская оптимизация. Автоматизированный подбор гиперпараметров с помощью GridSearchCV и RandomizedSearchCV. Стратегии ускорения поиска. Визуализация результатов оптимизации.

Раздел 5. Линейные модели

Тема 5.1. Линейная регрессия

Метод наименьших квадратов . Множественная линейная регрессия. Полиномиальная регрессия. Регуляризация: Ridge, Lasso, ElasticNet. Градиентный спуск и его модификации. Диагностика модели: анализ остатков, проверка предположений.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 5.2. Логистическая регрессия

Бинарная и многоклассовая логистическая регрессия . Функция логистических потерь. Интерпретация коэффициентов. Регуляризация в логистической регрессии. Калибровка вероятностей. Применение для задач кредитного скоринга и медицинской диагностики.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Раздел 6. Древовидные модели

Тема 6.1. Решающие деревья для классификации

Алгоритм построения дерева решений . Критерии разбиения: Gini, энтропия. Обрезка дерева для борьбы с переобучением. Визуализация и интерпретация деревьев. Преимущества и недостатки деревьев решений.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 6.2. Решающие деревья для регрессии

Особенности деревьев регрессии . Критерии разбиения: MSE, MAE. Экстраполяционные свойства деревьев. Сравнение с линейными моделями. Применение для задач прогнозирования.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 6.3. Критерии останова и регуляризация деревьев

Параметры контроля сложности дерева . Минимальное количество объектов в листе. Максимальная глубина дерева. Минимальное улучшение критерия. Постпруning и препруning. Влияние параметров на качество модели.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Раздел 7. Ансамблевые методы

Тема 7.1. Случайный лес (Random Forest)

Принцип работы Random Forest . Бэггинг и случайные подпространства признаков. Out-of-bag оценка. Важность признаков. Параметры настройки: количество деревьев, глубина, количество признаков. Применение для задач с большим числом признаков.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 7.2. Градиентный бустинг

Основы бустинга . Алгоритм градиентного бустинга. XGBoost, LightGBM, CatBoost - сравнение реализаций. Настройка гиперпараметров бустинга. Early stopping. Обработка категориальных признаков.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 7.3. Стекинг и блендинг моделей

Методы комбинирования моделей . Усреднение предсказаний. Взвешенное голосование. Стекинг с мета-алгоритмом. Блендинг на валидационной выборке. Создание ансамблей для соревнований по машинному обучению.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Раздел 8. Вероятностные и дистанционные методы

Тема 8.1. Наивный байесовский классификатор

Теорема Байеса и ее применение . Гауссовский наивный Байес. Мультиномиальный и бернуллиевский варианты. Сглаживание Лапласа. Применение для классификации текстов и фильтрации спама.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 8.2. Метод опорных векторов (SVM)

Линейный SVM для разделимых и неразделимых выборок . Kernel trick и нелинейные SVM. Типы ядер: RBF, полиномиальное, сигмоидальное. Параметры C и gamma. SVM для регрессии (SVR). Многоклассовая классификация.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Раздел 9. Практическое применение машинного обучения

Тема 9.1. Классификация текстовых данных

Предобработка текста: токенизация, лемматизация, удаление стоп-слов. Векторизация текста: Bag of Words, TF-IDF, Word2Vec. Применение различных алгоритмов для классификации текстов. Анализ тональности. Классификация документов.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

Тема 9.2. Построение комплексных ML-решений

Создание end-to-end pipeline для машинного обучения . Автоматизация предобработки данных. Feature engineering и отбор признаков.

Тема 9.3. Развертывание ML-моделей

Сериализация обученных моделей с pickle и joblib. Создание REST API для моделей с Flask/FastAPI. Развертывание в облаке. Разработка графического интерфейса в Streamlit.

