数据挖掘导论复习大纲

第一章 绪论

1 分析能力的八个等级

2 数据挖掘的基本任务及建模过程

第二章 数据

1 数据的类型、支持的操作类型、数据集类型、相似度、相异度、密度

第三章数据探索

1 数据质量

2 数据特征分析

3 Python 主要数据探索函数

4 统计作图函数

第四章 数据预处理

1 数据预处理的主要任务

2 处理缺失值、异常值的方法

3 数据集成

4 数据变换

5 Python 主要数据处理函数

第五章 数据挖掘建模之分类与预测-背景及决策树

1 概念及实现过程

2 评价

3 决策树：基本流程、不纯性度量方法及计算、过拟合、欠拟合

第六章 数据挖掘建模之分类与预测-回归

1 线性回归、logistic 回归

2 偏差、方差及其关系

3 回归方法：岭回归、套索回归、弹性回归

4 SVM、软边缘、核技术

第七章 数据挖掘建模之分类与预测-集成学习

1 集成学习概念

2 Bootstrap 抽样方法

3 Bagging：基本原理、误差分析

4 Boosting：基本原理、与 Bagging 对比

5 组合策略、集成学习的错误率

6 Stacking：基本原理

7 随机森林：方法、泛化误差、

8 AdaBoost：基本原理

9 GBDT：基本原理、损失函数、优缺点

10 XGBoost、LightGBM：基本原理

第八章 数据挖掘建模之分类与预测-神经网络

1 神经元结构

2 优化：反向传播、激活函数、学习率、优化方法、过拟合

第九章 数据挖掘建模之分类与预测-深度学习

1 基本原理

2 AutoEncoder、CNN

第十章 聚类分析-背景及 k-means

1 概念

2 评价指标

3 k-means：原理

第十一章 聚类分析-其他算法

1 密度聚类：核心点、边界点、噪音点、密度直达、密度可达、密度相连、非密度相连、DBSCAN 参数影响及优缺点

2 层次聚类：两种类型、簇之间距离计算方法

3 谱聚类：相似度矩阵计算、Laplacian 矩阵（拉普拉斯矩阵）、Graph Cut、Min Cut、Ratio Cut、Normalized Cut、谱聚类的意义

第十二章 关联规则

1 概念：项集、支持度计数、支持度、频繁项集、关联规则、支持度、置信度

2 Apriori 算法：先验原理、候选集产生与剪枝、支持度计数