МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Тульский государственный университет»

Институт прикладной математики и компьютерных наук
Кафедра «Прикладная математика и информатика»

Утверждено на заседании кафедры
«Прикладная математика и информатика»
24 января 2022 г., протокол № 5

Заведующий кафедрой

_____М.В. Грязев

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ) «Инструменты прикладной статистики»

основной профессиональной образовательной программы высшего образования – программы магистратуры

по направлению подготовки
01.04.02 Прикладная математика и информатика

с направленностью (профилем)
Перспективные методы искусственного интеллекта в сетях передачи и обработки данных

Форма обучения: очная

Идентификационный номер образовательной программы: 010402-01-22

Тула 2022 год

1. Место дисциплины (модуля) в структуре ОПОП ВО:
Дисциплина (модуль) относится к части дисциплин основной профессиональной образовательной программы, формируемых участниками образовательных отношений

2. Входные требования для освоения дисциплины (модуля), предварительные условия (если есть):
Учащиеся должны владеть знаниями о принципах работы традиционных компьютерных сетей, программно-конфигурируемых компьютерных сетей в объеме, соответствующем основным образовательным программам бакалавриата по укрупненным группам направлений и специальностей 01.00.00 «Математика и механика», 02.00.00 «Компьютерные и информационные науки»

3. Результаты обучения по дисциплине (модулю):

Планируемые результаты обучения по дисциплине (модулю)	
Формируемые компетенции (*код и наименование компетенции*)	**Результаты обучения** ***(знания, умения)***
ПК-8. Способен разрабатывать и модернизировать программное и аппаратное обеспечение технологий и систем искусственного интеллекта с учетом требований информационной безопасности в различных предметных областях.	ПК-8.1. Разрабатывает программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях.
	ПК-8.2. Модернизирует программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях.

4. Объем дисциплины (модуля) составляет 3 з.е., в том числе 24 академических часа контактная работа с преподавателем - 12 академических часа занятий лекционного типа, 12 академических часов занятий практического типа, 82 академических часов на самостоятельную работу обучающихся.

5. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий:
5.1. Структура дисциплины (модуля) по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий (в строгом соответствии с учебным планом)

Наименование разделов и тем дисциплины (модуля), **Форма про-**	**Номинальные трудозатраты обучающегося**		**Всего академических часов**	**Форма текущего контроля успеваемости* (наименование)**
	Контактная работа (работа во взаимодействии с преподавателем) Виды контактной работы, академические часы	**Самостоятельная работа обучающегося, академиче-**		

межуточной аттестации по дисциплине (модулю)	Занятия лекционного типа	Практические занятия	**ские часы**		
Тема 1. Цели и задачи анализа данных; Теория вероятностей и статистика как формализмы	2	2	14	18	контрольная работа
Тема 2. Типологизация задач восстановления плотности; Типологизация проверки гипотез	2	2	14	18	контрольная работа
Тема 3. Множественная проверка гипотез	2	2	14	18	контрольная работа
Тема 4. Анализ зависимостей	2	2	14	18	контрольная работа
Тема 5. Линейная регрессия, обобщения регрессии	2	2	14	18	контрольная работа
Тема 6. Анализ временных рядов. Основы теории измерений	2	2	14	18	контрольная работа
Другие виды самостоятельной работы (отсут-	—	—			—

ствуют)					
Промежуточная аттестация (зачет)					
Итого	*12*	*12*	*84*	108	—

5.2. Содержание разделов (тем) дисциплины

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
1.	Тема 1. Цели и задачи анализа данных; Теория вероятностей и статистика как формализмы	Роль теории вероятностей и статистики в анализе данных. Понятие об инструментах прикладной статистики и фундаментальных задачах интеллектуального анализа данных. Классификация инструментов статистики и фундаментальных задач интеллектуального анализа данных. Базовые законы теории вероятностей. Способы задания распределений. Числовые характеристики распределений. Основные распределения. Основные интерпретации вероятности.
2.	Тема 2. Типологизация задач восстановления плотности; Типологизация проверки гипотез	Задачи точечного оценивания. Задачи интервального оценивания. Псевдовыборки. Проверка без альтернативы, проверка параметрических гипотез. Проверка непараметрических гипотез. Проверка с альтернативой, ROC-анализ описаний объектов и стратегий распознавания. Стратегии распознавания при наличии механизма смешивания классов.
3.	Тема 3. Множественная проверка гипотез	Множественная проверка гипотез
4.	Тема 4. Анализ зависимостей	Дисперсионный анализ, Корреляционный анализ, Перестановочные тесты
5.	Тема 5. Линейная регрессия, обобщения регрессии	Виды системной информации, вещаемой в соте. Классификация сот. Процедуры выбора сети. Состояния абонентского терминала, процедуры, выполняемые терминалом в этих состояниях.
6.	Тема 6. Анализ временных рядов. Основы теории измерений	Анализ выживаемости, цензурированные данные. Основы теории тестов, валидация шкал.