Текущий контроль

Вид (форма) контроля
Просмотр выполненных работ

6. Порядок проведения промежуточной аттестации

Порядок проведения промежуточной аттестации регламентируется локальными нормативными актами Университета:
- Положением о формах, периодичности, порядке текущего контроля успеваемости и промежуточной аттестации обучающихся по основным образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры в федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный университет геодезии и картографии», п.6 «Организация и проведение промежуточной аттестации».
- Положением о порядке осуществления образовательного процесса, проведения текущего контроля успеваемости и промежуточной аттестации обучающихся с применением электронного обучения и дистанционных образовательных технологий в федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный университет геодезии и картографии», п.8 «Проведение промежуточной аттестации с применением ЭО и ДОТ».

Промежуточная аттестация: Экзамен, 3 семестр.
Экзамен по дисциплине проводится в формате тестирования для итоговой оценки знаний.

Студентам будет предложен тест, состоящий из тридцати вопросов закрытого типа с несколькими вариантами ответов. На выполнение всей работы отводится шестьдесят минут, в течение которых запрещается использование любых вспомогательных материалов и электронных устройств.

Итоговая оценка выставляется в зависимости от процента правильных ответов. Для получения оценки «отлично» необходимо верно ответить на 90-100% вопросов (27-30 правильных ответов). Оценка «хорошо» ставится за 70-89% правильных ответов (21-26), а «удовлетворительно» — за 55-69% (17-20). Результат ниже 55% правильных ответов признается неудовлетворительным.

7. Оценочные материалы текущего контроля

Раздел 1. Математические основы анализа данных
Контролируемые ИДК: ОПК-2.3

Тема 1.1. Библиотека NumPy для численных вычислений
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 1.2. Статистические вычисления в NumPy
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 2. Визуализация и анализ данных
Контролируемые ИДК: ОПК-2.3

Тема 2.1. Визуализация данных с Matplotlib
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 2.2. Интерактивная визуализация с Plotly
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 2.3. Работа с табличными данными в Pandas
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 2.4. Предобработка данных с Pandas
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 4. Методы обучения на основе расстояний
Контролируемые ИДК: ОПК-2.3

Тема 4.1. Метод k-ближайших соседей (KNN)
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 5. Линейные модели
Контролируемые ИДК: ОПК-2.3

Тема 5.1. Линейная регрессия
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 5.2. Логистическая регрессия
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 6. Древовидные модели
Контролируемые ИДК: ОПК-2.3

Тема 6.1. Решающие деревья для классификации
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 6.2. Решающие деревья для регрессии
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 6.3. Критерии останова и регуляризация деревьев
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 7. Ансамблевые методы
Контролируемые ИДК: ОПК-2.3

Тема 7.1. Случайный лес (Random Forest)
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 7.2. Градиентный бустинг
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 7.3. Стекинг и блендинг моделей
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 8. Вероятностные и дистанционные методы
Контролируемые ИДК: ОПК-2.3

Тема 8.1. Наивный байесовский классификатор
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 8.2. Метод опорных векторов (SVM)
Форма контроля/оценочное средство: Просмотр выполненных работ

Раздел 9. Практическое применение машинного обучения
Контролируемые ИДК: ОПК-2.3

Тема 9.1. Классификация текстовых данных
Форма контроля/оценочное средство: Просмотр выполненных работ

Тема 9.3. Развертывание ML-моделей
Форма контроля/оценочное средство: Просмотр выполненных работ

8. Оценочные материалы промежуточной аттестации

3 семестр, Экзамен
Контролируемые ИДК: ОПК-2.3
Вопросы/Задания:

1. Модель классификации показала точность (precision) 0.8 и полноту (recall) 0.6. Рассчитайте значение F1-меры. Ответ округлите до сотых.

2. Как называется явление, при котором модель показывает высокое качество на обучающей выборке, но значительно худшее качество на новых, ранее не встречавшихся данных?

3. Рассчитайте манхэттенское расстояние между точками (1, 2) и (4, 6).

4. Дополните предложение термином:
Правило NumPy, определяющее, каким образом операции могут выполняться поэлементно над массивами разной формы путём их согласованного расширения без физического копирования данных, называется _____ (broadcasting).

