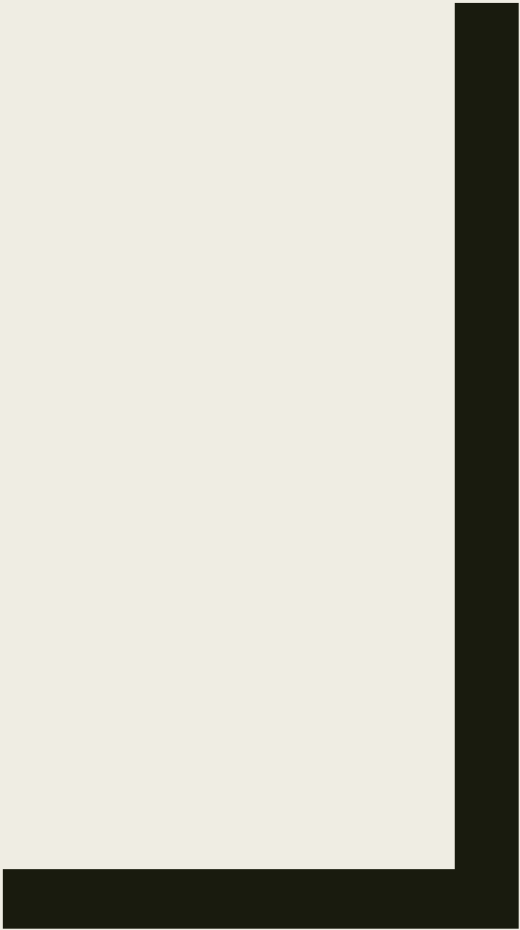


数据仓库&数据挖掘

谢希科

Outlines

1. 划重点
2. 课程报告要求



划重点

Overview

1. 概要
2. 数据的概念: 统计描述、可视化、距离度量
3. 预处理: 清洗、聚集、降维、离散化、二值化
4. 关联规则
5. 分类概念、决策树
6. 聚类

1. 概要

- 为什么要做数据挖掘?
- 数据挖掘需要解决什么问题?
- 数据挖掘的发展过程
- 数据挖掘的主要步骤

重要指数: ★ ★

2. 数据的概念：统计描述、可视化、距离度量

- 数据的基本概念
 - 标称、序数、二元、数值
- 数据常见的统计特征有哪些 分别是怎么计算
 - 均值，中位数，众数，分位数
- 怎么度量距离和相似性
 - 标称属性和二元属性的邻近性度量
 - 数值属性： L_p 距离

重要指数：★★★

3. 数据预处理

- 为什么要进行数据预处理
- 数据清洗主要解决什么问题
- 数据规约
 - 基本原理：小波，主成分分析

重要指数：★★★

4. 数据仓库

- 基本概念：
 - cell (单元格) cuboid (方体) cube (立方体)
 - 数据立方体: 维度, 度量, 格
 - 度量: 分布的、代数的、整体的
- 基本操作 (上卷、下钻、切片、切块)
- 数据立方体物化
 - 全物化, 半物化
 - 聚集路径的选择, 优化 (例题)

重要指数: ★★★★★

5. 关联规则

- 关联规则
 - 支持度和置信度的定义和计算
- Apriori算法的原理及实现
 - K -项集
 - 极大频繁项集、闭频繁项集的概念
- FP-growth算法（原理）
- 量化关联规则（对Apriori算法的简单扩展，了解原理）

重要指数：★★★★★

6. 分类：概念、决策树、最近邻、贝叶斯、集成学习

- 分类的概念
- 决策树
 - 什么优点 有什么缺点
 - 原理及实现
 - 算法（怎么选择最佳分裂点、信息增益、gini指数等计算、怎么处理不同类型的属性、分裂终止条件）
- 混淆矩阵（精度、召回率计算）
- 最近邻（原理）、贝叶斯（原理）、支持向量机（原理）
- 集成学习（原理，为什么好于单分类器）

重要指数：★★★★★

7. 聚类：概念、划分聚类、层次聚类

■ 聚类概念

- 几种聚类间距离计算(平均值, 最大最小距离, 期望值等)
- 聚类质量评价方法

■ K-means、K-Medoids 聚类方法（原理，区分，优化）

- 贪心策略与全局最优
- 参数选择

■ CF-Tree, BIRCH 算法（原理，优缺点）

■ DBSCAN 算法原理

- 基本概念（密度直接可达，密度可达，密度相连等）
- 优缺点

重要指数：★★★★★

课程报告

重要指数：★★★★★

评价标准

- 项目类型（提交：代码+报告）
 - A类：Spark系统级开发（满分65分）
 - B类：数据仓库或数据挖掘应用开发（满分50分）
 - C类：学术论文评讲（满分35分）
- 报告完整性：（20%）
 - 数据预处理、特征工程、模型选择、模型评价等步骤是否完整，方法是否合理。
- 代码实际效果：（35%）
 - 实验结果的效果展示。
- 创新性：（15%）
 - 有没有采用创新的方法。
 - 有没有针对现有方法进行过改进。
 - 研究的问题是不是一个新问题，没人研究过。
- 小组汇报（30%）
 - 组内成员分工是否明确，报告时全部人员到场。（10%）
 - PPT是否思路清晰，报告未超时。（10%）

报告时间

- 12月30日、31日两场。
- 每场约14支队伍。
- PPT报告（约8分钟），提问约3分钟（2-3个问题）。
- 报告顺序（组号）由助教在群里公布。
- 报告时候全组成员都必须上讲台。
- 来自听众的高质量“挑战性”问题可获全组额外加分。
- 所有队伍12月30日上午10点前将报告PPT发送至：

dmdw_ustc2019 at 163.com

附件名：组号+报告题目

未按时发送不得参加报告

The image features two large, thick, black L-shaped brackets. One is positioned on the left side, with its vertical line extending from the bottom and its horizontal line extending to the right at the top. The other is on the right side, with its vertical line extending from the top and its horizontal line extending to the left at the bottom. These brackets frame the central text.

GOOD LUCK!