6. Фонд оценочных средств (ФОС, оценочные и методические материалы) для оценивания результатов обучения по дисциплине (модулю).

6.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости, критерии и шкалы оценивания (в отсутствие утвержденных соответствующих локальных нормативных актов на факультете)

Вопросы для контрольных работ
Контрольная работа № 1 (теория вероятностей)
Задача 1. Для стандартного нормального распределения выразить p-value через функцию распределения. Т. е. надо вывести формулу, в которой одна функция выражается через дру-

гую. Построить график p-value. Обязательно надо надписать оси и отметить характерные значения.
Задача 2. Известно, что у мадагаскарских варанов длина имеет нормальное распределение с матожиданием 120 см и дисперсией 100 см2.

1. Какова вероятность, что длина мадагаскарского варана будет более 125 см?
2. Какова вероятность, что в группе из 25 мадагаскарских варанов средняя длина будет более 125 см?
3. Какова вероятность, что в двух группах каждая из 25 мадагаскарских варанов средние длины будут отличаться более чем на 5 см?

Контрольная работа № 2 (поиск критерия и проверка гипотезы)
Задача 1. Известно, что у мадагаскарских варанов длина имеет нормальное распределение с матожиданием 120 см и дисперсией 100 см2.У британских учёных возникло подозрение, что вараны бывают не только мадагаскарскими. Найдите критерий с уровнем значимости 5%, определяющий по длине варана его принадлежность к мадагаскарским варанам. Дайте явную формализацию этой задачи.
Задача 2.Ученые поехали на Мадагаскар изучать варанов. Британские ученые работали на севере Мадагаскара, французские – на юге. Каждые отловили и измерили по 100 варанов. Результаты измерений длины варанов записаны в файлы north.txt и south.txt. На уровне значимости 5% проверить, что южная и северная популяции варанов имеют одно и то же распределение длины. Дайте хотя бы две различные формализации этой задачи. Отметим, что о виде распределения ничего не говорилось.

Контрольная работа № 3 (множественная проверка, анализ зависимостей)
Задача 1. Ученые поехали на Мадагаскар изучать варанов. Измерили 100 варанов. Результаты измерений длины варанов записаны в файл length.txt. Пусть известно, что у варанов длина имеет нормальное распределение с матожиданием 120 см и дисперсией 100 см2.На уровне значимости 5% проверить, что ученым попадались только вараны. Если все 100 животных – вараны, это хорошо. Если хоть одно из 100 животных – не варан, это плохо. Дайте явную формализацию этой задачи.
Задача 2. 72 пациента проходили лечение от гипертонии. Для лечения использовались три вида лекарств, при этом их эффект изучался как при использовании специальной диеты, так и без диеты. Кроме того, в половине случаев применялась психотерапия. Изучаемая переменная — артериальное давление пациента по окончании лечения. Данные находятся в файле hypertension.txt. Требуется сравнить эффективность методов лечения гипертонии разными способами.

1. Нарисуйте и проинтерпретируйте диаграммы взаимодействия
2. Проведите трехфакторный дисперсионный анализ, используя все взаимодействия. Что можно сказать о значимости тройного взаимодействия?
3. Для пациентов, проходящих психотерапию, проведите двухфакторный дисперсионный анализ с целью выяснения значимых факторов, которые влияют на давление человека.

Реферат
Реферат пишется всеми студентами согласно одной и той же структуре. Студенты самостоятельно выбирают набор данных. Данные должны быть структурированными (классическая модель отношения) и содержать признаки всех изученных типов: категориальные, упорядоченные, арифметические.
Структура реферата (этапы статистического исследования)

1. Содержательная задача

2. Структура данных
3. Формальная задача
4. Разбиение выборки
5. Дескриптивная статистика
6. Анализ до распознавания
7. Параметрический подход
8. Непараметрический подход
9. Сравнение подходов и выводы

6.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации по дисциплине, критерии и шкалы оценивания

