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
Ни томский государственный университет

**ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ**

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

«УТВЕРЖДАЮ»
Проректор
А. Л. Степанченко
24 марта 2025 г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ (МОДУЛЮ) «Б1.О.08 ВВЕДЕНИЕ В ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

Уровень высшего образования: магистратура

Направление подготовки: 09.04.02 Информационные системы и технологии

Направленность (профиль) подготовки: Использование открытого программного обеспечения для разработки геосервисов

Квалификация (степень) выпускника: магистр

Форма обучения: очная

Год набора (приема на обучение): 2025

Срок получения образования: 1 года 10 месяца(-ев)

Объем: в зачетных единицах: 3 з.е.
в академических часах: 108 ак.ч.

Москва, 2025

Разработчики:

Заведующий кафедрой теоретических основ информатики, д.т.н., профессор Замятин А. В.

Ассистент кафедры теоретических основ информатики Карев С. В.

Фонд оценочных средств по дисциплине (модулю) составлен в соответствии с требованиями ФГОС ВО по направлению подготовки 09.04.02 Информационные системы и технологии, утвержденного приказом Минобрнауки от 19.09.2017 № 917

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

1. Планируемые результаты обучения, соотнесенные с планируемыми результатами освоения образовательной программы

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

ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий

Знать:

ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных

ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач

ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы

Уметь:

ОПК-2.1/Ум1 Работать с системами хранения, обработки и анализа пространственных данных

ОПК-2.1/Ум2 Формализовать прикладную задачу из профессиональной области в терминах интеллектуального анализа данных

ОПК-2.1/Ум3 Проектировать блок-схемы и псевдокод оригинальных алгоритмов, комбинируя различные методы для решения нестандартных задач

Владеть:

ОПК-2.1/Нв1 Навыками комплексно использовать геоинформационные, телекоммуникационные и мультимедийные технологии

ОПК-2.1/Нв2 Методологией экспериментирования: навыком сравнения разработанного алгоритма с базовыми моделями с помощью статистических тестов и метрик качества

ОПК-2.1/Нв3 Инструментарием визуализации этапов работы алгоритма для верификации корректности его логики

ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей

Знать:

ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных

ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей

ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора

Уметь:

ОПК-2.2/Ум1 Проектировать и реализовывать нейронные сети для задач ГИС, обрабатывать геоданные

ОПК-2.2/Ум2 Разрабатывать программный код для создания, компиляции и обучения нейросетевых моделей под конкретный тип данных

ОПК-2.2/Ум3 Интегрировать обученную нейросетевую модель в простое программное приложение или скрипт для автоматизации решения задачи

Владеть:

ОПК-2.2/Нв1 Инструментами разработки ПО, методами визуализации данных

ОПК-2.2/Нв2 Практическими навыками работы с современными фреймворками глубокого обучения для построения прототипов программных средств

ОПК-2.2/Нв3 Навыками диагностики состояния нейросети в процессе обучения и методами борьбы с ними

2. Показатели и критерии оценивания компетенций

Планируемые результаты освоения программы	Уровень овладения	Критерии оценивания (перечень сформированных ИДК/ЗУВ по уровням освоения)
ОПК-2 Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач	Повышенный	ИДК ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий Знать: ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы Уметь: ОПК-2.1/Ум1 Работать с системами хранения, обработки и анализа пространственных данных ОПК-2.1/Ум2 Формализовать прикладную задачу из профессиональной области в терминах интеллектуального анализа данных ОПК-2.1/Ум3 Проектировать блок-схемы и псевдокод оригинальных алгоритмов, комбинируя различные методы для решения нестандартных задач Владеть: ОПК-2.1/Нв1 Навыками комплексно использовать геоинформационные, телекоммуникационные и мультимедийные технологии ОПК-2.1/Нв2 Методологией экспериментирования: навыком сравнения разработанного алгоритма с базовыми моделями с помощью статистических тестов и метрик качества ОПК-2.1/Нв3 Инструментарием визуализации этапов работы алгоритма для верификации корректности его логики ИДК ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей Знать: ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных

