Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КАЛИНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

А. Б. Тристанов

ПРИКЛАДНАЯ СТАТИСТИКА И АНАЛИЗ ДАННЫХ

Учебно-методическое пособие по изучению дисциплины
для студентов магистратуры по направлению
09.04.01 Информатика и вычислительная техника

Калининград,
Издательство ФГБОУ ВО «КГТУ»
2023

Рецензент:
кандидат физико-математических наук, и.о. заведующего кафедрой
прикладной математики и информационных технологий
А.И. Руденко

Тристанов, А. Б.
Прикладная статистика и анализ данных: учебно-методическое пособие по изучению дисциплины для студентов магистратуры по направлению 09.04.01 Информатика и вычислительная техника / А. Б. Тристанов. – Калининград: Изд-во ФГБОУ ВО «КГТУ», 2023. – 26 с.

Учебно-методическое пособие является руководством по изучению дисциплины «Прикладная статистика и анализ данных» для студентов магистратуры по направлению подготовки 09.04.01 Информатика и вычислительная техника и содержит характеристику дисциплины (цель и планируемые результаты изучения дисциплины, место дисциплины в структуре основной профессиональной образовательной программы), тематический план с описанием для каждой темы форм проведения занятия, вопросов для изучения, методических материалов к занятию.
Рис. 1, табл. 1, список лит. – 6 наименований

Учебно-методическое пособие рекомендовано к использованию в качестве локального электронного методического материала в учебном процессе методической комиссией ИЦТ 5 июля 2023 г., протокол № 8

ОГЛАВЛЕНИЕ

1. ВВЕДЕНИЕ

Данное учебно-методическое пособие предназначено для студентов магистратуры по направлению 09.04.01 Информатика и вычислительная техника, изучающих дисциплину «Прикладная статистика и анализ данных».

Цель освоения дисциплины: ознакомление студентов с теоретическими основами статистики и с основными областями применения статистических методов; формирование у студентов практических навыков применения статистического анализа в прикладных задачах; овладение инструментальными средствами, моделями и методами интеллектуального анализа данных в задачах поиска информации, обработки и анализа данных, а также приобретения навыков исследователя данных (data scientist).

В результате освоения дисциплины ожидается, что студенты получат целостное представление о месте технологий обработки и анализа больших данных в профессиональной области, а также освоят ряд инструментальных средств анализа данных.

Далее в пособии представлен тематический план, содержащий перечень изучаемых тем, обязательных лабораторных/практических работ, мероприятий текущей аттестации. При формировании личного образовательного плана на семестр следует оценивать рекомендуемое время на изучение дисциплины; возможно, вам потребуется больше времени на выполнение отдельных заданий или проработку отдельных тем.

В разделе «Содержание дисциплины» приведены подробные сведения об изучаемых вопросах, по которым вы можете ориентироваться в случае пропуска каких-либо занятий, а также методические рекомендации преподавателя для самостоятельной подготовки. Каждая тема имеет ссылки на литературу (или иные информационные ресурсы), а также контрольные вопросы для самопроверки.

Раздел «Текущая аттестация» содержит описание обязательных мероприятий контроля самостоятельной работы и усвоения разделов или отдельных тем дисциплины. Далее изложены требования к завершающей аттестации – зачету и курсовой работе.

Помимо данного пособия, студентам следует использовать материалы, размещенные в соответствующем данной дисциплине разделе ЭИОС, в которые более оперативно вносятся изменения для адаптации дисциплины под конкретную группу.

В ходе изучения дисциплины, выполнения лабораторных работ и расчетно-графической работы используются: аналитическая платформа Loginom, бесплатная академическая версия которой может быть свободно загружена с сайта компании-разработчика ООО «Аналитические технологии» – https://loginom.ru/download и установлена на персональные компьютеры, отвечающие техническим требованиям, и дистрибутив Python

Ananconda (https://www.anaconda.com/download), включающий пакет Jupyter Notebook или JupyterLab.

2. ТЕМАТИЧЕСКИЙ ПЛАН

	Раздел (модуль) дисциплины	**Тема**
		1 семестр
Теоретическое обучение (лекции)		
1.1	Теория вероятностей и математическая статистика	Основные понятия теории вероятностей
1.2		Основные теоремы
1.3		Случайные величины
1.4		Предельные теоремы
1.5		Основные понятия математической статистики
1.6		Статистическое оценивание параметров распределения
1.7		Проверка статистических гипотез
1.8		Разведочный статистический анализ и визуализация данных
		2 семестр
2.1	Классификация и снижение размерности	Задача классификации
		Классификации с учителем
		Классификации без учителя
		Снижение размерности признакового пространства и отбор наиболее информативных признаков
3.1	Исследование зависимостей	Корреляционный, регрессионный и дисперсионный анализ
Практические (лабораторные занятия)		
1.1	Теория вероятностей и математическая статистика	Основы работы в JupyterLab. Основные статистические библиотеки Python
1.2		Исследование точечных оценок
1.3		Исследование интервальных оценок
1.4		Проверка гипотез о параметрах и виде распределений
1.5		Библиотеки визуализации данных
2.1	Классификация и снижение размерности	Решение задачи классификации с использованием библиотеки Sklearn
2.2		Решение задачи кластеризации ядерными методами с использованием библиотеки Sklearn
2.3		Исследование иерархической кластеризации с использованием библиотеки Sklearn
2.4		Метод главных компонент
3.1	Исследование зависимостей	Регрессионный анализ

3. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ И УКАЗАНИЯ К ИЗУЧЕНИЮ

3.1.РАЗДЕЛ 1. ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

3.1.1. Тема 1.1 Основные положения теории вероятностей

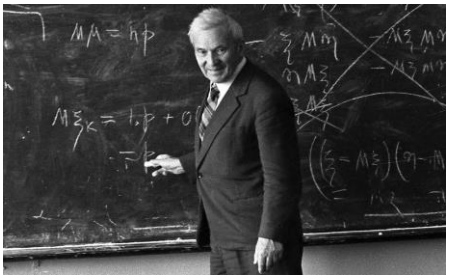
Перечень изучаемых вопросов:

