

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

Рабочая программа дисциплины

Введение в интеллектуальный анализ данных

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. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-3. Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-3.1. Демонстрирует навыки применения современного математического аппарата для построения адекватных математических моделей реальных процессов, объектов и систем в своей предметной области.

ИОПК-3.2. Демонстрирует умение собирать и обрабатывать статистические, экспериментальные, теоретические и т.п. данные для построения математических моделей, расчетов и конкретных практических выводов.

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 факультативным дисциплинам и входит в факультативный модуль по выбору «Введение в искусственный интеллект».

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

Пятый семестр, зачет с оценкой.

5. Входные требования для освоения дисциплины

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

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Математическая статистика».

6. Язык реализации

Русский.

7. Объем дисциплины

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

-лекции: 32 ч.

-лабораторные: 16 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение. Основные понятия. Терминология. Области и примеры применения. Тема 2. Этапы Data Science. Машинное обучение, общая постановка задачи. CRISP-DM. Регрессия, переобучение.

Тема 3. Топологии нейросетей и задачи для них. Нейросетевая классификация, Deep Learning. Сверточные нейронные сети.

Тема 4. Кластеризация (k-means). Метрики расстояний. Критерии точности (Карра, ROC, RMSE), ошибки I/II рода, гипотеза A/B.

Тема 5. Предварительная обработка данных. Оптимизация признакового пространства

Тема 6. Классификация (деревья решений). Классификация (статистическая, байесовский подход). SVM (метод опорных векторов). Регуляризация (L1, L2).

Тема 7. Ассоциативные алгоритмы (ассоциация, последовательная ассоциация).

Тема 8. Высокопроизводительная обработка данных (принципы и модели).

Критерии эффективности.

Тема 9. Многоуровневое машинное обучение. Визуализация.

Тема 10. Обработка естественного языка. Программные среды и сервисы (Hadoop, MapReduce, Spark, Yarn, Cassandra).

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится на основе контроля посещаемости, подготовки и защиты рефератов, а также по результатам выполнения лабораторных работ и фиксируется в форме контрольной точки не менее одного раза в семестр.

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/....

Задание №1: Проведение разведочного анализа данных.

Цель: Провести разведочный анализ выбранного набора данных.

Задачи: Выбрать набор табличных данных (например на сайте kaggle.com), набор данных должен содержать не менее 10 столбцов и 1000 строк различных типов. Определить характеристики набора данных (типы данных, размер набора данных, наличие пропусков, дубликатов, категориальных переменных). Провести корреляционный анализ признаков и вывести корреляционную матрицу. Построить различные визуализации для признаков (не менее 3 различных). Сделать выводы по полученным результатам.

Описание отчета: Название «№Лабораторной_Ф_И_И_№группы», отчет в формате html/pdf должен содержать исходный код, полученные результаты, комментарии, выводы.

Задание №2: Предварительная обработка данных и оптимизация признакового пространства.

Цель: провести предварительную обработку и оптимизировать признаковое пространство.

Задачи: выбрать набор табличных данных (например на сайте Kaggle.com), набор данных должен содержать пропуски и категориальные переменные. Оценить наличие пропусков, дубликатов, категориальных переменных и их характеристики. Построить корреляционную матрицу, оценить возможность использовать ее для отбора признаков. Заполнить пропуски в данных (попробовать 3 различных способа). Закодировать

категориальные переменные (попробовать 3 различных способа). Оценить наличие выбросов (z-оценка или ящик с усами), удалить выбросы. Используя метод главных компонент, сжать признаковое пространство (до 3х компонент), визуализировать результат относительно целевой переменной (если целевой переменной нет, то в качестве ее взять один из категориальных признаков). Сделать выводы по полученным результатам.

Описание отчета: Название «№Лабораторной_Ф_И_И_№группы», отчет в формате html/pdf должен содержать исходный код, полученные результаты, комментарии, выводы.

Задание №3: Классификация данных. Неконтролируемая классификация (кластеризация).

Цель: Провести кластеризацию геоданных и классификацию табличных данных.

Задачи: для решения задачи кластеризации выбрать набор содержащий геоданные (координаты) (например на сайте Kaggle.com). Провести кластеризацию этих данных (двумя моделями), настроить гиперпараметры, оценить качество кластеризации, визуализировать результат с использованием карты. Для решения задачи классификации выбрать набор табличных данных, обучить три модели, настроить гиперпараметры моделей (использовать grid-search/random-search), оценить качество классификации, построить гос-кривую, выбрать лучшую модель. Сделать выводы по полученным результатам.

Описание отчета: Название «№Лабораторной_Ф_И_И_№группы», отчет в формате html/pdf должен содержать исходный код, полученные результаты, комментарии, выводы.

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.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» - <https://www.tsu.ru/sveden/education/eduop/>.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Видом промежуточной аттестации является зачет с оценкой. Экзаменационный билет состоит из трёх частей, каждая в виде вопроса по одной из тем, освещенной на лекциях. К зачету с оценкой допускается обучающийся, успешно выполнивший все лабораторные работы, подготовивший и защитивший реферат по теме, согласованной с преподавателем.

Темы лекционных модулей (вопросы для зачета с оценкой):

1. Введение. Основные понятия. Терминология. Области и примеры применения
2. Этапы Data Science
3. Машинное обучение, общая постановка задачи
4. CRISP-DM
5. Регрессия, переобучение

6. Топологии нейросетей и задачи для них
7. Нейросетевая классификация, Deep Learning
8. Сверточные нейронные сети
9. Кластеризация (k-means)
10. Метрики расстояний
11. Критерии точности (Карра, ROC, RMSE), ошибки I/II рода, гипотеза A/B
12. Предварительная обработка данных
13. Оптимизация признакового пространства
14. Классификация (деревья решений)
15. Классификация (статистическая, байесовский подход)
16. SVM (метод опорных векторов)
17. Регуляризация (L1, L2)
18. Ассоциативные алгоритмы (ассоциация, последовательная ассоциация)
19. Высокопроизводительная обработка данных (принципы и модели)
20. Критерии эффективности
21. Многоуровневое машинное обучение. Визуализация
22. Обработка естественного языка
23. Программные среды и сервисы (Hadoop, MapReduce, Spark, Yarn, Cassandra)

Критерии оценивания:

Результаты зачета с оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» выставляется, если даны правильные ответы на все вопросы без ошибок.

Оценка «хорошо» выставляется, если имеются незначительные неточности в ответах или незначительный дефицит в детализации ответа.

Оценка «удовлетворительно» выставляется, если имеются значительные неточности в ответах или значительный дефицит в детализации ответа.

Оценка «неудовлетворительно» выставляется, если отсутствует понимание предмета.

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» - <https://www.tsu.ru/sveden/education/eduop/>.

11. Учебно-методическое обеспечение

- а) Электронный учебный курс по дисциплине в LMS iDo;
- 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 Bartolo 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) профессиональные базы данных:

– 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>

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа и занятия для проведения лабораторных работ, оснащенные компьютерной техникой с установленным соответствующим программным обеспечением и оборудованием, поддерживающим проведение презентаций, построение проектной документации, выход в сеть Интернет.

Аудитории для проведения индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

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