		Уметь: ОПК-2.2/Ум1 Проектировать и реализовывать нейронные сети для задач ГИС, обрабатывать геоданные ОПК-2.2/Ум2 Разрабатывать программный код для создания, компиляции и обучения нейросетевых моделей под конкретный тип данных ОПК-2.2/Ум3 Интегрировать обученную нейросетевую модель в простое программное приложение или скрипт для автоматизации решения задачи Владеть: ОПК-2.2/Нв1 Инструментами разработки ПО, методами визуализации данных ОПК-2.2/Нв2 Практическими навыками работы с современными фреймворками глубокого обучения для построения прототипов программных средств ОПК-2.2/Нв3 Навыками диагностики состояния нейросети в процессе обучения и методами борьбы с ними
	Базовый	ИДК ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий Знать: ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных Уметь: ОПК-2.1/Ум3 Проектировать блок-схемы и псевдокод оригинальных алгоритмов, комбинируя различные методы для решения нестандартных задач ОПК-2.1/Ум2 Формализовать прикладную задачу из профессиональной области в терминах интеллектуального анализа данных ОПК-2.1/Ум1 Работать с системами хранения, обработки и анализа пространственных данных ИДК ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей Знать: ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных

		Уметь: ОПК-2.2/Ум1 Проектировать и реализовывать нейронные сети для задач ГИС, обрабатывать геоданные ОПК-2.2/Ум3 Интегрировать обученную нейросетевую модель в простое программное приложение или скрипт для автоматизации решения задачи ОПК-2.2/Ум2 Разрабатывать программный код для создания, компиляции и обучения нейросетевых моделей под конкретный тип данных
	Пороговый	ИДК ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий Знать: ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных ИДК ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей Знать: ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных

3. Шкала оценивания

Оценка	Критерии оценивания (перечень сформированных ИДК/ЗУВ)

Отлично	ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы ОПК-2.1/Ум1 Работать с системами хранения, обработки и анализа пространственных данных ОПК-2.1/Ум2 Формализовать прикладную задачу из профессиональной области в терминах интеллектуального анализа данных ОПК-2.1/Ум3 Проектировать блок-схемы и псевдокод оригинальных алгоритмов, комбинируя различные методы для решения нестандартных задач ОПК-2.1/Нв1 Навыками комплексно использовать геоинформационные, телекоммуникационные и мультимедийные технологии ОПК-2.1/Нв2 Методологией экспериментирования: навыком сравнения разработанного алгоритма с базовыми моделями с помощью статистических тестов и метрик качества ОПК-2.1/Нв3 Инструментарием визуализации этапов работы алгоритма для верификации корректности его логики ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора ОПК-2.2/Ум1 Проектировать и реализовывать нейронные сети для задач ГИС, обрабатывать геоданные ОПК-2.2/Ум2 Разрабатывать программный код для создания, компиляции и обучения нейросетевых моделей под конкретный тип данных ОПК-2.2/Ум3 Интегрировать обученную нейросетевую модель в простое программное приложение или скрипт для автоматизации решения задачи ОПК-2.2/Нв1 Инструментами разработки ПО, методами визуализации данных ОПК-2.2/Нв2 Практическими навыками работы с современными фреймворками глубокого обучения для построения прототипов программных средств ОПК-2.2/Нв3 Навыками диагностики состояния нейросети в процессе обучения и методами борьбы с ними

Хорошо	ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы ОПК-2.1/Ум1 Работать с системами хранения, обработки и анализа пространственных данных ОПК-2.1/Ум2 Формализовать прикладную задачу из профессиональной области в терминах интеллектуального анализа данных ОПК-2.1/Ум3 Проектировать блок-схемы и псевдокод оригинальных алгоритмов, комбинируя различные методы для решения нестандартных задач ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора ОПК-2.2/Ум1 Проектировать и реализовывать нейронные сети для задач ГИС, обрабатывать геоданные ОПК-2.2/Ум2 Разрабатывать программный код для создания, компиляции и обучения нейросетевых моделей под конкретный тип данных ОПК-2.2/Ум3 Интегрировать обученную нейросетевую модель в простое программное приложение или скрипт для автоматизации решения задачи
Удовлетворительно	ОПК-2.1 Разрабатывает оригинальные алгоритмы с использованием интеллектуальных систем и технологий ОПК-2.1/Зн1 Способы обработки и представления геопространственных данных ОПК-2.1/Зн2 Классификацию задач интеллектуального анализа данных и формальные постановки этих задач ОПК-2.1/Зн3 Методы предварительной обработки и преобразования данных и их влияние на алгоритмы ОПК-2.2 Разрабатывает программные средства с использованием искусственных нейронных сетей ОПК-2.2/Зн1 Основы теории нейронных сетей, архитектуры и алгоритмы обучения, методы предобработки данных ОПК-2.2/Зн2 Особенности подготовки данных для нейросетевых моделей ОПК-2.2/Зн3 Типовые архитектуры нейросетей, применимые для решения задач классификации и регрессии в рамках вводного курса, и критерии их выбора
Неудовле-творительно	Результаты обучения не сформированы на достаточном уровне