Элементарные сведения из теории множеств (понятие множества, объединение, пересечение, дополнение множеств). Случайное событие. Алгебра событий. Совместные и несовместные, зависимые и независимые события. Полная группа событий. Аксиоматика Колмогорова[1]. Вероятностное пространство. Опыт с конечным числом исходов. Классическое определение вероятности. Непосредственный подсчет вероятностей. Схема выбора с возвращением и без возвращения элементов. Частота или статистическая вероятность события. Геометрическая вероятность. Аксиомы теории вероятностей и их следствия. Теорема сложения и умножения вероятностей. Условная вероятность события. Независимость событий. Формула полной вероятности. Теорема Байеса (формула Байеса). Независимые испытания. Формула Бернулли. Наивероятнейшее число успехов в схеме Бернулли. Локальная и интегральная предельные теоремы. Теорема Пуассона.

Методические указания к изучению:

Данная тема дается в относительно кратком изложении, поскольку предполагается, что Вы освоили курс математической статистики и теории вероятностей в программе бакалавриата либо в ходе подготовки к обучению в магистратуре. Тем не менее основные положения данных разделов математики следует углубить и расширить.

Теория вероятностей, как и многие другие математические дисциплины, в своей основе имеет теорию множеств, в связи с этим вам



1 Колмогоров Андрей Николаевич (12 (25) апреля 1903, Тамбов — 20 октября 1987, Москва) — выдающийся советский математик, один из крупнейших математиков XX века. Один из основоположников современной теории вероятностей. Им получены фундаментальные результаты в топологии, геометрии, математической логике, классической механике, теории турбулентности, теории сложности алгоритмов, теории информации, теории функций, теории тригонометрических рядов, теории меры, теории приближения функций, теории множеств, теории дифференциальных уравнений, теории динамических систем, функциональном анализе и в ряде других областей математики и её приложений. Автор новаторских работ по философии, истории, методологии и преподаванию математики.

следует повторить соответствующие разделы математики, которые изучались ранее.

Основополагающим понятием теории вероятностей является понятие случайного события. Дайте определения элементарному событию (исходу), достоверному и невозможному событию, множеству (пространству) элементарных исходов. Следует уяснить, что любое случайное событие есть подмножество множества элементарных исходов. Рассмотрите простые примеры с бросанием игрального кубика и извлечением карт из колоды. Рассмотрите понятие совместных и несовместных событий. Наглядным является иллюстрация событий с помощью кругов Эйлера. Дайте определение суммы и произведения событий.

Современная теория вероятностей оперирует понятием «вероятностное пространство». Для построения этого понятия потребуется разобраться с более сложным математическим объектом — σ-алгебра.

σ-алгебра — это совокупность некоторого множества и операций над элементами этого множества, такая, что операции замкнуты относительно этого множества. Конкретизируем это понятие для теории вероятностей.

Дадим определение: σ-алгебра событий Σ — это совокупность множества всех подмножеств пространства элементарных исходов и заданных операций сложения и умножения событий. Покажите, что достоверное Ω и невозможное события содержаться в Σ.

Бытовым пониманием вероятности события является степень уверенности в том, что событие произойдет или не произойдет. Более строго вероятность определяется через систему аксиом как мера на вероятностном пространстве.

Аксиоматика Колмогорова. Вероятностью (вероятностной мерой) называют числовую функцию, заданную на σ-алгебра событий Σ, такую что:

1. $P(A) \geqslant 1$
2. $P(\Omega) = 1$
3. Для любых несовместных событий A и B справедливо $P(A + B) = P(A) + P(B)$

Вероятностным пространством называют совокупность пространства элементарных исходов, алгебры событий и вероятностной меры - (Ω, Σ, P).

Далее следует рассмотреть частные случаи задания вероятностных мер — классическую, геометрическую и статистическую вероятности. Обязательно приведите примеры типовых задач. Уясните общность задач для одних и тех же вероятностных пространств. Покажите, что, например, задачи «на классическую вероятность» эквивалентны задачам про извлечение цветных шаров из урны.

Завершая изучение этого раздела темы, следует убедиться, что вы можете давать все указанные выше определения и приводить иллюстративные примеры «из жизни». В качестве предмета можно выбирать бросание игральных кубиков, извлечение карт из колоды или шаров из урны.

После того как мы изучили базовые понятия, аксиомы и свойства, можно переходить к формулированию основных теорем, которые из них следуют. Докажите теоремы о сложении и умножении вероятностей.

Дайте определение зависимым и независимым событиям, условной вероятности. Покажите, что условная вероятность удовлетворяет аксиомам Колмогорова. Как формулируется теорема о произведение вероятностей для зависимых событий?

Уделите достаточно времени для разбора формулы полной вероятности и связанной с ней формулы Байеса.

В продолжение темы разберите доказательства и примеры применения схемы Бернулли, локальной и интегральной предельных теорем Муавра - Лапласа, теоремы Пуассона.

Литература:
[1] гл.1, [3] гл. 1-2 [1] гл. 1-5, [3] гл. 3.

Контрольные вопросы:

1. Что называют суммой случайных событий и произведением случайных событий?
2. Как интерпретировать дополнение случайного события до Ω?
3. Покажите, что для случайных событий справедливы законы де Моргана.
4. Что означает замкнутость множества относительно некоторой операции?
5. Проиллюстрируйте с помощью кругов Эйлера совместные и несовместные события.
6. Что такое классическая вероятность? Покажите, что для классической вероятности справедливы аксиомы Колмогорова.
7. В чем отличия классической и геометрической вероятностей? Какие меры множеств используются для вычисления геометрической вероятности? Обобщите свой ответ на n-мерные пространства.
8. Дайте формулировки всех обозначенных в теме 1.2 теорем.
9. Приведите примеры зависимых и независимых событий.
10. Как связаны понятия «совместные» и «зависимые события»?
11. Как вычислить вероятность суммы двух несовместных событий?

