

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДЕНО:
Директор
А. В. Замятин

Оценочные материалы по дисциплине

Анализ больших массивов данных (Big Data)

по направлению подготовки

01.03.02 Прикладная математика и информатика

Направленность (профиль) подготовки:
Математические методы в цифровой экономике

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
К.И. Лившиц

Председатель УМК
С.П. Сущенко

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-2. Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач.

ОПК-4. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности.

ПК-1. Способен осуществлять научно-исследовательские и опытно-конструкторские разработки как по отдельным разделам темы, так и при исследовании самостоятельных тем.

ПК-2. Способен анализировать и оценивать риски, разрабатывать отдельные функциональные направления управления рисками.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-2.1. Обладает навыками объектно-ориентированного программирования для решения прикладных задач в профессиональной деятельности.

ИОПК-2.2. Проявляет навыки использования основных языков программирования, основных методов разработки программ, стандартов оформления программной документации.

ИОПК-2.3. Демонстрирует умение отбора среди существующих математических методов, наиболее подходящих для решения конкретной прикладной задачи.

ИОПК-2.4. Демонстрирует умение адаптировать существующие математические методы для решения конкретной прикладной задачи.

ИОПК-4.1. Обладает необходимыми знаниями в области информационных технологий, в том числе понимает принципы их работы

ИОПК-4.2. Применяет знания, полученные в области информационных технологий, при решении задач профессиональной деятельности.

ИПК-1.1. Осуществляет проведение работ по обработке и анализу научно-технической информации и результатов исследований.

ИПК-2.2. Собирает и обрабатывает аналитическую информацию для анализа и оценки рисков.

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

- проект;
- лабораторные работы;
- реферат.

Реферат (на согласованную тему). К реферату необходимо сделать презентацию.

Примеры тем:

1. Современные нейронные сети в обработке данных (изображений, видео, технологических сигналов, музыки и т.п.).
2. Современные алгоритмы классификации (изображений, текстов и т.п.).
3. Интеллектуальная обработка данных в ... (промышленности, медицине, бизнесе, индустрии развлечений, досуга и др.).
4. Извлечение знаний из текстов.
5. Детектирование аномалий.
6. Разновидности сверточных нейронных сетей.
7. Интеллектуальные алгоритмы в ранней диагностике заболеваний.
8. Интеллектуальные алгоритмы в персонализированной медицине.
9. Интеллектуальные алгоритмы в робототехнике, транспортных системах и т.п..

10. Интеллектуальные алгоритмы в банковском деле/страховании/....

Проект (на согласованную тему). Реализовать небольшой проект по интеллектуальной обработке данных с использованием среды RapidMiner или одного из языков программирования (например, Python, R), с возможным использованием общедоступных баз данных (или данных из иных источников).

Этапы реализации проекта:

Поиск и подготовка набора данных;

Разработка технического задания;

Пилотная реализация одной модели, выбор метрики и оценка точности (фиксация полученной точности на этом этапе);

Реализация всех пунктов технического задания, настройка параметров моделей, оценка точности (точность, полученная на этом этапе должна быть больше чем на предыдущем):

Подготовка отчета (с описанием предметной области, выбранных алгоритмов и параметров моделей), презентации, публичная защита проекта;

Каждый студент реализует индивидуальный или групповой проект как последовательность лабораторных работ:

Лабораторная работа №1. Индивидуальное задание по теме «Анализ предметной области, формулировка целей и задач исследования. Извлечение и первичное сохранение данных».

Цель работы – научить студентов решать задачи анализа предметной области, ее адаптации для методов анализа данных с учетом принципиальных особенностей предметной области.

Лабораторная работа №2. Индивидуальное задание по теме «Предварительная обработка данных: очистка, интеграция, преобразование».

Цель работы – научить студентов решать задачи предварительной обработки данных, предполагающей трудоемкую процедуру очистки (исключение противоречий, случайных выбросов и помех, пропусков), интеграции (объединение данных из нескольких возможных источников в одном хранилище), преобразования (может включать агрегирование и сжатие данных, дискретизацию атрибутов и сокращение размерности и т.п.).

Лабораторная работа №3. Индивидуальное задание по теме «Содержательный анализ данных методами Data Mining».

Цель работы – научить студентов обоснованно применять базовые методы интеллектуального анализа данных, учитывая особенности как теоретического построения применяемых методов, так и выбранной предметной области.

Лабораторная работа №4. Индивидуальное задание по теме «Визуализация и интерпретация полученных результатов».

Цель работы – научить студентов выполнять визуализацию и интерпретация полученных результатов в виде, пригодном для принятия управленческих решений.

Примеры тем для самостоятельного изучения:

- Нейросетевые методы анализа данных, сверточные сети (convolution neural networks). глубинное обучение (deep learning).
- Методы интеллектуального анализа медиа (social media data mining).
- Методы машинного обучения в задачах финансовой аналитики.
- Методы машинного обучения в задачах ранней медицинской диагностики.
- Комбинирование моделей в анализе данных, бустинг.
- Метод анализа независимых компонент (independent component analysis).
- Методы визуализации данных высокой размерности.