4. Контрольные мероприятия по дисциплине

Вид контроля	Форма контроля/Оценочное средство
Текущий контроль	Тестирование
Промежуточная аттестация	Экзамен

№ п/п	Наименование раздела	Контролируемые ИДК	Вид контроля/ используемые оценочные материалы	
			Текущий	Промежут. аттестация
1	Основные проблемы построения систем	ОПК-2.1 ОПК-2.2	Тестирование	Экзамен
2	Предварительная обработка данных. Классификация	ОПК-2.1 ОПК-2.2	Тестирование	Экзамен
3	Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация	ОПК-2.1 ОПК-2.2	Тестирование	Экзамен
4	Высокопроизводительная обработка данных	ОПК-2.1 ОПК-2.2	Тестирование	Экзамен

5. 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, 2 семестр.

Экзамен оценивается по пятибалльной шкале:

- зачтено с оценкой «отлично» выставляется, если обучающийся имеет глубокие, исчерпывающие знания в объеме освоенной программы; знание основной (обязательной) литературы; дает правильные и уверенные ответы, свидетельствующие о наличии твердых знаний и навыков в использовании технических средств; полное, четкое, грамотное и логически стройное изложение материала; демонстрирует свободное применение теоретических знаний при анализе практических вопросов;
- зачтено с оценкой «хорошо» выставляется обучающемуся в случае соблюдения тех же требований, но наличии в ответе обучающегося по некоторым перечисленным показателям недостатков принципиального характера, что вызывает замечания или поправки преподавателя;
- зачтено с оценкой «удовлетворительно» выставляется обучающемуся в случае соблюдения тех же требований, но в случае наличия в ответе ошибок, необходимости существенной помощи в виде поправок и наводящих вопросов преподавателя;
- не зачтено («не удовлетворительно») – обучающийся не смог ответить на 2/3 вопросов; не справился с практическим заданием или выполнено менее 50 % практических заданий.

6. Оценочные материалы текущего контроля

Раздел 1. Основные проблемы построения систем

Контролируемые ИДК: ОПК-2.1 ОПК-2.2

Тема 1.1. Основные проблемы построения систем

Форма контроля/оценочное средство: Тестирование

Вопросы/Задания:

1. Что является основной целью интеллектуального анализа данных (Data Mining)?

1. Простая констатация фактов, произошедших в прошлом
2. Хранение больших объемов информации в структурированном виде
3. Выявление скрытых закономерностей и взаимосвязей для прогнозирования
4. Замена экспертов-аналитиков искусственным интеллектом

2. Какой из перечисленных методов относится к обучению без учителя (unsupervised learning)?

1. Регрессионный анализ
2. Кластеризация
3. Классификация методом опорных векторов
4. Прогнозирование временных рядов

3. Задача обнаружения аномалий (outlier detection) направлена на:

1. Поиск часто повторяющихся последовательностей событий
2. Выявление объектов, которые существенно отличаются от основной массы данных
3. Снижение размерности признакового пространства
4. Автоматическую очистку данных от пропусков

4. Что из перечисленного НЕ является этапом предварительной обработки данных?

1. Очистка данных от шумов
2. Трансформация и нормализация данных
3. Построение нейросетевой модели
4. Обработка пропущенных значений

5. Задача «Ассоциация» в интеллектуальном анализе данных подразумевает:

1. Объединение нескольких таблиц баз данных в одну
2. Поиск часто встречающихся наборов объектов или правил «если-то»
3. Построение линии тренда
4. Распределение данных по нескольким серверам

6. Какой тренд развития характерен для современных систем анализа данных?