3.1.2. Тема 1.2 Случайные величины

Перечень изучаемых вопросов:

Понятие случайной величины. Функция распределения случайной величины и ее свойства. Дискретные и непрерывные случайные величины. Плотность распределения непрерывной случайной величины и ее свойства. Биномиальное распределение. Распределение Пуассона. Геометрическое распределение. Равномерное распределение. Экспоненциальное

распределение. Нормальное распределение. Распределение Вейбулла. Гамма-распределение. Распределение хи-квадрат. Числовые характеристики случайных величин и их свойства: математическое ожидание, дисперсия, СКО, мода, медиана, начальные и центральные моменты высших порядков. Понятие многомерной случайной величины. Совместная функция распределения. Независимые случайные величины. Многомерное нормальное распределение. Функции случайных величин.

Методические указания к изучению:

Данная тема является достаточно объемной и потребует большой самостоятельной работы. Традиционно начать следует с изучения основных определений и понятий. Четко уясните понятия функции распределения. Известная функция распределения позволяет получить любую информацию о случайной величине, поэтому все усилия математической статистики, о которой пойдет речь дальше, связаны нахождением именно функции распределения или хотя бы установлением отдельных ее свойств. Случайные величины делятся на дискретные и непрерывные. Для дискретной случайной величины рассмотрите понятие закона распределения, как соответствия значения случайной величины вероятности появления этого значения. Для непрерывной — плотность распределения. Рассмотрите свойства биномиального распределение, распределения Пуассона, геометрического распределения, равномерного распределения, экспоненциального распределения, нормального распределения, распределения Вейбулла, гамма-распределение и распределение хи-квадрат. Составьте таблицу свойств распределений, зарисуйте графики, если это возможно, функции распределения и плотности распределения для различных параметров. Особое внимание уделите нормальному распределению.

Уяснив понятие функции распределения, перейдем к изучению числовых характеристик случайных величин. Целесообразно расчетные формулы характеристик изучать параллельно для дискретных и непрерывных случайных величин. Обратите внимание, что зачастую параметрами распределений являются именно значения числовых характеристик. Вам нужно знать следующие числовые характеристики и их свойства: математическое ожидание, дисперсия, СКО, мода, медиана. Рассмотрите обобщение числовых характеристик — центральные и начальные моменты высших порядков.

До сих пор мы имели дело с одномерной случайной величиной, но зачастую на практике случайные величины многомерны. Далее переходим к обобщению — многомерным случайным величинам. Дайте определение многомерной случайной величине и совместной функции распределения. Отдельно изучите многомерное нормальное распределение.

На практике придется столкнуться с понятием «функция случайной величины», т.е. случай, когда некоторая числовая величина является функцией (зависит) от одной или нескольких случайных величин. В общем

случае данная величина также является случайной, причем закон распределения этой новой случайной величины зависит как от законов распределения аргументов, так и от свойств самой функции. Рассмотрите задачи нахождения распределения такой величины, а также их числовых характеристик [3, разд. 6.7].

Литература:
[1] гл. 6-8, [3] гл. 4-8.

Контрольные вопросы:

1. Дайте определение случайной величине. Как связаны случайные величины и случайные события?
2. Докажите свойства функции распределения.
3. Как по известной функции распределения найти вероятность попадания случайной величины в интервал?
4. Как, зная плотность распределения, найти вероятность попадания случайной величины в интервал?
5. Приведите примеры дискретных случайных величин.
6. Перечислите и докажите свойства плотности распределения.
7. Как по известной плотности распределения найти функцию распределения?
8. Дайте характеристику равномерному закону распределения (плотность распределения, функция распределения, основные числовые характеристики).
9. Дайте характеристику нормальному закону распределения (плотность распределения, функция распределения, основные числовые характеристики).
10. Что такое функция Лапласа? Как по таблице значений функции Лапласа узнать вероятность попадания нормальной случайно величины в заданный интервал?
11. Дайте определение n-мерной случайной величине.
12. Дайте определение совместной функции распределения.
13. Что такое функция случайной величины?
14. Как найти функцию распределения функции от непрерывной случайной величины?
15. Как найти функцию распределения функции от дискретной случайной величины?

3.1.3. Тема 1.3 Предельные теоремы

Перечень изучаемых вопросов:

Последовательности случайных величин. Сходимость последовательности. Сходимость почти наверное, по вероятности, в среднеквадратичном. Сходимость функций распределения. Неравенства

Чебышёва. Закон больших чисел в форме Чебышёва, в форме Бернулли. Центральная предельная теорема.

Методические указания к изучению:

Завершая изучение теории вероятностей, мы обратимся к важной группе теорем, называемых предельными. Эти теоремы являются своеобразным мостом между теорией вероятностей и математической статистикой, объясняя предельное поведение относительной частоты событий.

Дайте определение последовательности случайных величин и рассмотрите основные типы сходимости: сходимость почти наверное, сходимость по вероятности, сходимость в среднем квадратичном. Рассмотрите сходимость последовательностей функций распределения. Рассмотрите примеры.

Далее переходите к изучению закона больших чисел. Предварительно рассмотрите с доказательством первое и второе неравенства Чебышёва. Дайте определение закону больших чисел как критерию устойчивости средних арифметических случайных величин. Уясните его практический смысл, заключающийся в том, что при предельном росте числа испытаний средние арифметические случайных величин ведут себя как неслучайные и совпадают со своими средними значениями.

Сформулируйте теоремы: закон больших чисел в форме Чебышёва и закон больших чисел в форме Бернулли. Второй является частным случаем первого.

Переходите к формулированию центральной предельной теоремы. Эта теорема играет важную практическую роль, обосновывая использование нормального закона распределения для моделирования различного вида шумов и погрешностей, как суммарного результата воздействия большого количества случайных факторов. В завершение сформулируйте интегральную теорему Муавра – Лапласа.