5. Дополните предложение термином:
Метод регуляризации линейной регрессии, штрафующий сумму абсолютных значений коэффициентов модели и способный полностью обнулять коэффициенты наименее значимых признаков, называется регуляризацией _____ (L1).

6. Объясните суть компромисса между смещением и разбросом (bias-variance tradeoff) при построении модели машинного обучения.

7. Объясните, каким образом один и тот же алгоритм k-ближайших соседей (KNN) применяется как для задачи классификации, так и для задачи регрессии, и в чём разница итогового правила агрегации ответов соседей в этих двух случаях.

8. Сравните регуляризацию Ridge (L2) и Lasso (L1) линейной регрессии, объясните, почему только Lasso может использоваться для отбора признаков.

9. Сравните предварительную обрезку (pre-pruning) и последующую обрезку (post-pruning) дерева решений как способы борьбы с переобучением.

10. Объясните, какое допущение лежит в основе наивного байесовского классификатора и почему оно называется «наивным».

11. Расположите этапы типового процесса разработки модели машинного обучения в правильной последовательности.

1. Обучение модели и подбор гиперпараметров по валидационной выборке.
2. Анализ исходных данных и выявление пропусков и выбросов.
3. Оценка качества итоговой модели на тестовой выборке.
4. Разделение данных на обучающую, валидационную и тестовую выборки.
5. Предобработка данных (обработка пропусков, кодирование категориальных признаков, масштабирование).

12. Расположите этапы подбора гиперпараметров модели методом Grid Search с использованием кросс-валидации в правильной последовательности.

1. Определение сетки (набора значений) для каждого настраиваемого гиперпараметра.
2. Выбор комбинации гиперпараметров, показавшей наилучшее среднее качество по кросс-валидации.
3. Обучение и оценка модели методом кросс-валидации для каждой возможной комбинации значений гиперпараметров из сетки.
4. Обучение окончательной модели на всей обучающей выборке с выбранной комбинацией гиперпараметров.

13. Расположите этапы построения дерева решений для задачи классификации в правильной последовательности.

1. Вычисление значения критерия разбиения (например, неопределённости Джини) для текущего узла.
2. Рекурсивное повторение процесса для каждого из двух дочерних узлов.
3. Разбиение объектов текущего узла на две группы по выбранному признаку и порогу.
4. Перебор возможных признаков и пороговых значений, дающих наибольшее уменьшение критерия разбиения.

14. Расположите этапы развёртывания обученной модели машинного обучения в виде веб-сервиса в правильной последовательности.

1. Сериализация (сохранение) обученной модели в файл при помощи pickle или joblib.
2. Создание REST API, загружающего сохранённую модель и принимающего запросы с новыми данными для предсказания.
3. Обучение и валидация модели на исторических данных.
4. Развёртывание созданного API на сервере (в том числе в облаке) для приёма запросов из сети.

15. Расположите этапы построения ансамбля моделей методом градиентного бустинга в правильной последовательности.

1. Вычисление остатков (ошибок) текущего ансамбля моделей на обучающей выборке.
2. Инициализация ансамбля простой базовой моделью (например, константным предсказанием).
3. Добавление новой обученной модели к ансамблю с некоторым весом (шагом обучения).
4. Обучение новой базовой модели, предсказывающей вычисленные остатки предыдущего шага.

16. Установите соответствие между метрикой качества модели и её определением.

MSE - Доля объектов, верно предсказанных положительными, среди всех объектов, предсказанных моделью как положительные
Precision (точность) - Средний квадрат отклонений предсказанных числовых значений от фактических
Recall (полнота) - Доля объясненной моделью дисперсии целевой переменной относительно среднего значения
R^2 (коэффициент детерминации) - Доля объектов, верно предсказанных положительными, среди всех фактически положительных объектов

17. Установите соответствие между метрикой расстояния и принципом её вычисления.