1. Отказ от визуализации результатов
2. Переход от больших данных к малым выборкам
3. Интеграция методов машинного обучения и высокопроизводительных вычислений
4. Использование только статистических методов XIX века

7. Какая программная среда НАИБОЛЕЕ популярна для задач прототипирования и интеллектуального анализа данных благодаря богатству библиотек (Scikit-learn, Pandas)?

1. Microsoft Excel
2. Python
3. Adobe Photoshop
4. AutoCAD

Раздел 2. Предварительная обработка данных. Классификация

Контролируемые ИДК: ОПК-2.1 ОПК-2.2

Тема 2.1. Предварительная обработка данных. Классификация

Форма контроля/оценочное средство: Тестирование

Вопросы/Задания:

1. Для чего используется нормализация данных?

1. Для удаления всех числовых значений из набора данных
2. Для приведения признаков к единому масштабу (например, от 0 до 1)
3. Для увеличения количества признаков

4. Для преобразования текста в изображения

2. Что такое «редукция признакового пространства» без трансформации?

1. Создание новых признаков на основе существующих (PCA)
2. Отбор наиболее информативных существующих признаков
3. Увеличение размерности данных
4. Автоматическая генерация синтетических данных

3. Какой алгоритм классификации относится к непараметрическим?

1. Линейная регрессия
2. Метод k-ближайших соседей (k-NN)
3. Наивный байесовский классификатор с известным распределением
4. Линейный дискриминантный анализ

4. Основная идея метода опорных векторов (SVM) заключается в:

1. Построении нескольких деревьев решений и их усреднении
2. Построении гиперплоскости, максимизирующей зазор (отступ) между классами
3. Минимизации среднеквадратичной ошибки на обучающей выборке
4. Итеративном уточнении весов нейронов

5. Что является результатом работы дерева решений для задачи классификации?

1. Уравнение прямой линии
2. Иерархическая структура правил (узлы и листья)
3. Матрица ковариации
4. Набор центроидов кластеров

6. В чем заключается опасность переобучения (overfitting) дерева решений?

1. Дерево слишком плохо классифицирует объекты
2. Дерево требует слишком много памяти для хранения
3. Дерево слишком точно запоминает шум в обучающих данных и плохо обобщает новые
4. Дерево строится слишком долго

7. При нетранзитивной редукции признакового пространства оценка информативности признака производится:

1. С учетом всех возможных комбинаций признаков
2. Независимо от других признаков
3. Только для бинарных признаков
4. После преобразования признаков в главные компоненты

8. Какой критерий чаще всего используется для оценки качества разделения в узлах дерева решений?

1. Среднеквадратичная ошибка (MSE)
2. Коэффициент детерминации (R^2)
3. Энтропия или индекс Джини
4. F-мера

9. Контролируемая классификация означает, что:

1. Процесс обучения контролируется администратором базы данных
2. Для обучения модели используются размеченные данные (известны правильные ответы)
3. Модель работает только с числовыми данными
4. Алгоритм самостоятельно разбивает данные на кластеры

10. Какая операция НЕ относится к предварительной обработке данных?

1. Удаление дубликатов записей
2. Замена пропущенных значений средним или модой
3. Фильтрация выбросов (аномалий)
4. Настройка гиперпараметров нейронной сети

Раздел 3. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация

Контролируемые ИДК: ОПК-2.1 ОПК-2.2

Тема 3.1. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация

Форма контроля/оценочное средство: Тестирование

Вопросы/Задания:

1. Основная цель регрессионного анализа:

1. Разделение данных на группы по схожести Основная цель регрессионного анализа:
2. Поиск правил "если-то"
3. Моделирование зависимости между зависимой переменной и одной или несколькими независимыми переменными
4. Визуализация многомерных данных на плоскости

2. Какой алгоритм относится к поиску ассоциативных правил?

1. Метод опорных векторов (SVM)
2. Apriori
3. K-means
4. Градиентный бустинг

3. Алгоритм GSP (Generalized Sequential Patterns) используется для решения задачи:

1. Классификации текстов
2. Поиска последовательных (sequential) ассоциаций
3. Детекции аномалий в реальном времени
4. Построения регрессионных кривых

4. Понятие «Поддержка» (support) в алгоритме Apriori для правила X → Y означает:

1. Достоверность правила
2. Долю транзакций, содержащих одновременно X и Y
3. Количество транзакций в базе данных
4. Вероятность того, что Y не встретится без X

5. Метод визуализации t-SNE (t-distributed Stochastic Neighbor Embedding) чаще всего используется для:

1. Отображения линий регрессии
2. Визуализации многомерных данных на двумерной плоскости
3. Построения деревьев решений
4. Отображения графов ассоциативных правил

6. Что является примером аномалии в данных о температуре больных в стационаре?