Несмотря на необязательность рассмотрения подробных доказательств теорем в рамках данного курса, тем не менее для студентов, претендующих на получение высоких оценок, изучение доказательств рекомендуется, так как позволяет понять механизмы сформулированных законов.

Литература:

[3] гл. 9.

Контрольные вопросы:

1. Дайте определение основным типам сходимости. В чем их отличие?
2. Сформулируйте первое и второе неравенства Чебышёва. Приведите примеры.
3. Приведите геометрическую интерпретацию первого и второго неравенств Чебышёва.

3.1.4. Тема 1.4 Основные понятия математической статистики

Перечень изучаемых вопросов:

Предмет математической статистики. Генеральная и выборочная совокупности. Выборка, объем выборки.

Методические указания к изучению:

Переходя к изучению математической статистики, следует обратить внимание на предмет данной науки – изучение свойств случайных величин, получаемых по результатам экспериментальных исследований. Следует еще раз проанализировать ситуацию проведения эксперимента и уяснить различия предметов теории вероятностей и математической статистики. Основоположниками математической статистики как науки являются Бернулли Я., Лаплас П., Пирсон К., дальнейшее развитие математическая статистика нашла в работах Крамер, Фишера Р., Неймана Ю., существенный вклад внесли наши соотечественники – П.Л. Чебышёв, А.М. Ляпунов, А.Н. Колмогоров и др.

Далее дайте определение генеральной и выборочной совокупности. Под выборкой далее будем понимать последовательность одинаково распределенных случайных величин, распределение которых совпадает с генеральной совокупностью.

Литература:

[1] гл. 15, [2] гл. 1.

Контрольные вопросы:

1. Сформулируйте основные задачи математической статистики.
2. Что такое генеральная и выборочная совокупность?
3. Приведите примеры практических проблем, приводящих к статистическим задачам.

3.1.5. Тема 1.5 Статистическое оценивание параметров распределения

Перечень изучаемых вопросов:

Определение оценки параметра распределения, свойства оценок, точечное оценивание, метод моментов, метод максимального правдоподобия, интервальное оценивание.

Методические указания к изучению:

Переходя к изучению данной темы, следует вспомнить такие понятия теории вероятностей, как функция распределения и числовые характеристики: математическое ожидание, дисперсия, мода, медиана, асимметрия, эксцесс, центральный и начальный моменты.

Далее дайте определение точечной оценке параметров распределения.

Под оценкой следует понимать некоторую функцию от значений выборки, которая в некотором смысле приближает реальное значение параметра генеральной совокупности. Следует уяснить, что любая оценка, будучи функцией от случайной выборки, есть случайная величина, поэтому, как и любая случайная величина, оценка имеет свой закон распределения и соответствующие числовые характеристики. Рассмотрите основные свойства точечных оценок: смещенность, состоятельность, эффективность.

Рассмотрите доказательство несмещенности, состоятельности и эффективности выборочного среднего. Покажите, что распределение выборочного среднего подчиняется нормальному закону. Переходите к изучению свойств выборочной и исправленной дисперсии. Покажите, что выборочная дисперсия – смещенная состоятельная оценка дисперсии генеральной совокупности.

В статистике разработано достаточно большое количество методов получения точечных оценок, рассмотрите метод моментов, метод максимального правдоподобия и метод наименьших квадратов.

Метод моментов, предложенный К. Пирсоном, заключается в следующем: в определении значения точечной оценки путем решения системы уравнений, полученной путем рассмотрения соответствующих выборочных моментов и их зависимости от оцениваемого параметра. Рассмотрите пример оценки параметра биномиального распределения (вероятности «успеха» в n независимых повторных экспериментах).

Метод максимального правдоподобия был предложен Р. Фишером. Дайте определение функции правдоподобия. Предположим, что функция правдоподобия известна с точностью до оцениваемого параметра, тогда оценкой максимального правдоподобия параметра называют такое его значение, при котором функция правдоподобия достигает максимальное значение. Если функция правдоподобия дифференцируема по искомому параметру, то максимальное значение будет достигаться в соответствующих критических точках.

Далее рассмотрим интервальное оценивание параметров распределения. В отличие от точечного значения, интервальная оценка дает вероятностную оценку точности оценивания неизвестного параметра. Т. е. получаемая оценка представляет собой не конкретное (вспомним – случайное) значение, а интервал, в который оцениваемый параметр попадает с заданной вероятностью. Дайте определение нижней и верхней границе интервальной оценки, коэффициенту доверия (доверительной вероятности, уровню доверия).

Построение интервальной оценки сводится к выполнению следующих шагов: построение центральной статистики с известной функцией распределения, нахождение соответствующих квантилей по известному значению уровня доверия, определение нижней и верхней границ интервалов.

Рассмотрите конкретные примеры построения интервальных оценок, например параметра экспоненциального распределения, математического ожидания нормального распределения при известной и неизвестной дисперсии.

Литература:
[1] гл. 16, [2] гл. 2-3.
Контрольные вопросы:

1. Какую оценку называют несмещенной, состоятельной, эффективной?
2. В чем отличие точечной и интервальной оценки?
3. В чем заключается метод моментов в построении точечной оценки?
4. Что такое функция правдоподобия?
5. В чем состоит метод максимального правдоподобия построения точечной оценки?
6. Что такое центральная статистика?
7. Какую статистику используют для построения интервальной оценки математического ожидания нормального распределения при известной дисперсии?

3.1.6. Тема 1.6 Проверка статистических гипотез

Перечень изучаемых вопросов:

Понятие статистической гипотезы, статистический критерий, ошибки первого и второго рода, критерий Неймана – Пирсона, отношение правдоподобия, критерий согласия.

