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
«10» ____мая_____ 20_22_ г.

Рабочая программа дисциплины

Статистические методы машинного обучения

по направлению подготовки

02.03.02 Фундаментальная информатика и информационные технологии

Направленность (профиль) подготовки:
Искусственный интеллект и разработка программных продуктов

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2022

Код дисциплины в учебном плане: Б1.В.02.06

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-2 – Способен применять компьютерные/суперкомпьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности.

– ПК-2 – Способен проектировать базы данных, разрабатывать компоненты программных систем, обеспечивающих работу с базами данных, с помощью современных инструментальных средств и технологий.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-2.3 Использует инструментальные средства высокопроизводительных вычислений в научной и практической деятельности.

ИОПК-2.2 Использует методы высокопроизводительных вычислительных технологий, современного программного обеспечения, в том числе отечественного происхождения.

ИОПК-2.1 Обладает необходимыми знаниями основных концепций современных вычислительных систем.

ИПК-2.2 Готов осуществлять поиск, хранение, обработку и анализ информации из различных источников и баз данных, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий.

2. Задачи освоения дисциплины

– Познакомить с основными методами машинного обучения, применяемыми при анализе данных в экономике, медицине, социологии и других областях.

– Научить решать задачи статистического анализа данных с помощью моделей машинного обучения.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Дисциплина входит в модуль Искусственный интеллект.

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

Шестой 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 4 з.е., 144 часов, из которых:
-лекции: 16 ч.
-практические занятия: 32 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение в статистический анализ и первичная статистическая обработка

Задачи и этапы статистического анализа. Типы и структуры данных. Предварительная обработка данных.

Тема 2. Критерии сравнения групп.

Параметрические критерии. Непараметрические критерии.

Тема 3. Корреляционный анализ

Количественная корреляция. Ранговая корреляция. Корреляционный анализ количественных данных.

Тема 4. Регрессионный анализ

Парная регрессия. Множественная регрессия.

Тема 5. Дисперсионный анализ

Однофакторный дисперсионный анализ. Двухфакторный дисперсионный анализ

Тема 6. Задачи классификации и кластеризации.

Методы классификации. Методы кластеризации

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине осуществляется на основании проверки практических заданий, выполняемых студентами на компьютерах в течение семестра.

Студенты получают у преподавателя или собирают самостоятельно статистические данные для дальнейшего анализа и построения математических моделей. Проводят предварительную обработку данных, выбирают адекватный метод анализа в зависимости от целей исследования и типов данных, реализуют выбранные методы на компьютере, делают выводы и интерпретацию полученных результатов.

Результаты текущего контроля фиксируются в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзаменационная оценка складывается из текущего посещения (не менее 80% занятий), в срок выполненных практических заданий и результатов тестирования (при онлайн обучении) или письменного коллоквиума по темам:

1. Предварительная обработка данных. Обработка пропущенных значений и выбросов.
2. Критерии проверки нормальности.
3. Параметрические критерии сравнения выборок.
4. Непараметрические критерии сравнения выборок.
5. Общая постановка МНК-оценивания параметров линейной регрессии. Оценивание дисперсии ошибок.
6. Свойства МНК-оценок параметров линейной регрессии.
7. Обобщение оценок параметров линейной регрессии для случая коррелированных гомоскедастичных наблюдений.
8. Обобщение оценок параметров линейной регрессии для случая коррелированных гетероскедастичных наблюдений.
9. Оценки параметров линейной регрессии при связывающих эти параметры ограничениях.
10. Нелинейные модели регрессии, допускающие линеаризацию. Проверка гипотезы об адекватности модели регрессии.