1. Значение температуры 36.6°C у большинства пациентов
2. Значение температуры 42.5°C у выздоравливающего пациента
3. Монотонный ряд температур в течение суток
4. Запись о температуре в текстовом поле "комментарий"

7. Основное отличие регрессии от классификации:

1. В регрессии используется меньше данных
2. Регрессия работает только с текстом
3. В регрессии предсказывается непрерывное числовое значение, а в классификации — метка класса
4. Регрессия — это метод обучения без учителя, а классификация — с учителем

8. Какая библиотека чаще всего используется для визуализации данных в Python?

1. NumPy
2. Matplotlib
3. Scikit-learn
4. TensorFlow

Раздел 4. Высокопроизводительная обработка данных

Контролируемые ИДК: ОПК-2.1 ОПК-2.2

Тема 4.1. Высокопроизводительная обработка данных

Форма контроля/оценочное средство: Тестирование

Вопросы/Задания:

1. Что такое SMP-система (Symmetric Multiprocessing)?

1. Система с асимметричным доступом к жесткому диску
2. Многопроцессорная система с общей памятью, где все процессоры равноправны
3. Кластер слабосвязанных компьютеров без общей памяти
4. Система для последовательной обработки данных

2. Расшифровка аббревиатуры MPMD (Multiple Program Multiple Data):

1. Одна программа работает с несколькими потоками данных
2. Несколько различных программ параллельно обрабатывают свои наборы данных
3. Одна программа копируется на несколько машин с одними данными
4. Поток команд шире, чем поток данных

3. Модель SPMD (Single Program Multiple Data) предполагает, что:

1. Каждый процессор выполняет свою уникальную программу
2. Все процессоры выполняют одну и ту же программу, но на разных данных
3. Данные хранятся в одном экземпляре и копируются по запросу
4. Программа пишется только на языке Java

4. Какой фреймворк чаще всего используется для распределенной обработки больших данных в модели MapReduce?

1. Apache Hadoop
2. Microsoft Word
3. Adobe Premiere
4. MySQL (в одиночном режиме)

5. Что позволяет обеспечить масштабирование вычислений при интеллектуальном анализе больших данных?

1. Использование только одного мощного суперкомпьютера
2. Применение параллельных алгоритмов и распределенных файловых систем
3. Запрет на использование облачных технологий
4. Хранение всех данных в оперативной памяти одного сервера

7. Оценочные материалы промежуточной аттестации

2 семестр, Экзамен

Контролируемые ИДК: ОПК-2.1 ОПК-2.2

Вопросы/Задания:

1. Дайте определение интеллектуального анализа данных (Data Mining). Объясните, чем вызван рост актуальности этой дисциплины в последние десятилетия. Какие тенденции развития систем анализа данных наблюдаются сегодня?

2. Перечислите и кратко охарактеризуйте основные задачи, решаемые методами интеллектуального анализа (классификация, регрессия, кластеризация, ассоциация, обнаружение аномалий). Приведите примеры практических ситуаций для каждой задачи.

3. Опишите типовой процесс проведения анализа данных (например, методология CRISP-DM). Какие этапы входят в этот процесс? Какова роль предварительной обработки данных?

4. Какие проблемы качества данных (пропуски, шум, выбросы, дубликаты) встречаются на практике? Опишите методы их обнаружения и устранения (очистка, трансформация, нормализация, дискретизация).

5. В чем принципиальное различие между обучением с учителем и без учителя? Приведите примеры алгоритмов для каждого подхода. Когда целесообразно применять тот или иной тип обучения?

6. Что такое «проклятие размерности»? Объясните разницу между отбором признаков (feature selection) и выделением признаков (feature extraction). Опишите метод главных компонент (PCA) как пример выделения признаков.

7. Опишите алгоритм построения дерева решений для классификации. Какие критерии (энтропия, индекс Джини) используются для выбора разбиения? Как интерпретировать полученное дерево?