Методические указания к изучению:

Предположим, что на основе некоторой априорной информации известно значение параметра распределения, или вид распределения наблюдаемых данных. На основе этой информации формируется гипотеза о том, что это значение совпадает с теоретическим. Такие гипотезы называются статистическими, а соответствующие методы проверки статистических гипотез позволяют определить, можно ли доверять полученным данным и следует ли принять выдвинутую гипотезу.

Общая схема проверки гипотез заключается в построении некоторой функции – статистического критерия, распределение которой известно для случая, когда гипотеза верна. Тогда можно определить некоторое множество значений (критическое множество) данного критерия, вероятность появления которого в случае, если гипотеза верна, маловероятна и считать, что если значение критерия попадает в это множество, то гипотезу следует отвергнуть, в противном случае принять.

Рассмотрите ситуации принятия неверной гипотезы и отказа от верной гипотезы, дайте определение ошибкам первого и второго рода и соответствующим им вероятностям.

Рассмотрите подробно критерий Неймана – Пирсона (критерий отношения правдоподобия). Рассмотрите пример построения наиболее мощного критерия проверки гипотезы о значении математического ожидания нормального распределения и значения параметра экспоненциального распределения.

Изучите вопрос определения объема выборки, обеспечивающего заданные вероятности ошибок первого и второго рода.

Рассмотренные выше методы предполагают известную форму распределения генеральной совокупности, т. е. гипотезы строились относительно параметров известного распределения. Перейдем к непараметрическим гипотезам.

Критерий согласия – статистический критерий, предназначенный для обнаружения расхождений между теоретической статистической моделью и экспериментальными данными, которые, по версии исследователя, должны описываться соответствующей теоретической моделью.

Рассмотрите критерий Колмогорова, проверяющий простую гипотезу о совпадении непрерывной теоретической и эмпирической функций распределения. Далее переходите к рассмотрению критерия согласия хи-квадрат.

Зачастую возникает вопрос проверки взаимосвязи или взаимной независимости двух наборов данных. Рассмотрите критерий Спирмена.

Литература:

[1] гл. 19, [2] гл. 4-5.

Контрольные вопросы:

1. Что называют уровнем значимости критерия?
2. Что называют мощностью критерия?
3. В соответствии с теоремой Неймана – Пирсона какой критерий является наиболее мощным?
4. Какими свойствами обладает ранговый коэффициент корреляции Спирмена?
5. Какие критерии называют критериями согласия?
6. В чем заключается критерий Колмогорова проверки гипотез?

3.1.7. Тема 1.7 Разведочный статистический анализ и визуализация данных

Перечень изучаемых вопросов:

Сторителлинг. Таблицы. Диаграмма. Инфографика. Дашборды.

Методические указания к изучению:

Переходя к более сложным статистическим темам, остановимся на вопросах визуализации данных и подбора подходящих средств визуализации для необходимых задач иллюстрации результатов и разведочного анализа.

В бизнесе визуализация необходима для решения широкого спектра задач, начиная от кадровых вопросов и заканчивая предоставлением скидки конкретному покупателю. Инструменты для визуализации бизнес-данных относятся к классу BI-систем.

Одной из проблем анализа больших данных является сложность представления результатов данного анализа в понятном для конечного пользователя виде. В рамках данного раздела дисциплины предлагается изучить современные подходы к визуализации и представлению данных на основе платформы визуализации Yandex.DataLens. В ходе лабораторных работ изучите основные концепции платформы, используя документацию разработчика.

Изучите основные «чарты», применяемые в визуализации: линейные диаграммы, колонки и гистограммы, круговые диаграммы, полярные графики, точечные графики, карты, деревья, временные диаграммы, пр. Рекомендуется обращать внимание на английский эквивалент наименований чартов. Обозначьте области применения различных диаграмм и таблиц. Найдите не менее 10 примеров различных чартов и их комбинаций. Приведите примеры удачной визуализации и неудачной. Изучите приведенную ниже схему (Рисунок 1).

Далее рассмотрите понятие «сторителлинг» в анализе данных и понятие «инфографика». Раскройте смысл контекста при представлении результата, какие цели с точки зрения визуализации преследуются при подготовке докладов по результатам анализа данных? Дайте определения понятию «дашборд» (информационная панель); какова задача интерактивности при представлении результатов? Приведите примеры не менее 10 доступных дашбордов, изучите их структуру.

