2022秋

数据科学与大数据导论
期末复习

曹劲舟

助理教授

深圳技术大学

大数据与互联网学院

2022年12月

期末考试题型

- 闭卷
- 题型：
  - 选择题（30分，10题）
  - 简答题（25分，5题）
  - 计算题（15分，1题）
  - 问答题（30分，3题）
  - 附加题（30分）

讲授内容

- 大数据概述：特征、性质、价值、发展趋势
- 数据科学基础：基本概念、流程与基本步骤
- 大数据分析算法：分类、聚类、回归、关联分析、异常检测
- 大数据可视化：概念、常用工具、案例
- 城市大数据科学
- 图数据计算：中心性计算、Page Rank、社区检测
- 文本挖掘：文本表示（TF-IDF）

大数据概述

- 大数据的特征
- 大数据处理步骤
- 大数据的类型

数据科学基础

- 数据科学的处理流程
- 数据采集
- 数据类型：数据的种类

数据科学基础

- 数据预处理：
  - 数据清洗：如何做？如何处理异常值（分箱操作）
  - 数据集成：实体识别步骤、各类相似度（距离）度量计算、 Pearson相关系数、Spearman Rank相关系数
  - 数据变换：数据规范化、离散化、熵的概念及其计算
  - 数据规约：主成分分析概念

大数据分析算法

- 算法的类型：监督/非监督，机器学习/数据挖掘概念
- 聚类：相关性、k-means步骤与优缺点
- 分类：SVM、决策树算法流程、KNN，分类算法评价标准（Accuracy/Precision/Recall/F1 Score）计算
- 回归分析：一元线性回归步骤、损失函数、多元线性回归、
- 关联分析：步骤、频繁项集计算、支持度与置信度概念与计算、频繁关联规则提取、APriori算法

大数据可视化

- 数据可视化的概念、作用、类型
- 统计图表类型和使用场景

图数据计算

- 应用领域（案例）
- 图的表示方式
- 相关指标计算：聚类系数、中心性
- Page Rank概念及计算
- 社区探测概念、模块度计算、Louvain算法步骤和计算

文本挖掘

- 文本分析的任务
- 独热向量编码计算
- TF-IDF计算

题型示例

选择题： 下面购物篮能够提取的3-项集的最大数量是多少（C）

ID	购买项
1	牛奶，啤酒，尿布
2	面包，黄油，牛奶
3	牛奶，尿布，饼干
4	面包，黄油，饼干
5	啤酒，饼干，尿布
6	牛奶，尿布，面包，黄油
7	面包，黄油，尿布
8	啤酒，尿布
9	牛奶，尿布，面包，黄油
10	啤酒，饼干

A、1　　B、2　　C、3　　D、4

题型示例

- **简答题：**

  大数据的4V

  例举可视化图表类型

题型示例

- **计算题：**

  TF-IDF

  寻找异常点

  频繁项集计算

  点度中心性、中介中心性、接近中心性计算

  社区探测计算

  熵值计算

题型示例

- **问答题：**

  叙述关联规则学习的步骤

  KNN算法的内容以及优缺点