8. Дайте определения переобучения (overfitting) и недообучения (underfitting). Какие методы позволяют предотвратить переобучение в деревьях решений (стрижка, ограничение глубины) и в нейронных сетях (регуляризация, dropout, early stopping)?

9. Объясните основную идею метода опорных векторов для задачи классификации. Что такое максимизация зазора? Зачем используются ядра (kernel functions) и какие виды ядер существуют?

10. Опишите структуру многослойного персептрона (входной, скрытые, выходной слои). Каким образом происходит обучение такой сети с помощью алгоритма обратного распространения ошибки?

11. Перечислите основные метрики для оценки качества моделей классификации: accuracy, precision, recall, F-мера, ROC-AUC. Объясните, в каких ситуациях каждая из них наиболее информативна. Что такое матрица ошибок (confusion matrix)?

12. Что такое регрессия? Чем она отличается от классификации? Опишите основные этапы регрессионного анализа: выбор модели, оценка параметров, проверка адекватности.

13. Какие метрики используются для оценки точности регрессионных моделей? Поясните смысл среднеквадратичной ошибки (MSE), средней абсолютной ошибки (MAE) и коэффициента детерминации R^2.

14. Кратко охарактеризуйте линейную, полиномиальную, гребневую (ridge) регрессию и LASSO. В каких случаях применяются регуляризованные версии регрессии?

15. Дайте определение ассоциативным правилам. Где они применяются (рыночная корзина, рекомендательные системы)? Поясните понятия поддержки (support), достоверности (confidence) и лифта (lift).

16. Опишите принцип работы алгоритма Apriori для поиска часто встречающихся наборов. В чем заключается свойство анти-монотонности (Apriori property) и как оно используется для сокращения перебора?

17. Что понимают под последовательными паттернами (sequential patterns)? Чем алгоритм GSP отличается от Apriori? Приведите пример задачи, требующей поиска последовательностей.

18. Какие типы аномалий (точечные, контекстуальные, коллективные) существуют? Опишите основные подходы к обнаружению выбросов: статистические, на основе расстояния, методы машинного обучения (одноклассовая SVM, изолирующий лес).

19. Какова роль визуализации в процессе анализа? Какие методы используются для визуализации многомерных данных (t-SNE, параллельные координаты, матрицы диаграмм рассеяния)? Для чего нужна визуализация промежуточных результатов?

20. Почему для анализа больших данных необходимы высокопроизводительные вычисления? Опишите основные принципы организации параллельной обработки данных.

21. Что такое симметричные мультипроцессорные системы (SMP)? Каковы их преимущества и ограничения при решении задач анализа данных?

22. Объясните различия между моделями MPMD (Multiple Program Multiple Data) и SPMD (Single Program Multiple Data). Приведите примеры, когда каждая из них эффективна.

23. Какие программные средства используются для интеллектуального анализа данных? Опишите экосистему Python (Pandas, Scikit-learn, TensorFlow, PyTorch). Для чего применяются платформы Apache Hadoop и Apache Spark?

24. Сравните деревья решений, метод опорных векторов и нейронные сети по следующим критериям: интерпретируемость, требования к объему данных, склонность к переобучению, вычислительная сложность. В каких случаях какой метод предпочтителен?

25. Как оценивать качество модели, если классы в данных сильно несбалансированы? Какие метрики (PR-AUC, F-мера) и подходы (взвешивание классов, аугментация, SMOTE) используются для работы с дисбалансом?

26. Опишите процесс подготовки данных перед подачей в нейросеть: нормализация/стандартизация, кодирование категориальных признаков (one-hot encoding), разбиение на обучающую, валидационную и тестовую выборки.

27. Что такое ансамблирование моделей? Поясните принципы бэггинга (на примере случайного леса) и бустинга (на примере градиентного бустинга). В чем их преимущества перед одиночными моделями?

28. Какие проблемы возникают при анализе больших данных (Big Data)? Как высокопроизводительные вычисления и распределенные системы (Hadoop, Spark) помогают их решать?

29. Выберите одну из предметных областей (финансы, медицина, маркетинг, производство) и опишите, какие задачи интеллектуального анализа данных там востребованы. Приведите пример конкретного применения.

30. Каковы современные тенденции в области интеллектуального анализа данных? Какую роль играют глубокое обучение (deep learning) и автоматизированное машинное обучение (AutoML)? Какие новые вызовы стоят перед специалистами?