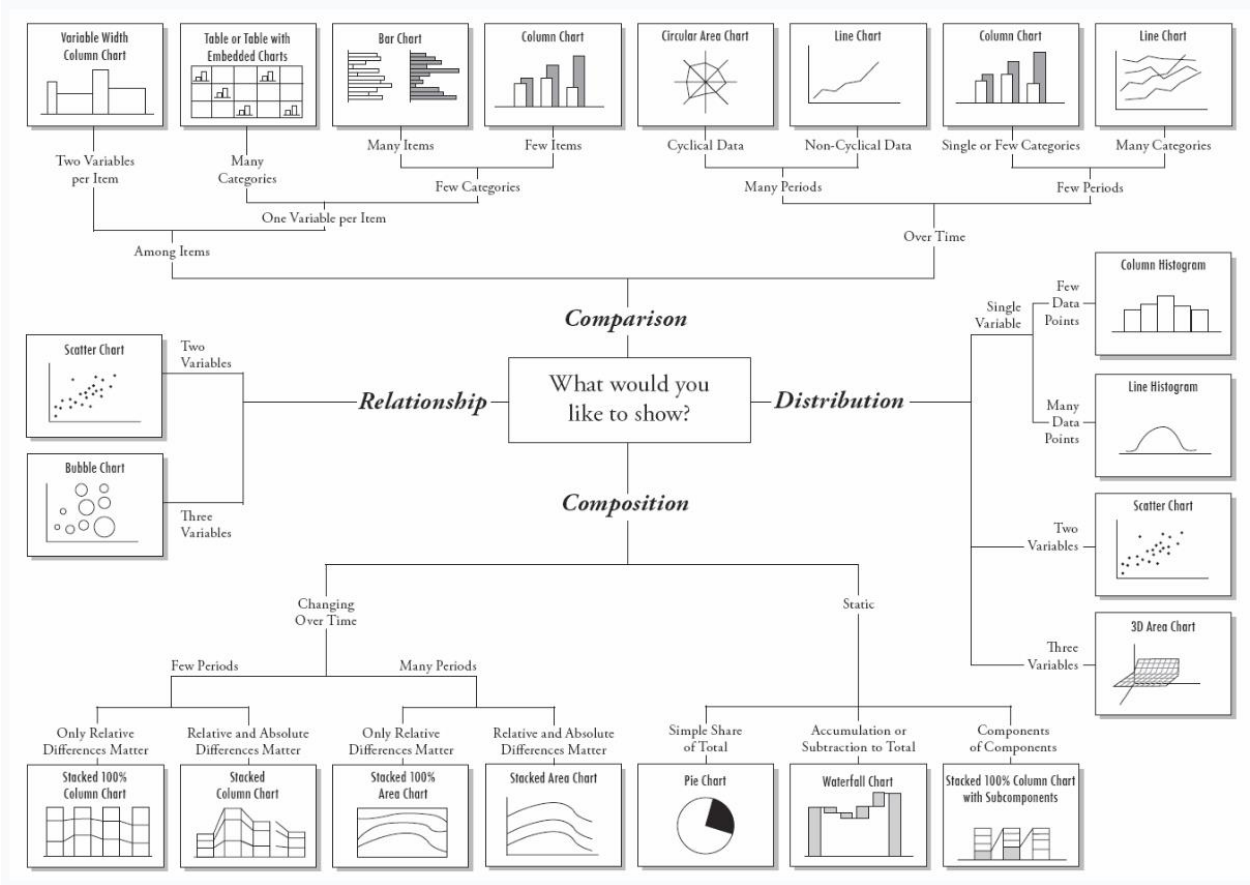


Рисунок 1. Рекомендуемые типы «чартов» в зависимости от визуализируемой информации (заимствовано: https://www.labnol.org/software/find-right-chart-type-for-your-data/6523/)

Литература:

[5] гл. 18-20.

Контрольные вопросы:

1. Что такое BI-система? Приведите примеры не менее трех BI-систем.
2. В чем преимущество интерактивных панелей (дашбордов) перед статической презентацией?
3. В чем отличие «подключения» от «датасета» на платформе Yandex.DataLens?
4. В чем преимущества использования облачных платформ визуализации? Приведите не менее пяти преимуществ.
5. Предложите схему организации взаимодействия пользователей (аналитика, системного администратора, клиента, специалиста отдела продаж) платформы Yandex.DataLens при подготовке и использовании информационных панелей для некоторого модельного предприятия. Рассмотрите реальные кейсы внедрения.

3.2. РАЗДЕЛ 2. КЛАССИФИКАЦИЯ И СНИЖЕНИЕ РАЗМЕРНОСТИ

3.2.1. Тема 2.1 Задача классификации

Перечень изучаемых вопросов:

Сущность задач классификации и снижения размерности. Многомерный статистический анализ. Дискриминантный анализ. Типовые практические задачи. Типология математических постановок задач классификации и снижения размерности. Основные этапы решения задачи классификации и снижения размерности.

Методические указания к изучению:

Данная тема является вводной к изучению широкого класса задач классификации и поиска общих свойств во множестве объектов. Приступая к изучению данного раздела, остановимся на ключевых методологических принципах: 1. Эффект существенной многомерности, 2. Возможность лаконичного объяснения природы анализируемых многомерных структур, 3. Использование «обучения» в подборе математических моделей классификации и 4. Оптимизационную природу формулировки задачи классификации. Рассмотрите каждый из принципов и приведите иллюстрирующие примеры. Изучите шесть типовых задач классификации, а также типовые задачи снижения размерности представленных в [5].

Процедура решения задачи классификации и снижения размерности может быть сведена к восьми этапам:

1. Установочный (предметно-содержательное определение целей исследования).
2. Постановочный (определение типа прикладной задачи в пределах используемого математического аппарата).
3. Информационный (составление плана сбора исходной информации и его реализация).
4. Априорный математико-постановочный (выбор базовой модели).
5. Разведочный (выбор и применение специальных методов статической обработки исходных данных).
6. Апостериорный математико-постановочный (уточнение выбора базовой математической модели).
7. Вычислительный (реализация численных алгоритмом идентификации модели).
8. Итоговый (формулирование результатов).

Сопоставьте данную процедуру с методологией CRISP-DM.

Литература:

[5] гл. 1.

Контрольные вопросы:

1. Что означает эффект существенной многомерности в задачах классификации и какие вызовы он представляет?
2. Как принцип возможности лаконичного объяснения связан с задачами классификации? Приведите пример.
3. Какие основные этапы включает процедура решения задачи классификации и снижения размерности?
4. Какие примеры задач классификации можно назвать для каждого из шести типов задач классификации?
5. Что такое оптимизационная природа задачи классификации и какие критерии оптимизации могут использоваться?
6. Какие три этапа охватывает процесс выбора базовой модели классификации?
7. Какие методы статистической обработки данных могут быть использованы на этапе разведочного анализа?
8. Как апостериорный математико-постановочный этап влияет на выбор модели классификации?
9. Какие метрики качества могут быть использованы для оценки результатов задачи классификации и какие аспекты они измеряют?

3.2.2. Тема 2.2 Классификации с учителем

Перечень изучаемых вопросов:

Классификация при известном распределении классов. Теоретические основы и практика применения дискриминантного анализа.

Методические указания к изучению:

Изучение данной темы начинается с рассмотрения правила классификации на основе отношения правдоподобия для случая двух классов. Рассмотрите основные модели распределений признаков. Сформулируйте понятие байесовского классификатора как частный случай отношения правдоподобия. Качество классификации в рассматриваемом случае определяется ошибками первого и второго рода. Вспомните соответствующие разделы ранее изученных курсов (технология Data Mining). Байесовский классификатор минимизирует вероятность принятия ошибочного решения. Далее рассмотрите вопросы классификации для случая более сложного разделения классов путем задания границы критической области: случай линейной гиперплоскости, кусочно-линейные классификаторы. Рассмотрите классификацию на основе минимизации функции потерь.

