

# 数据挖掘 复习 · 总结

---



# Outline

1

课程考核规划

2

理论内容复习

3

考试题型

4

注意事项

# 课程考核规划

- 课程主要内容

- 数据挖掘的基本概念、经典问题和算法、工具和应用
- 频繁模式挖掘、分类、聚类的算法及实践

<https://spoc.bnu.edu.cn/>

- 考核方式：

- 平时作业（60%）： 签到(5)+随堂作业与讨论（25）  
+编程实验（30）
- 期末考试（40%）： 闭卷考试

kaggle

Compete

# Outline

1

课程考核规划

2

理论内容复习

3

考试题型

4

注意事项

# 第一讲 概述

- **数据挖掘**：从海量数据中发现有趣模式的过程。作为知识发现过程，它通常包括数据清理、数据集成、数据选择、数据变换、模式发现、模式评估和知识表示。（挖掘什么类型的数据）
- **数据挖掘功能**：指定数据挖掘任务发现的模式或知识类型，包括特征化和区分，频繁模式、关联和相关性挖掘，聚类分析等。
- **成功应用**：如商务智能、Web搜索、生物信息学、金融、数字图书馆和数字政府等。

# 第三讲 认识数据

## 认识数据



### 基本概念

数据集  
数据对象  
属性



### 属性的类别

标称属性、二元属性、序  
列属性、数值属性  
连续属性、离散属性



### 数据的统计描述

中心趋势、离散趋势、图  
形展示 可视化



### 数据的相异度

数据矩阵 相异度矩阵  
相异度计算

# 第三讲 认识数据

## 相异度 计算

- 数据矩阵、相异度矩阵
- 不同类型变量/属性，采用不同相异度计算方法
- 数值型属性--> 标准化--> 欧几里得距离
- 二元属性--> 相依表--> 对称 VS 非对称
- 分类属性--> 不匹配率
- 序数属性--> 基于秩计算相异度
- 向量对象--> 余弦距离

# 第五讲 预处理

数据  
清理

数据  
归约

数据  
集成

数据变换  
与离散化

- 数据清理
  - 缺失值处理、噪声处理
- 数据集成与变换
  - 相同实体发现、冗余与冲突分析
  - 属性值规范化
- 数据归约
  - 维归约、数量归约
- 离散化与概念分层产生



# 第六讲 频发模式 挖掘

- ✓ 项集、频繁项集
- ✓ 闭频繁项集、极大频繁项集
- ✓ 关联规则：支持度、置信度
- ✓ 相关分析、相关度量

## 频繁项 集挖掘

FP增长

Apriori  
算法

使用垂直  
数据格式

## 相关分析 度量

- ✓ 提升度 (lift)
- ✓ 卡方  $\chi^2$
- ✓ 全置信度 (all-conf)
- ✓ 余弦 (cosine)
- ✓ 最大置信度 (max-confidence)
- ✓ Kulc度量 (配合IR)
- ✓ . . .

本章涵盖基础概念和基本计算，建议认真领会阅读教材内容

# 第八讲 分类基础

## 分类

### 评估分类和预测 方法的五条标准

-----

- ✓ 准确率
- ✓ 计算速度
- ✓ 鲁棒性
- ✓ 可伸缩性
- ✓ 可解释性

基于后验概率的  
贝叶斯定理

### 决策树算法

-----

ID3、C4.5、  
CART

朴素贝叶  
斯分类

逻辑回归  
分类

### 评估分类准 确率的方法

- ✓ 推荐方法：分层的k-折交叉确认
- ✓ 提高整体准确率方法：装袋和提升
- ✓ 准确率度量的替换：灵敏性、特效性和精度

# 第九讲 分类高级

贝叶斯信念  
网络分类

基于后验概率的  
贝叶斯定理

用后向传播  
分类

支持向量机

- 惰性学习法(或从近邻学习)
  - k-最近邻分类、基于案例的推理
- 其他分类方法
  - 遗传算法、粗糙集方法、模糊集方法
- 关于分类的其他问题
  - 多类分类、半监督分类
  - 主动学习、迁移学习

# 第十讲 聚类基础

◆ **簇**：是数据对象的集合，同一簇中的对象彼此相似，而不同簇中对象彼此相异。

◆ **聚类**：将物理或抽象对象的集合划分为相似对象的类的过程称为。

◆ **聚类算法**：

- **划分方法**: k-means, k-medoids
- **层次方法**: BIRCH, CHAMELEON
- **基于密度的方法**: DBSCAN

◆ **聚类评估**：

- **估计聚类趋势**：霍普金斯统计量评估是否存在非均匀分布
- **确定簇数**：
- **测定聚类质量**：外在方法，内在方法（轮廓系数）

# Outline

1

课程考核规划

2

理论内容复习

3

考试题型

4

注意事项

# 考试题型

一、选择 15\*2分

二、简答 3 \* 5分

三、计算 3 \* 5分

四、算法分析与流程图： 3 \* 8分

五、应用题： 16分

# 注意事项

认真准备  
开心考试

- 闭卷考试
- 100分钟
- 独立完成
- 相互支持：双0分



付出就有回报  
相信自己