Критерии оценивания реферата:

Результаты подготовки и защиты реферата определяются оценками «зачтено», «не зачтено».

Оценка «зачтено» выставляется, если реферат подготовлен на достаточно высоком методическом уровне, а тема в достаточной степени раскрыта в пояснительной записке.

Критерии оценивания лабораторных работ:

Оценка «зачтено» выставляется, если лабораторные работы выполнены и защищены в полном объеме.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Видом промежуточной аттестации является зачет. Билет состоит из трёх частей, каждая в виде вопроса по одной из тем, освещенной на лекциях.

Вопросы к зачету:

1. Основные понятия, терминология;
2. Data Mining / Data Science;
3. Big Data (основные понятия и свойства);
4. Дедукция и индукция;
5. Интеллектуальный анализ данных в бизнесе примеры применения;
6. Интеллектуальный анализ данных в решении сложных прикладных задач;
7. Интеллектуальный анализ данных в ранней диагностике опасных заболеваний;
8. Интеллектуальный анализ данных в индустриальной предиктивной аналитике;
9. Основные задачи и классификация методов анализа данных;
10. Принципиальные основы машинного обучения;
11. Предварительная обработка данных;
12. Оптимизация признаков пространства;
13. Постановка задачи классификации;
14. Контролируемая непараметрическая классификация;
15. Контролируемая непараметрическая нейросетевая классификация;
16. Классификация по методу машины опорных векторов;
17. Деревья решений;
18. Неконтролируемая классификация (кластеризация);
19. Регрессия (понятие регрессии, основные этапы регрессионного анализа, методы восстановления регрессии);
20. Ассоциация;
21. Последовательная ассоциация (алгоритмы семейства «Априори», алгоритм GSP);
22. Многоуровневое машинное обучение (бутстрэппинг, бэггинг, стекинг, бустинг);
23. Обнаружение аномалий;
24. Визуализация в Data Mining;
25. Функции активации;
26. Основные типы искусственных нейронных сетей;
27. Сверточные нейронные сети;
28. Среды и фреймворки глубинного обучения;
29. Основные задачи обработки текста;
30. Этапы предварительной обработки текста;
31. Метрики качества классификации;
32. Гипотеза A/B, Каппа-индекс согласия, ROC-кривая;
33. Метрика качества прогноза временного ряда;
34. Метрики качества кластеризации;
35. Принципы высокопроизводительных вычислений;

36. Особенности построения вычислительного кластера;
37. Среды и инструменты высокопроизводительных вычислений;
38. Инструменты data mining.

Балльные оценки для элементов контроля

Элементы учебной деятельности	Максимальный балл с начала семестра
Реферат по теме с презентацией	20
Реализация проекта	40
Опрос на занятиях	10
Зачет	30

Сумма баллов, набранная студентом в течение семестра, переводится в оценку промежуточной аттестации успеваемости студента по приведенной ниже шкале.

Пересчет баллов в оценки для промежуточной аттестации

Баллы на дату контрольной точки	Оценка
$\geq 90\%$ от максимальной суммы баллов	5 (зачтено)
От 70% до 89% от максимальной суммы баллов	4 (зачтено)
От 60% до 69% от максимальной суммы баллов	3 (зачтено)
$< 60\%$ от максимальной суммы баллов	2 (не зачтено)

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Проверка остаточных знаний проводится в форме теста.

Примеры вопросов к тесту:

1. Выберите все верные утверждения:
 - а) Последовательность имеет место в случае, если несколько событий связаны друг с другом.
 - б) Отнесение нового объекта к какому-либо из существующих классов выполняется путем классификации.
 - в) В случае, если несколько событий связаны друг с другом во времени, имеет место тип зависимости, именуемый ассоциация.
 - г) Хранимая ретроспективная информация позволяет определить еще одну закономерность, заключающуюся в поиске существующих кластеров.
2. Выявление лояльных или нелояльных держателей кредитных карт относится к задаче
 - а) Контролируемой классификации
 - б) Ассоциации
 - в) Прогнозирования
 - г) Неконтролируемой классификации
3. Для построения алгоритма машинного обучения требуется три типа выборок:
 - а) Обучающая
 - б) Валидационная
 - в) Тестовая
 - г) Стратифицированная

- д) Квотная
- 4. Нейросетевые классификаторы относят к:
 - а) Параметрическим подходам
 - б) Непараметрическим подходам
 - в) Прагматическим подходам
 - г) Эклектический подход
- 5. Индуктивный подход ...
 - а) к исследованию данных позволяет сформулировать гипотезу и найти с ее помощью новые пути аналитических решений
 - б) к исследованию данных предполагает наличие некоторой сформулированной гипотезы, подтверждение или опровержение которой после анализа данных позволяет получить некоторые частные сведения

Ключи: 1 в) 2 г) 3 а)б)в) 4 б) 5 а)

Информация о разработчиках

Замятин Александр Владимирович, д-р техн. наук, профессор, заведующий кафедрой теоретических основ информатики института прикладной математики и компьютерных наук НИТГУ, директор ИПМКН.