Вопросы к экзамену

1. Цели и задачи анализа данных в разных видах деятельности. Методология анализа данных как науки и индустрии. Связь анализа данных с фундаментальной и прикладной математикой.
2. Трехуровневая классификация аналитических задач и технологий. Уровень сбора и хранения информации. Уровень запросов к данным, описания данных и проверки гипотез. Уровень генерации новых гипотез и выявления закономерностей.
3. Понятие об инструментах прикладной статистики и фундаментальных задачах статистического распознавания.
4. Основные модели в анализе данных. Понятие модели данных. Понятие информационной модели.
5. Классификация инструментов статистики и фундаментальных задач интеллектуального анализа данных: по наличию целевых признаков, по типу признаков, по существованию распределения, по модели исходных данных.
6. Теория вероятностей и статистика как формализмы. Понятие эксперимента в теории вероятностей. Основные задачи теории вероятностей. Понятие о параметрических и полупараметрических моделях эксперимента.
7. Понятие случайной величины. Совместные распределения, маргинальные распределения, условная вероятность, теорема произведения, формула Байеса
8. Виды распределений. Способы задания распределений, функция распределения и ее свойства, плотность распределения.
9. Дополнительные способы задания распределений: квантили, p-value.
10. Числовые характеристики распределений: матожидание, дисперсия, моменты.
11. Основные дискретные и непрерывные распределения.
12. Интерпретация вероятности (классическая, геометрическая, частотная, субъективная)
13. Задача точечного оценивания
14. Статистики как функции выборки: вариационный ряд, выборочное среднее, выборочная дисперсия, выборочная медиана
15. Свойства точечных оценок: состоятельность, несмещенность, эффективность, робастность
16. Принципы статистики: принцип максимального правдоподобия, принцип максимальной апостериорной вероятности, принцип максимальной обоснованности. Сравнение разных принципов на одной и той же задаче
17. Метод максимального правдоподобия как метод получения точечных оценок

18. Метод наименьших квадратов, его связь с методом максимального правдоподобия. Регуляризация при настройке линейных моделей регрессии: ridge, lasso, elasticnet.
19. Свойства точечных оценок в западной культуре: accuracy, precision, trueness
20. Задача разделения смеси распределений. Идентифицируемые распределения. EM-алгоритм.
21. Задача интервального оценивания, доверительный интервал, уровень надежности
22. Методы построения распределения точечной оценки (параметрический, наивный, бутстреп), построение доверительных интервалов по распределению точечной оценки
23. Непараметрическое восстановление распределений, метрические методы, ядровое сглаживание
24. Понятие классов и их традиционные наименования в статистике. Типы задач в проверке гипотез.
25. Фишеровская задача распознавания, нулевая гипотеза, функция правдоподобия, p-значение и его использование, ошибки первого рода и специфичность, уровень значимости.
26. Неймановская задача распознавания, альтернативная гипотеза, отношение правдоподобия, ошибки второго рода и чувствительность, мощность критерия. Минимаксная задача распознавания, равный уровень ошибок.
27. Совместное распределение, априорная вероятность, апостериорная вероятность. Функция потерь, средний риск, байесовская задача распознавания.
28. Матрица ошибок.Основные показатели качества в задачах классификации и восстановления регрессии, доступные показатели качества в разных типах задач. Парадоксы их использования (проблема группирования, проблема редких событий), проблемы теории рационального выбора.
29. Важнейшие функции потерь, соответствующие байесовские стратегии.
30. Средние и эмпирические показатели качества
31. Эмпирический риск, обобщающая способность стратегии. Явления недообучения и переобучения. Роль обучающей, валидационной и контрольной выборок при обучении по прецедентам. Кросс-валидация.
32. Разложение среднего риска на части, дилемма смещения-дисперсии, теоретическое обоснование ансамблей классификаторов.
33. Проверка параметрических гипотез и проверка непараметрических гипотез
34. Многоэтапная диагностика
35. Метрические модели в распознавании. Парзеновские окна. Метод k ближайших соседей.
36. Множественная проверка гипотез
37. Последовательный анализ
38. Дисперсионный анализ как инструмент статистики
39. Корреляционный анализ как инструмент статистики
40. Регрессионный анализ как инструмент статистики
41. Меры качества регрессионных моделей как инструмент статистики
42. ROC-анализ описаний объектов, индекс Джини (сравнение распределения признака между классами без согласования их размера)
43. ROC-анализ классификаторов, ROC-AUC и его свойства.
44. Отбор и генерация признаков на основе операционных характеристик признаков (информативности)
45. Перестановочные тесты

46. Обобщенные линейные модели. Логистическая регрессия. Переход от линейных моделей к нелинейным при помощи ядровой функции.
47. Анализ временных рядов
48. Анализ выживаемости как задача на цензурированных данных. Основы модели Кокса.
49. Теория измерений (метрологии), характеристики средства измерений, выпуклые комбинации предикторов.
50. Теория тестов, валидация шкал.