Евклидово расстояние - Обобщение через параметр степени, включающее евклидово и манхэттенское расстояния как частные случаи
Манхэттенское расстояние - Квадратный корень из суммы квадратов разностей координат точек по каждому измерению

Расстояние Минковского - Сумма модулей разностей координат точек по каждому измерению

18. Установите соответствие между библиотекой Python и её основным назначением в анализе данных.

NumPy - Работа с табличными данными: загрузка, группировка, объединение, обработка пропусков

Pandas - Построение статических графиков с гибкой настройкой осей, легенд и подграфиков

Plotly - Эффективные операции над многомерными числовыми массивами и линейная алгебра

Matplotlib - Построение интерактивных и анимированных графиков и дашбордов

19. Установите соответствие между типом регуляризации линейной регрессии и её особенностью.

Ridge (L2) - Комбинирует штрафы по сумме квадратов и по сумме модулей коэффициентов одновременно

Lasso (L1) - Штрафует сумму квадратов коэффициентов, плавно уменьшая их, но не обнуляя полностью

ElasticNet - Штрафует сумму модулей коэффициентов, способна полностью обнулять коэффициенты малозначимых признаков

20. Установите соответствие между древовидным алгоритмом машинного обучения и его характерным принципом построения.

Одиночное дерево решений - Ансамбль деревьев строится последовательно, каждое новое дерево исправляет ошибки предыдущего ансамбля

Случайный лес - Ансамбль деревьев строится независимо и параллельно, каждое на своей случайной подвыборке данных и признаков

Градиентный бустинг - Единственная модель строится путём рекурсивного разбиения обучающей выборки по значениям признаков

9. Материально-техническое и учебно-методическое обеспечение дисциплины

9.1. Перечень основной и дополнительной учебной литературы

Основная литература

1. Митяков Е. С. Искусственный интеллект и машинное обучение: учебное пособие для вузов / Е. С. Митяков, А. Г. Шмелева, А. И. Ладынин. - Санкт-Петербург: Лань, 2025. - 252 с. - 978-5-507-51465-6. - Текст: электронный // Издательство Лань: [сайт]. - URL: https://e.lanbook.com/book/450827 - Режим доступа: по подписке

2. Титов А. Н. Обработка данных в Python. Основы работы с библиотекой Pandas: учебно-методическое пособие / А. Н. Титов, Р. Ф. Тазиева. - Казань: КНИТУ, 2022. - 116 с. - 978-5-7882-3164-8. - Текст: электронный // Издательство Лань: [сайт]. - URL: https://e.lanbook.com/book/331013 - Режим доступа: по подписке

Дополнительная литература

1. Пальмов С. В. Интеллектуальный анализ данных: учебное пособие / С. В. Пальмов. - Интеллектуальный анализ данных - Самара: Поволжский государственный университет телекоммуникаций и информатики, 2017. - 127 c. - 2227-8397. - Текст: электронный // IPR SMART: [сайт]. - URL: https://www.iprbookshop.ru/75376.html - Режим доступа: по подписке

9.2. Профессиональные базы данных и ресурсы «Интернет», к которым обеспечивается доступ обучающихся

Профессиональные базы данных

Не используются.

Ресурсы «Интернет»

1. https://docs.python.org/3/index.html - Техническая документация по языку программирования Python

2. https://sdo.miigaik.ru/ - Система дистанционного обучения МИИГАиК

9.3. Программное обеспечение и информационно-справочные системы, используемые при осуществлении образовательного процесса по дисциплине

Перечень программного обеспечения
(обновление производится по мере появления новых версий программы)

1. Anaconda;
2. Microsoft Visual Studio;

Перечень информационно-справочных систем
(обновление выполняется еженедельно)

1. 1С Предприятие;

9.4. Оборудование и технические средства обучения

Компьютерный класс

Маркерная доска - 1 шт.
Моноблок - 96 шт.
Проектор - 8 шт.
Автоматизированное рабочее место (АРМ) - 40 шт.
Плоттер - 1 шт.
Сканер - 1 шт.
Интерактивная панель - 1 шт.

