

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 - II

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки :

Big Data and Data Science

Форма обучения

Очная

Квалификация

Магистр

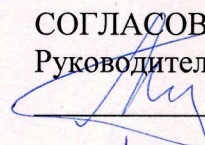
Год приема

2022

Код дисциплины в учебном плане: Б1.П.В.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 части образовательной программы, формируемой участниками образовательных отношений .

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

Второй семестр, зачет

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 и курса «Статистические методы машинного обучения – I».

6. Язык реализации

Английский.

7. Объем дисциплины

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

-лекции: 10 ч.

-лабораторные: 20 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Множественная регрессия.

Основные понятия и задачи регрессионного анализа, Общая постановка задачи множественной регрессии. Метод наименьших квадратов оценки параметров регрессии. Теорема Гаусса-Маркова. Оценка дисперсий. Проверка качества модели множественной регрессии. Нелинейные модели и линеаризация.

Тема 2. Дополнительные вопросы регрессионного анализа.

Фиктивные переменные. Случай смещенного шума. Случай коррелированных наблюдений Гетероскедастичность. Мультиколлинеарность.

Тема 3. Задачи классификации.

Основные понятия и задачи классификации. Бинарная классификация и логистическая регрессия. Метрики качества. ROC-анализ.

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

Зачет во втором семестре проводится в форме итогового тестирования. Тест состоит из 10-15 вопросов. Продолжительность экзамена 30 минут.

Примерный перечень теоретических вопросов и тем для подготовки к экзамену:

1. Нелинейные модели и линеаризация.
2. Случай смещенного шума.
3. Случай коррелированных гомоскедастичных наблюдений.
4. Случай некоррелированных гетероскедастичных наблюдений.
5. Мультиколлинеарность.
6. Фиктивные переменные.
7. Постановка задачи классификации.
8. Логистическая регрессия.
9. Метрики качества бинарного классификатора.
10. ROC-анализ.

Примеры заданий для лабораторных работ

Лабораторная работа. Множественная регрессия. Фиктивные переменные

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.
3. Построить и провести анализ множественной модели регрессии целевой переменной от всех представленных количественных и порядковых факторов.
4. Провести обработку и кодирование категориальных факторов.
5. Построить и провести анализ множественной модели регрессии с учетом всех предложенных факторов.
6. Удалить незначимые факторы. Построить окончательную модель.
7. Проверить остатки модели на нормальность.
8. Задать новое наблюдение со своими значениями признаков и построить прогноз целевого показателя для него.

Лабораторная работа. Логистическая регрессия.
Задание.

Сформировать наблюдения, связанные однофакторной логистической регрессией.

1. Задать объем выборки $n = 20 : 50$.
2. Значения фактора x сформировать как реализацию целочисленной равномерно распределенной случайной величины в интервале $[a, b]$.
3. Задать нормально распределенный шум $\varepsilon \sim N(0, \sigma)$.
4. Определить регрессионную модель

$$P(x) = \frac{e^{\theta_0 + \theta_1 x + \varepsilon}}{1 + e^{\theta_0 + \theta_1 x + \varepsilon}}$$

5. Значение бинарной зависимой переменной определить как

$$y_i = \begin{cases} 0, & P(x_i) < \frac{1}{2}; \\ 1, & P(x_i) \geq \frac{1}{2}. \end{cases}$$

Все параметры задать самостоятельно, ориентируясь на диаграмму рассеяния.

6. Оценить параметры модели.
7. Проверить общее качество модели.

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

Для теста из 10 вопросов. За каждый вопрос в зависимости от его сложности можно получить от 1 до 3 баллов. Максимально 20.

отлично	От 18 до 20 баллов
хорошо	От 14 до 17 баллов
удовлетворительно	От 11 до 13 баллов
неудовлетворительно	От 0 до 10 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/cafiere/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 OneNote, 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.