

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Научно-образовательный центр «Высшая ИТ школа»

УТВЕРЖДЕНО:
Исполнительный директор НОЦ ВИТШ

Т.С.Кетова

Рабочая программа дисциплины

Прикладная статистика
(ПрикСтат)

по направлению подготовки
09.03.04 (33.04) Программная инженерия

Направленность подготовки:
«Программная инженерия»

Форма обучения
Очная

Квалификация
Программный инженер

Год приема
2023

СОГЛАСОВАНО:
Руководитель ОП
О.А.Змеев

Председатель УМК
Д.О. Змеев

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующих компетенций и результатов обучения:

БК-5	Способен использовать прикладную вычислительную математическую теорию в форме построения логической цепочки математических суждений для разделения вычислительных и аналитических задач на более мелкие подзадачи с последующих их решением и обобщения результатов на общую задачу	<p>Знает: Основы дискретной математики, основы вычислительной математики, основы теории чисел, основы теории доказательств, основы линейной алгебры и геометрии</p> <p>Умеет: Использовать построение логической цепочки суждений для построения доказательств математических, или сводимых к математическим задач; использовать разные вычислительные методы и приёмы; объяснять собственные математические выкладки заинтересованным сторонам; находить ошибки в логике доказательств математических задач</p>
ПК 2.1	Способен к использованию и имплементации наукоемких алгоритмов и вычислительных компонентов для решения практико-ориентированных задач, в рамках разработки программного обеспечения	<p>Знать: Основные виды и правила использования вычислительных алгоритмов, основанных на машинном обучении</p> <p>Правила интеграции наукоёмкого модуля в общую структуру модулей программного обеспечения</p> <p>Уметь: Имплементировать основные виды наукоемких алгоритмов и вычислительных компонентов систем</p>
ПК-2.2	Способен трансформировать вычислительную задачу из бизнес-постановки в формальное определение алгоритмической задачи, и обратно преобразовать полученные результаты в бизнес-термины	<p>Знает: Правила логического и вывода и преобразования информации описывающих объекты и правила реального мира в формальные определения для алгоритмов</p> <p>Умеет: Преобразовать формулировку задачи с уровня бизнес-терминов в уровень формальный алгоритмический</p> <p>Интерпретировать полученные результаты вычислительного алгоритма с точки зрения бизнес-терминов</p>

2. Задачи освоения дисциплины

- Научиться 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 использования и применения специализированных библиотек языков программирования и приложений для статистической обработки данных.

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплины (модули)» части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Для внесения оценок в зачетные книжки обучающихся принимается сокращенное название дисциплины «ПрикСтат».

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине Семестр 5, Экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются знания основ математического анализа, линейной алгебры, методов оптимизаций, теории вероятностей и математической статистики, а также основ программирования.

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 4 з.е., 144 часов, из которых:

– лекции: 30.0 ч.;

– практические занятия: 30.0 ч.;

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), 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 данных. Типы данных. Графические и табличные способы представления данных. Предварительная обработка данных. Оценки параметров и числовых характеристик. Проверка статистических гипотез.

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. Анализ таблиц сопряженности. Коэффициент Крамера.

Тема 5. Регрессионный анализ. Постановка задачи регрессионного анализа. 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. Фиктивные переменные. Случай смещенного шума. Случай коррелированных наблюдений Гетероскедастичность. Мультиколлинеарность.

Тема 6. 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

Экзамен во втором семестре проводится в форме итогового тестирования. Тест состоит из 10-15 вопросов. Продолжительность экзамена 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.
16. Нелинейные модели и линеаризация.
17. Случай смещенного шума.
18. Случай коррелированных гомоскедастичных наблюдений.
19. Случай некоррелированных гетероскедастичных наблюдений.
20. Мультиколлинеарность.
21. Фиктивные переменные.
22. Постановка задачи классификации.
23. Логистическая регрессия.
24. Метрики качества бинарного классификатора.
25. ROC-анализ.

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» - <https://moodle.tsu.ru/course/view.php?id=00000>

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:

1. Кабанова Т.В. Применение пакета R для решения задач прикладной статистики: учебное пособие: [для студентов и аспирантов университетов]. – Томск: Издательский Дом Томского государственного университета. 2019. 124 с.

2. Мьльников Л.А. Статистические методы интеллектуального анализа данных. – СПб.: БХВ-Петербург, 2021. – 240 с.

3. Дж. Д. Лонг, Пол Титор. R. Книга рецептов: Проверенные рецепты для статистики, анализа и визуализации данных / пер. с англ. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 510 с.

б) дополнительная литература:

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

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

3. 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.

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

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

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

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

в) ресурсы сети Интернет:

– Статистика в Data Science – исчерпывающий гид для амбициозных практиков ML <https://habr.com/ru/company/skillfactory/blog/526972/>

– Введение в Data Science и машинное обучение <https://stepik.org/course/4852>

– 10 примеров использования статистических методов в проекте машинного обучения <https://www.machinelearningmastery.ru/statistical-methods-in-an-applied-machine-learning-project/>

– Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных www.machinelearning.ru/

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.

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/>
- 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 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 смешанном формате («Актру»).

15. Информация о разработчиках

Кабанова Татьяна Валерьевна,
кандидат физико-математических наук, доцент,
доцент кафедры теории вероятностей и математической статистики

Ким Константин Станиславович,
кандидат физико-математических наук,
доцент НОЦ «Высшая ИТ школа»