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	2 (не зачтено)	3 (зачтено)	4 (зачтено)	5 (зачтено)
Знания *(виды оценочных средств: опрос, тесты)*	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения *(виды оценочных средств: практические задания)*	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки (владения, опыт деятельности) *(виды оценочных средств: выполнение и защита курсовой работы, отчет по практике, отчет по НИР и т.п.)*	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач

7. Ресурсное обеспечение:

7.1. Перечень основной и дополнительной литературы

Основная литература

1. Hlavac V. Ten lectures on statistical and structural pattern recognition. – Springer Science; Business Media, 2013)

Дополнительная литература

1. Лагутин, М.Б. Наглядная математическая статистика. — М.: П-центр, 2003.
2. Кобзарь, А.И. Прикладная математическая статистика. — М.: Физматлит, 2006.
3. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004 (Schlesinger M. I.,
4. Вальд, А. Последовательный анализ. — М.: Физматлит, 1960.
5. Bishop C. M. Pattern recognition and machine learning. – Springer, 2006.
6. Max Kuhn, Kjell Johnson. Applied Predictive Modeling. — Springer, 2013.

7. Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — 2nd ed. — Springer-Verlag, 2009. — 746 p. — ISBN 978-0-387-84857-0
8. Tabachnick, B.G., Fidell, L.S. Using Multivariate Statistics. — Boston: Pearson Education, 2012.
9. Bonnini, S., Corain, L., Marozzi, M., Salmaso S. Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R. — Hoboken: John Wiley & Sons, 2014.
10. Agresti, A. Categorical Data Analysis. — Hoboken: John Wiley & Sons, 2013.
11. Bilder, C.R., Loughin, T.M. Analysis of Categorical Data with R. — Boca Raton: Chapman and Hall/CRC, 2013.
12. Cameron, A.A., Trivedi, P.K. Regression Analysis of Count Data. — Cambridge: Cambridge University Press, 2013.
13. Bretz, F., Hothorn, T., Westfall, P. Multiple Comparisons Using R. — Boca Raton: Chapman and Hall/CRC, 2010.
14. Chihara, L., Hesterberg, T. Mathematical Statistics with Resampling and R — Hoboken: John Wiley & Sons, 2011.
15. Kanji, G.K. 100 statistical tests. — London: SAGE Publications, 2006.
16. Mukhopadhyay, N., de Silva, B. M. Sequential methods and their applications. — Boca Raton: Chapman and Hall/CRC, 2009.
17. Olsson, U. Generalized Linear Models: An Applied Approach. — Lund: Studentlitteratur, 2004.
18. Pearl J., Glymour M., Jewell N.P. Causal Inference in Statistics: A Primer. — Chichester: John Wiley & Sons, 2016.
19. Wooldridge, J. Introductory Econometrics: A Modern Approach. — Mason: South-Western Cengage Learning, 2013.

7.2. Перечень лицензионного программного обеспечения, в том числе отечественного производства

При реализации дисциплины может быть использовано следующее программное обеспечение:

1. Операционная система ALT Linux MATE Starter kit 9 лицензияGPL
2. Программный продукт Python 3.5.1 (64-bit)Python Software Foundation
3. Операционная система Microsoft Windows 10 Education академическая лицензия

7.3. Перечень профессиональных баз данных и информационных справочных систем

1. http://www.edu.ru – портал Министерства образования и науки РФ
2. http://www.mon.gov.ru - Министерство образования и науки Российской Федерации
3. http://www.fasi.gov.ru - Федеральное агентство по науке и инновациям

7.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. www.machinelearning.ru
2. www.kaggle.com
3. archive.ics.uci.edu/ml/index.php

7.5. Описание материально-технического обеспечения.

Образовательная организация, ответственная за реализацию данной Программы, располагает соответствующей материально-технической базой, включая современную вычислительную технику, объединенную в локальную вычислительную сеть, имеющую выход в Интернет. Используются специализированные компьютерные классы, оснащенные современным оборудованием. Материальная база соответствует действующим санитарно-техническим нормам и обеспечивает проведение всех видов занятий (лабораторной, практической, дисциплинар-

ной и междисциплинарной подготовки) и научно-исследовательской работы обучающихся, предусмотренных учебным планом.

8. Соответствие результатов обучения по данному элементу ОПОП результатам освоения ОПОП указано в Общей характеристике ОПОП.

9. Рабочая программа внедрена в соответствии с Соглашением о предоставлении из федерального бюджета грантов в форме субсидий на разработку программ бакалавриата и программ магистратуры по профилю «искусственный интеллект, а также на повышение квалификации педагогических работников образовательных организаций высшего образования в сфере искусственного интеллекта, заключённым «29» сентября 2021 г. № 075-15-2021-1036 между МГУ имени М.В.Ломоносова и Минобрнауки России.