

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:
Математическое моделирование и информационные системы

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
К.И. Лившиц

Председатель УМК
С.П. Сущенко

Томск – 2024

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-3 Способен формализовывать, согласовывать и документировать требования к системе и подсистеме, обрабатывать запросы на изменение требований к системе и подсистеме, выявлять и формализовывать риски, анализировать проблемные ситуации..

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-3.1 Реализовывает построение формализованной математической модели системы (подсистемы), введение целевой функции системы, подсистемы и ограничений, соответствующих требованиям к системе (подсистеме).

ИПК-3.2 Адаптирует формализованную математическую модель системы (подсистемы) к изменению требований (ограничений к целевой функции) к системе (подсистеме).

ИПК-3.3 Выявляет и формализовывает в виде математической модели возникающие при функционировании системы (подсистемы) риски; выявляет и анализирует проблемные ситуации.

2. Оценочные материалы текущего контроля и критерии оценивания

Типовые задания для проведения текущего контроля успеваемости по дисциплине

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

Примеры тем:

Современные нейронные сети в обработке данных (изображений, видео, технологических сигналов, музыки и т.п.);

Современные алгоритмы классификации (изображений, текстов и т.п.);

Интеллектуальная обработка данных в ... (промышленности, медицине, бизнесе, индустрии развлечений, досуга и др.);

Извлечение знаний из текстов;

Детектирование аномалий;

Разновидности сверточных нейронных сетей;

Интеллектуальные алгоритмы в ранней диагностике заболеваний;
Интеллектуальные алгоритмы в персонализированной медицине;
Интеллектуальные алгоритмы в робототехнике, транспортных системах и т.п.;
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.

Методические материалы для оценки текущего контроля успеваемости по дисциплине.

Рейтинговая система для оценки текущей успеваемости обучающихся

Таблица – Балльные оценки для элементов контроля

Элементы учебной деятельности	Максимальный балл с начала семестра	Оцениваемая компетенция
Реферат по теме с презентацией	20	ОПК-2, ОПК-4, ПК-3
Реализация проекта	40	ОПК-2, ОПК-4, ПК-3
Опрос на занятиях	10	ОПК-2, ОПК-4, ПК-3
Зачет с оценкой	30	ОПК-2, ОПК-4, ПК-3

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

Типовые задания для проведения промежуточной аттестации по дисциплине

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 признакового 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.

Методические материалы для проведения промежуточной аттестации по дисциплине.

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
≥ 65% от максимальной суммы баллов	зачтено
< 65% от максимальной суммы баллов	не зачтено

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

Список вопросов для оценки остаточных знаний

1. Основные понятия, терминология;
2. Дедукция и индукция;
3. Интеллектуальный анализ данных в ранней диагностике опасных заболеваний;
4. Принципиальные основы машинного обучения;
5. Постановка задачи классификации;
6. Классификация по методу машины опорных векторов;
7. Регрессия (понятие регрессии, основные этапы регрессионного анализа, методы восстановления регрессии);
8. Многоуровневое машинное обучение (бутстрэппинг, бэггинг, стекинг, бустинг);
9. Функции активации;
10. Среды и фреймворки глубинного обучения;

Информация о разработчиках

Замятин Александр Владимирович, д-р техн. наук, профессор, заведующий кафедрой теоретических основ информатики ТГУ, директор ИПМКН.