- 项集，频繁闭项集，最大频繁项集的关系 =
- 机器学习：聚类，分类 =
- 数据预处理：标准化， =
- 抽样 =
- 频繁项集（重要） =
- 聚类：层次，模糊 =
- 离群点检测：基于一元正态分布 =
- 多选：数据挖掘的预测建模任务，影响k-means聚类算法的因素 =
- 分类器的评价：召回率，准确率 =
- 马氏对于欧式距离的优点 =
- 哪些要素，哪些特征对聚类分析有影响 =
- 填空：什么是频繁项集，决策树，信息熵（越大，不确定性越大）=
- 关联规则挖掘 =
- 特征选择，特征提取：类内散布矩阵，类间散布矩阵 =

a4内容

测试题

概述

数据挖掘的概念：从海量的数据中通过相关的算法来发现隐藏在数据中的规律和知识的过程。
数据挖掘的流程：明确问题、数据准备、数据挖掘、结果解释和评估。
数据挖掘的预测建模任务主要包括：回归和分类
数据挖掘任务分为：模式挖掘、描述建模、预测建模。

数据

五数概括，盒图，离群点

数据预处理

数据预处理的任务：数据清洗、数据集成、数据归约、数据变换。

数据清洗

马氏距离的优缺点：量纲无关，排除变量之间的相关性的干扰。

数据集成

卡方检验

数据规约

- 数量归约：通过直方图、聚类和数据立方体聚集等非参数方法，使用替代的、较小的数据表示形式替换原数据。
- 属性子集选择：检测并删除不相关、弱相关或冗余的属性。
- 抽样：使用比数据小得多的随机样本来表示大型的数据集。

抽样方法：不放回简单随机取样、放回随机简单取样、聚类取样、分层取样。

数据变换

概念分层
标准化：

- 小数定标规范化
- 最小-最大规范化
- z-score规范化 属性构造
  概念分层
  离散化：
- 分箱：等深分箱、等宽分箱。

概率分类

贝叶斯决策

12页 例4.1
20页 例4.2
37页 例4.3
67页 例4.4

- 最小风险贝叶斯决策规则公式
- 贝叶斯分类器的错误率
- 错误率的估计（56）

特征选择与特征提取

特征选择和提取的目的：经过选择或变换，组成识别特征，尽可能保留分类信息，在保证一定分类精度的前提下，减少特征维数，使分类器的工作即快又准确。

特征选择和特征提取的异同：

- 特征选择：从L个度量值集合{x1,x2,...,xL}中按一定准则选出供分类用的子集，作为降维（m维，m < L）的分类特征。
- 特征提取：使一组度量值（x1,x2,...,xL）通过某种变换$h_i(.)$产生新的m个特征(y1,y2,...ym),作为降维的分类特征，其中i=1,2,...,m;m<L。

类内散布矩阵：表示各样本点围绕均值的散布情况 ——该类分布的协方差矩阵。

类间散布矩阵：表示c类模式在空间的散布情况，记为Sb。
公式参考

基于类内散布矩阵的单类模式特征提取

27页 例5.2 39页 例5.3

频繁模式挖掘

频繁模式

- 项集：包含0个或者多个项的集合
- 支持度s：事务中同时包含集合A和集合B的百分比
- 置信度c：事务中同时包含集合A和集合B的事务数与包含集合A的事务数的百分比
- 频繁模式：支持度满足了最小支持度阈值的项集

Apriori
gt-growth

压缩频繁项集

- 挖掘闭模式
  如果X⊆Y ，且Y中至少有一项不在X中，那么Y是X的真超项集。如果在数据集中不存在频繁项集X的真超项集Y，使得X、Y的支持度相等，那么称项集X是这个数据集的闭频繁项集。
- 剪枝的策略
  - 项合并
    如果包含频繁项集X 的每个事务都包含项集Y，但不包含Y 的任何真超集，则X∪Y 形成一个闭频繁项集，并且不
  - 子项集剪枝 如果频繁项集X 是一个已经发现的闭频繁项集Y 的真子集，并且两者的支持度计数相等，则X 和Y 的所有后代都不可能是闭频繁项集，因此可以剪枝。
- 极大频繁项集
  - 如果在数据集中不存在频繁项集X的真超项集Y，使得X属于Y并且Y也是频繁项集，那么称项集X是这个数据集的极大频繁项集。
  - 可以推导出极大频繁项集是闭频繁项集，而闭频繁项集不一定是极大频繁项集。

回归

拟合优度检验

- 最小二乘法公式
- 离差
- 回归差
- 残差
- R^{2}
- 显著性检验

聚类

无监督
哪些要素，哪些特征对聚类分析有影响：离群点，噪声，簇数，初始点选取，数据集大小，密度
数据挖掘对聚类的典型要求如下：

- 处理不同属性类型的能力
- 可伸缩性
- 对于确定输入参数的领域知识的要求
- 发现任意形状的簇
- 处理噪声数据的能力
- 增量聚类和对输入次序不敏感
- 聚类高维数据的能力
- 基于约束的聚类

- 可解释性和可用性

聚类过程遵循的基本步骤：

- 特征选择
- 近邻测度
- 准则定义
- 算法调用
- 结果验证
- 结果判定

影响聚类算法效果的主要原因有：

- 特征选取
- 模式相似性测度
- 分类准则

基于划分的聚类

- k-means
- k-中心点（PAM）

基于层次的聚类

- DIANA：层次分裂聚类算法
- AGNES: 凝聚层次聚类方法

基于密度的聚类

- DBSCAN

基于网格的聚类

- STING
- CLIQUE

离群点

基于统计学的离群点检测

- 正态分布：
  13页 例9.1

基于近邻的离群点检测

基于聚类的方法

基于分类的方法

模糊模式识别

模糊集合

- 核
- 支集
- 幂集
- 表示方法（19页）
- 运算（35页）
- 截集

模糊关系

55页
58页
60页
64页

模糊模式分类的直接方法和间接方法

模糊K-均值算法

108页

分类

分类器评价

$$ 召回率 = \frac{将正类预测为正类}{原本的正类} = \frac{TP}{TP+FN} $$

$$ 准确率 = \frac{将正类预测为正类}{预测的正类} = \frac{TP}{TP+FP} $$

决策树

信息熵 $$ Ent = -\sum_{i=1}^{n}p_i log_2 p_i $$ 信息熵增益 $$ Gain(D,a) = Ent(D) - \sum_{v=1}^{V}\frac{|D^{v}|}{|D|}Ent(D^{v}) $$