

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

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. Цель и планируемые результаты освоения дисциплины

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. 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 Блоку 1 «Дисциплины (модули)».

Дисциплина относится к обязательной части образовательной программы. Дисциплина входит в модуль «Математика».

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

Восьмой семестр, зачет.

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 2 з.е., 72 часов, из которых:

-лекции: 16 ч.

-лабораторные: 16 ч.

в том числе практическая подготовка: 16 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 больших данных. Предварительная обработка данных. Классификация. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация. Высокопроизводительная обработка больших данных. Программные среды для интеллектуального анализа больших 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.

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 больших данных

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

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

Промежуточная аттестация проводится в форме зачета. Результаты зачета – оценки «зачтено», «не зачтено» проставляется по результатам сдачи практических работ.

За каждый элемент учебной деятельности (реферат по теме с презентацией; реализация проекта – лабораторных работ; опрос на занятиях; зачет) студент получает баллы, которые суммируются:

«зачтено» – студент заработал от 60% и выше от максимальной суммы баллов;

«не зачтено» – студент заработал менее 60% от максимальной суммы баллов.

Во время зачета студент может повысить свою оценку, сдав заново соответствующую лабораторную работу.

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 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) 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.

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.

Аудитории для проведения лабораторных занятий, оснащенные компьютерной техникой с установленным соответствующим программным обеспечением и другим оборудованием, поддерживающим 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 института прикладной математики и компьютерных наук НИТГУ, директор ИПМКН.