Далее переходите к рассмотрению трех и более классов. Дайте формальное описание задачи классификации в данном случае. В данном случае задача построения байесовского классификатора сводится к построению байесовских классификаторов для всех пар классов.

Литература:
[5] гл. 2.

Контрольные вопросы:

1. Каким образом вычисляется отношение правдоподобия для двух классов в задаче бинарной классификации?
2. Какие свойства распределения Гаусса делают его подходящим для моделирования признаков в задачах классификации?
3. Какие еще модели распределений часто используются для признаков в задачах классификации и в каких случаях они применимы?
4. Как алгоритм SVM (Support Vector Machine) решает задачу линейной классификации с помощью гиперплоскости? Какие объекты играют роль опорных векторов?
5. В чем состоит принцип кусочно-линейных классификаторов, таких как решающие деревья или случайные леса?
6. Какой метод оптимизации может использоваться для обучения параметров кусочно-линейных классификаторов?
7. Какова роль функции потерь в задаче классификации и как она связана с определением оптимальных параметров модели?
8. Какие примеры функций потерь вы можете привести и в каких сценариях они применяются?
9. Какие метрики используются для измерения качества классификации и каковы их интерпретации? Укажите примеры ситуаций, когда одна метрика более предпочтительна, чем другая.
10. Какие стратегии можно использовать для снижения ошибки первого рода или ошибки второго рода в задачах классификации?
11. Подробно опишите, как сводится задача построения байесовского классификатора для трех и более классов к построению байесовских классификаторов для всех пар классов.

3.2.3. Тема 2.3 Классификации без учителя

Перечень изучаемых вопросов:

Постановка задачи классификации в условиях отсутствия обучающих выборок. Задача расщепления смесей вероятностных распределений. Автоматическая классификация, основанная на описании классов «ядрами». Иерархическая классификация. Процедуры кластер-анализа и разделения смесей распределения при наличии априорных ограничений; выбора метрики и сокращения размерностей в задачах кластер-анализа. Интерпретация результатов автоматической классификации.

Методические указания к изучению:

Данный раздел дисциплины касается изучения задачи кластеризации или автоматической классификации, т.е. случая, когда данные обучающей выборки не содержат меток классов.

Начать изучение данной темы следует с формулирования задачи классификации объектов или признаков в условиях отсутствия разметки. Дайте определение понятию метрики между объектами и меры близости объектов друг к другу, дайте определению расстояния между классами. Вспомните известные метрики и их свойства. Изучите общую модель смеси вероятностных распределений и общую схему решения задачи автоматической классификации в рамках данной модели. Дайте понятие идентифицируемости смеси распределений.

Далее переходите к алгоритмам кластеризации. Рассмотрите эвристические алгоритмы, алгоритмы, использующие понятие центра тяжести, алгоритмы метода динамических сгущений.

Отдельный класс алгоритмов образован иерархической классификацией. Рассмотрите дивизимные и агломеративные алгоритмы. Изучите их графическую интерпретацию.

Рассмотрите алгоритмы решения задачи кластеризации при наличии априорных ограничений: при наличии неполных обучающих выборок, при ограничениях на связи между объектами.

Далее рассмотрите методы целенаправленного проецирования данных в пространство небольшой размерности с сохранением кластерной структуры.

Литература:
[5] гл. 5-10.

Контрольные вопросы:

1. Какие преимущества и ограничения имеет метод автоматической классификации на основе ядерных функций?
2. Что представляет собой иерархическая классификация и как она может быть применена для решения задач классификации с множеством классов?
3. Какие методы могут быть использованы для решения задачи разделения смесей распределения при наличии априорных ограничений?
4. Как выбор метрики влияет на результаты кластер-анализа и какие критерии следует учитывать при выборе метрики?
5. Какие методы визуализации и анализа можно применить для более наглядного понимания результатов классификации?

3.2.4. Тема 2.4 Снижение размерности признакового пространства и отбор наиболее информативных признаков

Перечень изучаемых вопросов:

Сущность задачи. Метод главных компонент. Модели и методы факторного анализа. Многомерное шкалирование.

Методические указания к изучению:

Сформулируйте задачу снижения размерности. Рассмотрите метод главных компонент и его графическую интерпретацию. Рассмотрите модели факторного анализа: общий вид модели и ее связь с главными компонентами, алгоритмы идентификации модели факторного анализа.

Рассмотрите статистическую модель метрического шкалирования. Дайте понятие погрешности аппроксимации. Рассмотрите структурную модель многомерного шкалирования.

Отдельно рассмотрите методы анализа и визуализации неколичественных данных: понятие анализа соответствий, множественный анализ соответствий.

Литература:

[5] гл. 13-17.

Контрольные вопросы:

1. В чем заключается метод главных компонент (PCA) и какой основной целью он служит при анализе данных?
2. Что означает «главные компоненты» в контексте метода главных компонент (PCA)?
3. Каким образом метод PCA позволяет снизить размерность данных?
4. Какие шаги включает процесс применения метода главных компонент для снижения размерности данных?
5. Что такое факторный анализ и какие типы данных исследуются с его помощью?
6. В чем заключается метод многомерного шкалирования (MDS)?
7. Какие метрики или меры сходства используются при применении MDS?
8. Чем отличается факторный анализ от метода главных компонент?

3.3. РАЗДЕЛ 3. ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

3.3.1. Тема 3.1 Корреляционного, регрессионного и дисперсионного анализа.

Перечень изучаемых вопросов:

Задачи корреляционного, регрессионного и дисперсионного анализа. Выборочный коэффициент корреляции, уравнение регрессии, однофакторный анализ.

Методические указания к изучению:

Установление зависимостей между наблюдаемыми данными является важной частью науки и инженерной практики. Выделим основные задачи

данного раздела: выявление наличия взаимосвязей между отдельными группами переменных (корреляционный анализ), установление аналитической зависимости вида Y=F(X), когда переменные носят количественный характер (регрессионный анализ), анализ влияния некоторых качественных параметров X на некоторую величину Y (дисперсионный анализ).

Вспомните определение понятий ковариация и корреляция, понятие зависимых и независимых случайных величин. Далее рассмотрите понятие корреляционного поля и корреляционной таблицы, дайте точечную и интервальную оценку коэффициента корреляции. Рассмотрите примеры данных имеющих разные значения коэффициентов корреляции и интерпретацию результатов.

Далее переходите к следующей задаче исследования зависимостей. Функцию, описывающую зависимость условного среднего значения выходной переменной Y от заданных фиксированных значений входных переменных X, называют функцией регрессии. Установление реального вида данной функции не всегда возможно, поэтому, как правило, ограничиваются некоторым приближением, например, линейным. Изучите уравнение линейной регрессии. Рассмотрите метод наименьших квадратов как способ получения коэффициентов уравнения по имеющимся экспериментальным данным. Рассмотрите применение регрессионного анализа в планировании экспериментов. Обзорно рассмотрите задачу проверки адекватности модели регрессии, значимости ее коэффициентов.

В завершении изучения раздела познакомьтесь с задачами дисперсионного анализа, как группой методов, позволяющим установить наличие, например влияния изменения некоторого качественного параметра на экспериментальный объект.

Литература:
[1] гл. 20, [2] гл. 6-8, [6].

Контрольные вопросы:

1. Перечислите задачи изучения статистических зависимостей.
2. Приведите примеры статистических зависимостей и практических проблем, приводящих к соответствующим разделам факторного анализа.
3. Какой статистический критерий используется для проверки гипотезы о равенстве 0 коэффициента корреляции?
4. Что называют коэффициентом детерминации?
5. Запишите формулу для вычисления оценки коэффициента корреляции.
6. Приведите пример оценки параметров линейной регрессии.
7. В чем отличие однофакторного и двухфакторного дисперсионного анализа?

4. ТРЕБОВАНИЯ К АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

4.1. ТЕКУЩАЯ АТТЕСТАЦИЯ

В ходе изучения дисциплины студентам предстоит пройти следующие этапы текущей аттестации: защита лабораторных работ на протяжение всего курса и выполнение расчетно-графической работы во втором семестре.

Преподаватель вправе выбрать методику оценивания знаний студентов: традиционная зачетно-экзаменационная либо балльно-рейтинговая.

4.2. ПОРЯДОК ПРИМЕНЕНИЯ РЕЙТИНГОВОЙ СИСТЕМЫ

В рамках балльно-рейтинговой системы выставляется оценка за качество выполнения и защиту лабораторных и контрольных работ.

Таблица 1. Виды деятельности и соотношение трудоемкости

Вид деятельности	Доля, %	Кол-во ед.	Макс. балл за ед.	Всего
Обязательные виды деятельности				
1 семестр				
Посещаемость занятий	40	N1	=400/N1	400
Выполнение лаб. работ (защита)	60	2	600	600
Итого:	100			1000
2 семестр				
Посещаемость занятий	20	N2	=200/N2	200
Выполнение лаб. работ (защита)	40	2	200	400
Контрольная работа (РГР)	40	1	400	400
Итого:	100			1000
Всего				2000
Дополнительные задания (по выбору студента в каждом семестре)				
Подготовка реферата (видео-доклада)	20		200	200
Решение дополнительных задач контрольной работы	10		100	100
Выполнение задания в рамках НИРС	50		500	500

4.3. УСЛОВИЯ ПОЛУЧЕНИЯ ПОЛОЖИТЕЛЬНОЙ ОЦЕНКИ

Для получения оценки «зачтено» в первом семестре необходимо выполнить все лабораторные работы и защитить их или в рамках БРС набрать не менее 60 % установленных баллов.

Во втором семестре выставляется дифференцированная оценка в рамках БРС:

<60 % – неудовлетворительно,

60-75 % – удовлетворительно,

75-85 % – хорошо,

>85 % – отлично.

5. ЛИТЕРАТУРА

1. Гмурман, В. Е. Теория вероятностей и математическая статистика: учебник для прикладного бакалавриата / В. Е. Гмурман, 12-е изд., Москва: Издательство Юрайт, 2015.

2. Горяинов, В. Б. Математическая статистика / В. Б. Горяинов, И. В. Павлов, Г. М. Цветкова. Москва: Изд-во МГТУ им. Н.Э. Баумана, 2001.

3. Печенкин, А. В. Теория вероятностей / А. В. Печенкин, О. И. Тескин, Г. М. Цветкова. 3-е изд., Москва: Изд-во МГТУ им. Н.Э. Баумана, 2004.

4. Айвазян, С. А. Прикладная статистика: Основы моделирования и первичная обработка данных: справ. изд. / С. А. Айвазян, В. М. Бухштабер, Л. Д. Мешалкин, Москва: Финансы и статистика, 1983. 471 с.

5. Айвазян, С. А. [и др.]. Прикладная статистика: Классификация и снижение размерности: справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин, Москва: Финансы и статистика, 1989. 607 с.

6. Айвазян, С. А. Прикладная статистика: Исследование зависимостей: справ. изд. / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин, Москва: Финансы и статистика, 1985. 487 с.

Локальный электронный методический материал

Александр Борисович Тристанов

ПРИКЛАДНАЯ СТАТИСТИКА И АНАЛИЗ ДАННЫХ

Редактор М. А. Дмитриева

Уч.-изд. л. 1,2. Печ. л. 1,6.

Издательство федерального государственного бюджетного
образовательного учреждения высшего образования
«Калининградский государственный технический университет».
236022, Калининград, Советский проспект, 1