# 复习提纲

## CH1 绪论

- 基本概念
  - 什么是大数据
  - 什么是数据挖掘
- 大数据4V特征
- 数据挖掘主要任务
- KDD过程(数据挖掘是核心)
- 挑战

## CH2 认识数据和数据预处理

- 数据属性类型
- 数据的统计描述
  - 中心性
    - 均值、众数、中位数、中列数
  - 散度
    - 极差、最大最小、四分位、百分位、方差
- 相似性度量
  - 标称型数据 Jaccard Distance
  - 数据型数据
    - 欧氏距离
    - 曼哈顿距离
  - 数据标准化、归一化
    - 归一化——最大最小法
    - 标准化——Z-Score法
  - 其他相似性
    - 余弦相似性
    - 马氏距离
    - 相关系数
    - KL散度
- 数据预处理
  - 清理
    - 缺失值
    - 噪声
  - 集成

        - 冗余分析
        - 卡方检验
    - 归约
        - 维度规约
            - PCA
            - 特征筛选
    - 变换
    - 离散化

# CH3 关联规则挖掘

- 定义
    - 关联规则挖掘
    - 频繁模式
    - 项集
    - 支持度
    - 支持度计数
    - 置信度
- Apriori算法
    - 剪枝基本思想
    - 算法流程、计算
    - 存在挑战和改进
- FP-Growth算法
    - 如何构造FP树
    - 如何挖掘
- 评估方法
    - 支持度
    - 置信度
    - 兴趣因子

# CH4 分类

- 基本概念
    - 监督学习、无监督学习
    - 生成模型、判别模型
- 分类算法
    - 决策树
        - 构造过程
        - 分裂属性选择
            - 信息增益、信息增益率、基尼指数
    - 过拟合问题
        - 如何避免
    - KNN
        - 基本思想
        - 优缺点
    - Naive Bayes
        - 理论

        - 优点
    - SVM
        - 支持向量
        - 小样本
        - 泛化能力
        - 基本思想
        - 非线性——核函数
    - ANN
- 集成学习
    - 集成准则
        - 准确性
        - 多样性
    - 集成策略
        - Bagging(RF)
        - Boosting(AdaBoost)
        - Stacking
- 评估
    - 准确度
    - 精度
    - 召回率
    - F1值
    - 类不平衡问题
        - 灵敏度
        - 特效性

# CH5 聚类和离群点检测

- 什么是聚类
- 聚类算法分类
- K-Means、DBSCAN
- 什么是离群点
- 离群点种类
    - 全局
    - 局部
    - 集体
- LOF

# CH6 大数据分析

- 哈希技术
    - 最小哈希
        - 签名矩阵计算
    - LSH(局部敏感哈希)
        - 在签名矩阵中寻找相似文档
- 数据流挖掘
    - 数据流挑战
    - 概念漂移

        - 检测方法
            - 基于分布
            - 基于错误率
    - 数据流分类
        - VFDT(霍夫丁上界)
    - 数据流聚类
        - Online
            - 维护微簇
        - Offline
            - KMEANS
            - DBSCAN
- Hadoop/Spark
    - 什么是Hadoop/Spark
    - Hadoop设计准则
    - Hadoop生态(MapReduce,HDFS)
        - HDFS存储
        - MapReduce计算
    - MapReduce vs Spark
    - Spark设计准则
        - RDD(弹性数据集)
        - 操作
            - Transformation
            - Action