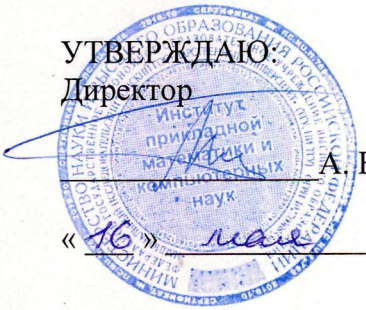
Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук



УТВЕРЖДАЮ:
Директор
_____ А. В. Замятин
« 16 » мая 20 22 г.

Рабочая программа дисциплины

Введение в интеллектуальный анализ данных

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки :
Интеллектуальный анализ больших данных

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: Б1.О.02.03

СОГЛАСОВАНО:
Руководитель ОП
_____А.В. Замятин

Председатель УМК
_____С.П. Сущенко

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

– УК-1 – способность осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий;

– ОПК-1 – способность решать актуальные задачи фундаментальной и прикладной математики.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИУК-1.3 Предлагает и обосновывает стратегию действий с учетом ограничений, рисков и возможных последствий.

ИУК-1.2 Осуществляет поиск, отбор и систематизацию информации для определения альтернативных вариантов стратегических решений в проблемной ситуации..

ИУК-1.1 Выявляет проблемную ситуацию, на основе системного подхода осуществляет её многофакторный анализ и диагностику.

ИОПК-1.1 Анализирует проблемы в области фундаментальной и прикладной математики.

2. Задачи освоения дисциплины

– изучить основные модели и методы разработки данных;

– научиться применять указанные модели и методы, а также программные средства, в которых они реализованы;

– приобрести опыт анализа реальных данных с помощью изученных методов.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательной части образовательной программы. Дисциплина входит в модуль Общепрофессиональные дисциплины.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Первый семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:

-лекции: 16 ч.

-лабораторные: 16 ч.

в том числе практическая подготовка: 0 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Основные проблемы построения систем

Краткое содержание темы. Актуальность, базовая терминология и тенденции развития. Основные задачи, этапы и классификация методов анализа данных. Предварительная обработка данных. Классификация. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация. Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных.

Тема 2. Предварительная обработка данных. Классификация

Краткое содержание темы. Основные методы и предварительная обработка данных. Оптимизация признакового пространства без трансформации пространства признаков. Контролируемая непараметрическая нейросетевая классификация. Классификация по методу машины опорных векторов. Деревья решений. Раздел 3. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация. Понятие регрессии. Основные этапы регрессионного анализа. Описание алгоритма ассоциации. Алгоритмы семейства «Априори». Алгоритм GSP. Обнаружение аномалий и методы визуализации.

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

Краткое содержание темы. Понятие регрессии. Основные этапы регрессионного анализа. Описание алгоритма ассоциации. Алгоритмы семейства «Априори». Алгоритм GSP. Обнаружение аномалий и методы визуализации.

Тема 4. Высокопроизводительная обработка данных

Краткое содержание темы. Принципы организации высокопроизводительных вычислений. SMP-системы. Модели параллельных вычислений MPMD, SPMD. Программные среды для интеллектуального анализа данных.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится на основе контроля посещаемости, составления и защиты рефератов, работы над групповым проектом и фиксируется в форме контрольной точки не менее одного раза в семестр.

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.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен в первом семестре проводится в письменной форме по билетам. Билет содержит два теоретических вопроса. Продолжительность экзамена 1,5 часа.

Примерный перечень теоретических вопросов:

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.

Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Для оценки этапов освоения компетенции используется балльно-рейтинговая система оценивания:

Элементы учебной деятельности	Максимальный балл с начала семестра	Оцениваемая компетенция
Реферат по теме с презентацией	20	УК-1, ОПК-1
Реализация проекта	40	УК-1, ОПК-1
Опрос на занятиях	10	УК-1, ОПК-1
Экзамен	30	УК-1, ОПК-1

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

Пересчет баллов в оценки для промежуточной аттестации

Баллы на дату контрольной точки	Оценка
≥ 90% от максимальной суммы баллов	5
От 70% до 89% от максимальной суммы баллов	4
От 60% до 69% от максимальной суммы баллов	3
< 60% от максимальной суммы баллов	2

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=22102

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Замятин А.В. Введение в интеллектуальный анализ данных : учебное пособие /А. В. Замятин. - Томск : Издательский Дом Томского государственного университета , 2016. - 118 с. – URL: http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000529594.-1

– Pocket Data Mining electronic resource : Big Data on Small Devices / by Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes. - Cham : : Springer International Publishing : : Imprint: Springer, , 2014. - 108 p. – URL: http://dx.doi.org/10.1007/978-3-319-02711.-1

- Principles of Data Mining electronic resource /by Max Bramer. - London : Springer London : Imprint: Springer, 2013. - 440 p. – URL: http://dx.doi.org/10.1007/978-1-4471-4884-5

б) дополнительная литература:

– Principles of Data Mining electronic resource /by Max Bramer.Bramer, Max. London : : Springer London :: Imprint: Springer, 2013, XIV, 440 p. [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1007/978-1-4471-4884-5.

– Pocket Data Mining electronic resource : Big Data on Small Devices / /by Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes.Gaber, Mohamed Medhat. Cham : : Springer International Publishing : : Imprint: Springer, , 2014. IX, 108 p. [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1007/978-3-319-02711-1.

– Миркин Б. Г. Введение в анализ данных : учебник и практикум для бакалавриата и магистратуры : [для студентов вузов, обучающихся по инженерно-техническим, естественно-научным и экономическим направлениям и специальностям] /Б. Г. Миркин ; "Высшая школа экономики" Национальный исследовательский университет. – Москва : Юрайт , 2015. – 173 с.

– Кулаичев А.П. Методы и средства комплексного анализа данных : учебное пособие. – Москва : Форум [и др.] , 2014. – 511 с.

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– C, C++, C#, Python, R-Studio, Rapid Miner, MS Azure.;

б) информационные справочные системы:

– Data Mining for Service electronic. Berlin, Heidelberg, Imprint: Springer, Springer eBooks VIII, 291 p. 2014 (edited by Katsutoshi Yada) [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1007/978-3-642-45252-9

– Data Mining for Geoinformatics electronic resource : Methods and Applications / /edited by Guido Cervone, Jessica Lin, Nigel Waters. New York, NY : : Springer New York : : Imprint: Springer, , 2014, 166 p. [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1007/978-1-4614-7669-6

в) профессиональные базы данных *(при наличии)*:

– Университетская информационная система РОССИЯ – https://uisrussia.msu.ru/

– Единая межведомственная информационно-статистическая система (ЕМИСС) – https://www.fedstat.ru/

14. Материально-техническое обеспечение

Для материально-технического обеспечения дисциплины требуется наличие компьютерной техники с установленным соответствующим программным обеспечением и другого оборудования, поддерживающего проведение презентаций, построение проектной документации, выход в сеть Интернет.

15. Информация о разработчиках

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