11. Итерационные алгоритмы оценивания параметров регрессии.
12. Доверительные интервалы для параметров регрессии. Интервалы предсказания.
13. Коэффициенты детерминации и парной корреляции, корреляционное отношение: определения и свойства.
14. Частный и множественный коэффициенты корреляции: определения и свойства.
15. Понятие ранговой корреляции. Основные типы задач анализа ранговых связей.
16. Коэффициенты ранговой корреляции Кендалла и Спирмена. Обобщенный коэффициент ранговой корреляции.
17. Проверка гипотезы о статистически зависимой ранговой связи.
18. Коэффициент конкордации и его свойства.
19. Категоризованные данные. Анализ зависимости признаков по таблицам сопряженности.
20. Общая постановка задачи дисперсионного анализа.
21. Однофакторный дисперсионный анализ. Проверка гипотезы о влиянии фактора на исследуемый объект.
22. Исследование влияния на объект уровней фактора методами множественного сравнения.
23. Общее решение задачи двухфакторного дисперсионного анализа.
24. Двухфакторный дисперсионный анализ с равным числом r>1 наблюдений в ячейке.
25. Двухфакторный дисперсионный анализ с неравным числом наблюдений в ячейке.
26. Неполные сбалансированные блоки в задачах дисперсионного анализа.
27. Решение задачи трехфакторного дисперсионного анализа.
28. Общая постановка задачи дискриминантного анализа.
29. Решение задачи параметрического дискриминантного анализа. Расщепление смесей распределений.
30. Типы расстояний и мер близости между объектами и между классами.
31. Типы функционалов качества разбиения множества объектов на классы.
32. Основные типы кластер-процедур.

Посещение и сданные практические задания являются условием для допуска к теоретической части. Оценка за теоретическую часть ставится на основании теста или письменного коллоквиума.

Тест из 15 вопросов. Максимум 30 баллов.

0-15	Неудовлетворительно
16-20	Удовлетворительно
21-25	Хорошо
26-30	Отлично

Письменный коллоквиум. Два вопроса.

Ответ не дан или дан неверно, имеются грубые ошибки в формулировках и выводах	Неудовлетворительно
Ответ дан, но не в полном объеме, имеются существенные недочеты	Удовлетворительно
Ответ дан практически полностью, имеются некоторые незначительные ошибки	Хорошо
Ответ дан в полном объеме, допускаются очень незначительные погрешности	Отлично

При недостаточном посещении в течение семестра или невыполненных в срок работах студент может получить на экзамене дополнительные вопросы по пропущенным темам или дополнительное задание по практике.

11. Учебно-методическое обеспечение

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

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине (Приложение 1).

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Джеймс Г., Уиттон Д., Хасти Е., Тибширани Р. Введение в статистическое обучение с примерами на языке R М.: ДМК Пресс, 2016. – 450 с.

– Кабанова Т. В. Применение пакета R для решения задач прикладной статистики : учебное пособие : [для студентов и аспирантов университетов]. Томск : Издательский дом Том. гос. ун-та, 2019. – 124 с.

– Марголис Н. Ю., Кабанова Т. В. Прикладная статистика: учебно-методическое пособие. Ч. 1. Том. гос. ун-т, 2007. – 46 с.

– Марголис Н. Ю., Кабанова Т. В. Прикладная статистика: учебно-методическое пособие. Ч. 2. Том. гос. ун-т, 2007. – 58 с.

б) дополнительная литература:

– М. Кендалл, А. Стьюарт Статистические выводы и связи. Наука.: Физматлит, 1973. – 432 с.

– С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин Прикладная статистика. Классификация и снижение размерности. Финансы и статистика, 1989. – 608 с.

– Айвазян С. А, Мхитарян В. С. Прикладная статистика. Основы эконометрики: Учебник для экономических специальностей вузов: В 2 т. . Т. 1. ЮНИТИ-ДАНА, 2001. – 270 с.

– Айвазян С. А. Прикладная статистика. Основы эконометрики: Учебник для экономических специальностей вузов: В 2 т. . Т. 2, ЮНИТИ-ДАНА, 2001. – 432 с.

в) ресурсы сети Интернет:

– http://statsoft.ru/#tab-STATISTICA-link

– https://www.r-project.org/

– http://www-01.ibm.com/software/ru/analytics/spss/index.html

– http://itmu.vsuet.ru/Posobija/MathCAD/gl13/index.htm#anc1323

– http://www.exponenta.ru/

– Общероссийская Сеть КонсультантПлюс Справочная правовая система. http://www.consultant.ru

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– MS Windows,

– MS Office,

– Mathcad,

– Statistica,

– R, R Studio.

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

– ЭБС Консультант студента – http://www.studentlibrary.ru/

– Образовательная платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения практических занятий, 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

Кабанова Татьяна Валерьевна, канд. физ.-мат. наук, доцент, доцент кафедры теории вероятностей и математической статистики ИПМКН ТГУ.