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 2022 г.

Рабочая программа дисциплины

Статистические методы машинного обучения - I

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки :
Big Data and Data Science

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: Б1.О.02

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:

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

– ОПК-2 – способность совершенствовать и реализовывать новые математические методы решения прикладных задач.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.3 Демонстрирует навыки использования основных понятий, фактов, концепций, принципов математики, информатики и естественных наук для решения практических задач, связанных с прикладной математикой и информатикой.

ИОПК-2.1 Использует результаты прикладной математики для освоения, адаптации новых методов решения задач в области своих профессиональных интересов.

ИОПК-2.2 Реализует и совершенствует новые методы, решения прикладных задач в области профессиональной деятельности.

ИОПК-2.3. Проводит качественный и количественный анализ полученного решения с целью построения оптимального варианта.

2. Задачи освоения дисциплины

- Научить студентов решать задачи статистического анализа данных, начиная от их формулирования исходных задач соответствующей предметной области на языке прикладной статистики, выбора методов решения и критериев качества полученных решений и заканчивая формулировкой полученных выводов на языке предметной области.
- Изучить основные методы статистического анализа данных.
- Сформировать навыки работы в программах статистической обработки данных.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательной части образовательной программы.

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, знание основ математического анализа, линейной алгебры, методов оптимизаций, теории вероятностей и математической статистики, а также основ программирования.

6. Язык реализации

Английский.

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 6 з.е., 216 часов, из которых:
-лекции: 20 ч.
-лабораторные: 44 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Математические основы.

Элементы линейной алгебры. Основы математического анализа. Основные вопросы методов оптимизаций. Элементы теории вероятностей.

Тема 2. Введение в статистический анализ.

Типы данных. Графические и табличные способы представления данных. Предварительная обработка данных. Оценки параметров и числовых характеристик.

Тема 3 .Критерии сравнения групп.

Параметрические критерии. t-критерий Стьюдента. Критерий Фишера. Дисперсионный анализ. Непараметрические критерии. Критерии Манна-Уитни, Вилкоксона, Краскала-Уолиса, Фридмана.

Тема 4. Корреляционный анализ.

Парный коэффициент корреляции Пирсона. Z-преобразование Фишера. Ранговая корреляция. Коэффициент Спирмена, Кендалла, конкордации Кендалла. Корреляционный анализ категоризованных данных.

Тема 5. Парная регрессия.

Определение простой регрессии. Метод наименьших квадратов оценки параметров простой регрессии. Условия Гаусса-Маркова. Теорема Гаусса-Маркова. Оценки дисперсий. Проверка качества модели регрессии, Коэффициент детерминации, его интерпретация, общая адекватность модели. Нелинейные модели и линеаризация.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, выполнения лабораторных работ, и фиксируется в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен в первом семестре проводится в форме итогового тестирования. Тест состоит из 15-20 вопросов. Продолжительность экзамена 1 академический час (45 минут).

Примерный перечень теоретических вопросов и тем для подготовки к экзамену:

1. Решение систем линейных уравнений.
2. Собственные векторы и собственные числа матрицы.
3. Функции многих переменных. Понятие градиента.
4. Метод градиентного спуска.
5. Формула полной вероятности и формула Байеса.
6. Типы данных и способы их представления.
7. Параметрические критерии сравнения групп.
8. Непараметрические критерии сравнения групп.
9. Корреляционный анализ количественных данных.
10. Ранговая корреляция.
11. Корреляционный анализ категоризованных данных.
12. Парная регрессии. Модель. МНК-оценки параметров.
13. Числовые характеристики оценок параметров парной регрессии.
14. Теорема Гаусса-Маркова для случая парной регрессии.

15. Проверка качества уравнения парной регрессии.

Примеры практических заданий по математическим основам
Задание. Решение систем линейных уравнений.
Для заданной системы уравнений найти решение методом обратной матрицы.

Задание. Собственные числа и собственные векторы.
Для заданной матрицы найти собственные числа и собственные векторы.

Задание. Функции многих переменных.
Для заданной функции многих переменных вычислить градиент, найти экстремумы функции.

Примеры заданий для лабораторных работ по статистическому анализу

Лабораторная работа. Предварительная обработка данных

Задание.

1. Импортировать заданный набор данных.
2. Проверить на наличие пропусков и выбросов.
3. Для количественных показателей построить гистограммы.
4. Найти оценки числовых характеристик.
5. Проверить гипотезу о нормальности.
6. Построить диаграммы размаха по группам на основании разбиения количественных показателей по уровням категориальных признаков.

Лабораторная работа. Анализ связи признаков

Выполняется в R.

Задание.

Импортировать таблицу с данными в R.

1. Построить графики для визуализации данных и их взаимосвязей.
2. Проверить связи факторов друг с другом и их влияние на зависимую целевую переменную, выбирая соответствующий критерий, в зависимости от типов данных.
3. Проверить гипотезы о значимости связи.

Лабораторная работа. Парная регрессия. Генерация.

Выполняется в R.

Задание.

1. Определить объем выборки n (от 50 до 150).
2. Сгенерировать вектор значений предсказывающей переменной.
3. Задать вектор шума, удовлетворяющий условиям Гаусса-Маркова.
4. Задать параметры регрессии.
5. Сформировать вектор значений зависимой переменной по линейной модели регрессии.
6. Построить диаграмму рассеяния и при необходимости скорректировать параметры.
7. Построить МНК-оценки параметров, проверить их значимость, сравнить с исходными значениями
8. Найти СКО остатков.
9. Проверить общую адекватность модели.

Лабораторная работа. Парная регрессия для реальных данных. Линейные и нелинейные модели.

Выполняется в R.

Задание.

Импортировать таблицу с данными в R.

1. Построить графики для визуализации данных и их взаимосвязей.
2. Проверить связь фактора с зависимой целевой переменной.
3. Построить и провести анализ парной модели регрессии целевой переменной от фактора.
4. Построить линейную, степенную, экспоненциальную, логарифмическую и обратную зависимости.
5. Оценить качество каждой модели.
6. Выбрать наиболее адекватную.

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

Для теста из 15 вопросов. За каждый вопрос в зависимости от его сложности можно получить от 1 до 3 баллов. Максимально 30.

отлично	От 26 до 30 баллов
хорошо	От 21 до 25 баллов
удовлетворительно	От 16 до 20 баллов
неудовлетворительно	От 0 до 15 баллов

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle»

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

в) Методические указания по проведению лабораторных работ.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература (на английском)

1. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) 1st ed. 2013, Corr. 7th printing 2017 Edition.
2. https://book.stat420.org/applied_statistics.pdf
3. http://thuvienso.bvu.edu.vn/bitstream/TVDHBRVT/15780/1/Applied-Statistics.pdf
4. http://wpage.unina.it/cafiero/books/stat.pdf
5. https://www.researchgate.net/publication/242692234_Statistical_foundations_of_machine_learning_the_handbook

б) дополнительная литература (на английском):

6. https://bookdown.org/ndphillips/YaRrr/
7. https://mml-book.github.io/book/mml-book.pdf

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

- Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);
- публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.)
- R The R Foundation, США свободно распространяемое.
- RStudio RStudio, PBC, США свободно распространяемое.
- JASP Амстердамский университет, Нидерланды свободно распространяемое.

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, 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.

Лаборатории, оборудованные персональными компьютерами, соответствующим необходимым программным обеспечением, выходом в интернет.

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешенном формате («Актру»).

15. Информация о разработчиках

Кабанова Татьяна Валерьевна, кандидат физ.-мат. наук, доцент, кафедра ТВиМС ИПМКН ТГУ.