10. Методические указания по освоению дисциплины (модуля)

С целью успешного освоения данного предмета студенту необходимо выполнять домашние работы по указанным преподавателям темам и представленным на занятиях материалам. Для более успешного выполнения работы студенту предлагается использование учебно-методических материалов, соответствующих указанной теме.

Самостоятельная работа студентов описывается и регулируется методическими рекомендациями по дисциплине.

Перечень вопросов для промежуточной аттестации:

1. Что является основной структурой данных в библиотеке NumPy?
2. Как в NumPy называется набор правил для выполнения математических операций над массивами разной формы?
3. Какие две основные структуры данных предоставляет библиотека Pandas?
4. Как называется процесс создания новых признаков из существующих данных при работе с Pandas?
5. Какой тип визуализации НЕ упоминается в качестве базового для библиотеки Matplotlib в предоставленном содержании?
6. Что из перечисленного НЕ является основной задачей машинного обучения?
7. Как называется проблема, когда модель отлично работает на обучающих данных, но плохо справляется с новыми, неизвестными данными?
8. Какая из этих метрик обычно используется для оценки качества моделей регрессии?
9. Какая пара метрик является ключевой для оценки моделей классификации и часто используется при построении матрицы ошибок?
10. Какова основная цель k-fold кросс-валидации?
11. Что означает параметр 'k' в алгоритме k-ближайших соседей (KNN)?

12. Какой метод оптимизации гиперпараметров перебирает все возможные комбинации из заданного набора значений?
13. Какой тип регуляризации в линейной регрессии способен полностью обнулять коэффициенты у наименее важных признаков?
14. Какой из перечисленных критериев используется для разделения узлов в дереве решений для задачи классификации?
15. Как называется процесс удаления ветвей из дерева решений для борьбы с переобучением?
16. В чем заключается основной принцип работы ансамблевого метода "Случайный лес" (Random Forest)?
17. Что из перечисленного является популярной реализацией алгоритма градиентного бустинга?
18. Какую роль выполняет "мета-алгоритм" в методе стекинга?
19. Для какой задачи, согласно содержанию, особенно хорошо подходит наивный байесовский классификатор?
20. В чем заключается "трюк с ядром" (kernel trick) в методе опорных векторов (SVM)?
21. Как расшифровывается аббревиатура TF-IDF, используемая при векторизации текстов?
22. Какая библиотека упоминается для сериализации (сохранения) обученных моделей машинного обучения?
23. Какая технология упоминается для создания контейнеров, позволяющих легко развертывать ML-модели?
24. Какая ключевая особенность библиотеки Plotly по сравнению с Matplotlib отмечена в программе?
25. Какой алгоритм оптимизации упоминается как фундаментальный для обучения линейных моделей?
26. В чем заключается основная идея "бустинга"?
27. Как называется процесс приведения слов к их базовой или словарной форме?
28. Компромисс между способностью модели минимизировать ошибку на обучающем наборе и ее способностью обобщаться на новые данные известен как:
29. Что из перечисленного является типом ядра, используемого в методе опорных векторов (SVM)?
30. Какова цель создания REST API для модели машинного обучения?

Описание возможностей изучения дисциплины лицами с ОВЗ и инвалидами

При необходимости обучения студентов-инвалидов и лиц с ограниченными возможностями здоровья аудиторные занятия могут быть заменены или дополнены изучением полнотекстовых лекций, презентаций, видео- и аудиоматериалов в электронной информационно-образовательной среде (ЭИОС) университета. Индивидуальные задания подбираются в адаптированных к ограничениям здоровья формах (письменно или устно, в форме презентаций). Выбор методов обучения зависит от их доступности для инвалидов и лиц с ограниченными возможностями здоровья.

Рабочая программа дисциплины для инвалидов и лиц с ограниченными возможностями здоровья разрабатывается индивидуально с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся.