Министерство науки и высшего образования Российской Федерации

Томский государственный университет
систем управления и радиоэлектроники

А.А. Захарова

АНАЛИЗ БОЛЬШИХ ДАННЫХ

Методические указания по самостоятельной работе студентов для направления 09.04.01 «Информатика и вычислительная техника»

Томск 2022

УДК 004.048
ББК 32.813
З-38

Рецензент:
Мицель А.А., профессор кафедры АСУ, докт. техн. наук

З-38 Захарова, Александра Александровна
Анализ больших данных: методические указания по самостоятельной работе студентов для направления 09.04.01 «Информатика и вычислительная техника» / А.А. Захарова. – Томск: ТУСУР, 2022. – 10 с.

Методические указания содержат задание и указания по самостоятельной работе по дисциплине «Анализ больших данных». Приведено описание основных разделов, перечень лабораторных работ, вопросы для контроля знаний.

Одобрено на заседании каф. АСУ протокол № 11 от 14.10.2021

УДК 004.048
ББК 32.813

ОГЛАВЛЕНИЕ

1 ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

Целью курса является освоение основных концепций и методов аналитики данных, особенностей областей применения и использования их как готового инструмента принятия решений при работе со структурированными и неструктурированными данными больших объемов. Целью преподавания данной дисциплины является формирование у студентов теоретических знаний и практических навыков по вопросам анализа данных; поиска управленческих решений; освоение студентами современных математических методов машинного обучения.

Задачи дисциплины:

– изучение основных понятий, процесса и технологий «больших данных»;

– изучение методов статистического анализа данных;

– изучение методов машинного обучения «с учителем» и «без учителя»;

– формирование у студентов знаний и умений, необходимых для эффективного управления техническими, организационными и экономическими системами.

Самостоятельная работа студента по дисциплине включает в себя проработку лекционного материала, подготовку к выполнению лабораторных работ, подготовку к защите отчетов по лабораторным работам, подготовку к зачету.

2 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Раздел 1 Введение в аналитику больших данных

Тема 1. Введение в анализ данных: понятие данных, типы шкал измерений, жизненный цикл данных, этапы решения задачи анализа данных и их взаимосвязи.

Литература 1,2

Тема 2. Концепция больших данных: предпосылки формирования тренда, определение больших данных, источники больших данных, основные характеристики больших данных (4V), качество исходных данных, структурированность данных, драйверы рынка Big Data, функции и задачи больших данных, составляющие аналитики больших данных, процесс и принципы аналитики Big Data, отличия Big Data и Business intelligence (BI-аналитики), техники и методы анализа Big Data, специалисты Data Scientists.

Литература 1,2

Тема 3. MapReduce и новый программный стек: потребности приложений для обработки Big Data, распределенные файловые системы, процесс вычислений в MapReduce, задачи-распределители, задачи-редукторы, группировка по ключу, комбинаторы, алгоритмы, в которых используется MapReduce, обобщения MapReduce для системы потоков работ, рекурсивные обобщения MapReduce, модель коммуникационной стоимости.

Литература 1,2

Раздел 2. Методы статистического анализа данных

Тема 4. Основные задачи и статистического анализа, корреляционный анализ, дисперсионный анализ, регрессионный анализ, факторный анализ, кластерный анализ, дискриминантный анализ, многомерное шкалирование.

Литература 2-5

Раздел 3. Методы машинного обучения

Тема 5. Методы машинного обучения без учителя: кластеризация, сетевой анализ

Литература 2-5

Тема 6. Методы машинного обучения с учителем: наивный Байесовский классификатор, комплексные модели, деревья решений, бустинг и бэггинг.

Литература 2-5

3 ЛАБОРАТОРНЫЕ РАБОТЫ

Лабораторные работы предназначены для формирования практических навыков и реализации методов и алгоритмов анализа данных с использованием электронных таблиц.

Лабораторная работа № 1.

Цель: научиться рассчитывать индекс экономики знаний (по методологии KAM Всемирного банка) на основе основных показателей и ограниченного количества стран.

Задачи:

1. Изучить методику Всемирного банка по расчету индекса экономики знаний и индекса знаний.

2. Рассчитать индексы экономики знаний по набору из 12 основных показателей для 15 стран, используя инструменты Excel (Calc).

Контрольные вопросы:

1. Понятие Индекса экономики знаний KEI, назначение, разработчик.
2. Структура Индекса экономики знаний KEI (подындексы и набор основных показателей, общее количество показателей)/

3. Методика расчета Индекса экономики знаний
4. Что понимается под ранжированием.
5. Для чего осуществляется нормирование.
6. Метод нормировки, применяемый для расчета KEI – основные этапы.
7. Какая шкала измерения используется в индексе экономики знаний?

Лабораторная работа № 2.

Цель лабораторной работы: научится строить регрессионную модель для классификации покупателей и оценивать параметры её качества.

Задание: Используя Excel или Calc постройте модель множественной регрессии, оцените качество модели.

Контрольные вопросы:

1. Понятие множественной регрессии.
2. Опишите процесс обучения регрессионной модели.
3. Показатели значимости регрессионной модели: коэффициент детерминации, критерии Фишера и Стьюдента.
4. Параметры качества регрессионной модели.
5. Кривая ошибок.

Лабораторная работа № 3.

Цель лабораторной работы: научиться создавать *модель наивного байесовского классификатора*

Задание: используя Excel или Calc проведите классификацию твитов по принадлежности к заданному типу.

Контрольные вопросы

1. Смысл теоремы Байеса.
2. Многомерная и мультиноминальная модель наивного Байесовского классификатора.
3. Для чего используется оценка апостериорного максимума.
4. Предположение условной независимости.
5. Проблема арифметического переполнения.
6. Оценка параметров Байесовской модели.
7. Проблема неизвестных слов.
8. Методика реализации наивного Байесовского классификатора.
9. Формирование вероятностного пространства.

Лабораторная работа № 4.

Цель лабораторной работы: научиться проводить кластерный анализ

методами k-средних и k-медиан.

Используя Excel или Calc проведите сегментирование клиентской базы компании (кластеризацию методом k-средних и k-медиан) для проведения таргетированных рассылок о предложениях компании.

Контрольные вопросы:

1. Классификация методов кластеризации.
2. Перечислите известные вам меры расстояния.
3. Метрики качества кластеризации. Силуэт.
4. Сущность метода k-средних.
5. Сущность метода k-медиан.

Лабораторная работа № 5.

Цель лабораторной работы: научиться проводить кластерный анализ на основе сетевых графов.

Задание: используя Excel или Calc проведите кластеризацию клиентской базы компании, использую модульную максимизацию сетевых графов.

Контрольные вопросы:

1. Понятие сетевого графа, вершины, ребра.
2. Матрица смежности.
3. Степень вершины графа.
4. Задача сетевого анализа.
5. Понятие сообщества в сетевом анализе.
6. Понятие случайного графа.
7. Методы разделения сети на кластеры.
8. Агломеративная и дивизионная кластеризация.
9. Понятие модульности.
10. Алгоритм «edge.betweenness.community».

Лабораторная работа № 6.

Цель лабораторной работы: научится строить комплексные модели (ансамбли моделей) для классификации покупателей и оценивать параметры качества.

Используя Excel или Calc на основе технологий бэггинга и бустинга постройте комплексные модели, оцените их качество.

Контрольные вопросы

1. Понятие ансамбля моделей.
2. Понятие бэггинга. Назначение, основные этапы.
3. Понятие бустинга. Назначение, основные этапы.

4. Как рассчитывается прогнозное классификационное значение ансамбля моделей при бэггинге?

5. Как рассчитывается прогнозное классификационное значение ансамбля моделей при бустинге? Параметр α.

6. Алгоритм «Случайный лес».

7. Загрязнение Джини.

4 ВОПРОСЫ ДЛЯ КОНТРОЛЯ ЗНАНИЙ ПО ДИСЦИПЛИНЕ

Раздел 1

1. Раскройте соотношение понятий: данные, информация, знания.
2. Дайте определения понятия «измерение» (в анализе данных).
3. Понятие шкалы измерений.
4. Классификация шкал измерений.
5. Номинальная шкала: определение и возможные операции в этой шкале.
6. Порядковая шкала: определение и возможные операции в этой шкале.
7. Шкала интервалов: определение и возможные операции в этой шкале.
8. Шкала отношений: определение и возможные операции в этой шкале.
9. Шкала разностей: определение и возможные операции в этой шкале.
10. Абсолютная шкала: определение и возможные операции в этой шкале.
11. Дайте понятие жизненный цикл данных.
12. Перечислите о охарактеризуйте основные этапы жизненного цикла данных.
13. Перечислите о охарактеризуйте основные тапы решения задачи анализа данных и их взаимосвязи.
14. История формирования тренда больших данных.
15. Что такое «большие данные».
16. Перечислите основные характеристики больших данных.
17. Проблема качества и структурированности исходных данных.
18. В чем особенность аналитики Больших данных.
19. Опишите основные составляющие аналитики больших данных.
20. Техники и методы анализа, применимые к Big data по McKinsey.
21. Кто такие Data Scientists, необходимые им знания, умения, навыки.
22. Понятие и источники открытых данных.
23. Процесс вычислений в MapReduce: задачи-распределители, задачи-редукторы, группировка по ключу, комбинаторы.
24. Алгоритмы анализа данных, в которых используется наиболее часто используется MapReduce.
25. Обобщения MapReduce для системы потоков работ .
26. Рекурсивные обобщения MapReduce.

27. Модель коммуникационной стоимости: сущность и важность.

Раздел 2

28. Корреляционный анализ: сущность и примеры решаемых задач в управлении организациями.
29. Дисперсионный анализ: сущность и примеры решаемых задач в управлении организациями.
30. Регрессионный анализ: сущность и примеры решаемых задач в управлении организациями.
31. Факторный анализ: сущность и примеры решаемых задач в управлении организациями.
32. Кластерный анализ: сущность и примеры решаемых задач в управлении организациями.
33. Дискриминантный анализ: сущность и примеры решаемых задач в управлении организациями.
34. Многомерное шкалирование.
35. Понятие множественной регрессии.
36. Опишите процесс обучения регрессионной модели.
37. Показатели значимости регрессионной модели: коэффициент детерминации, критерии Фишера и Стьюдента.
38. Параметры качества регрессионной модели.
39. Кривая ошибок.

Раздел 3

40. Классификация методов кластеризации.
41. Перечислите известные вам меры расстояния.
42. Метрики качества кластеризации.
43. Сущность метода k-средних.
44. Сущность метода k-медиан.
45. Понятие сетевого графа, вершины, ребра.
46. Матрица смежности.
47. Степень вершины графа.
48. Задача сетевого анализа.
49. Понятие сообщества в сетевом анализе.
50. Понятие случайного графа.
51. Методы разделения сети на кластеры.
52. Агломеративная и дивизионная кластеризация.
53. Понятие модульности.
54. Алгоритм «edge.betweenness.community».
55. Смысл теоремы Байеса.
56. Многомерная и мультиноминальная модель наивного Байесовского классификатора (NBC).
57. Для чего используется оценка апостериорного максимума в NBC?

58. Предположение условной независимости в NBC.
59. Проблема арифметического переполнения в NBC.
60. Оценка параметров Байесовской модели.
61. Проблема неизвестных слов.
62. Методика реализации наивного Байесовского классификатора.
63. Формирование вероятностного пространства в NBC.
64. Комплексные модели (ансамбли моделей).
65. Деревья решений.
66. Бустинг: понятие и суть метода.
67. Бэггинг: понятие и суть метода.
68. Алгоритм «Случайный лес»: сущность, достоинства и недостатки.

5 УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ

Основная литература

1. Юре, Л. Анализ больших наборов данных / Л. Юре, Р. Ананд, Д.У. Джеффри ; перевод с английского А.А. Слинкин. — Москва : ДМК Пресс, 2016. — 498 с. — ISBN 978-5-97060-190-7. — Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: https://e.lanbook.com/book/93571 (дата обращения: 01.10.2021). — Режим доступа: для авториз. пользователей.

2. Мицель, А. А. Прикладная математическая статистика: Учебное пособие [Электронный ресурс] / А. А. Мицель. — Томск: ТУСУР, 2019. — 113 с. — Режим доступа: https://edu.tusur.ru/publications/9151 (дата обращения: 01.10.2021.

Дополнительная литература:

3. Мицель, А. А. Прикладная математическая статистика: Практические работы [Электронный ресурс] / А. А. Мицель. — Томск: ТУСУР, 2019. — 81 с. — Режим доступа: https://edu.tusur.ru/publications/9153http://88.204.72.158/learning/090401e/d17/090401e-d17-lect.pdf (дата обращения 01.10.2021)

4. Теория вероятностей и математическая статистика: Тезисы лекций [Электронный ресурс] / — Томск: ТУСУР, 2012. — 77 с. — Режим доступа: https://edu.tusur.ru/publications/1764, (дата обращения 01.10.2021)

5. Фролов, А.Н. Краткий курс теории вероятностей и математической статистики : учебное пособие / А.Н. Фролов. — Санкт-Петербург : Лань, 2017. — 304 с. — ISBN 978-5-8114-2460-3. — Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — Режим доступа: https://e.lanbook.com/book/93706 (дата обращения: 01.10.2021).