**数据挖掘**

1. **数据预处理**
2. **考点**

数据清洗，集成，转换，降维

给定特征列表，根据数据特点建模，设计方案（比如分类）

1. **相关概念**
2. 为什么要预处理数据？

现实世界中的数据是脏的（不完全，噪音，不一致）

1. 数据预处理的形式

数据清理：补充缺失数据、平滑噪声数据、识别或删除离群点，解决不一致

数据集成：集成多个数据库、数据立方或文件

数据变换：规范化和聚集

数据归约：简化数据、但产生同样或相似的结果

1. 数据清理

缺失值处理（不完全）

-均值填补法：分别计算各组目标属性的均值

-随机填补法：采用某种概率抽样的方式，从有完整信息的元组中抽取缺失数据的填补值

-回归填补法：在现有观察值基础上，以含有缺失值的目标属性为因变量，以与目标属性相关性高的其它属性为自变量，建立最小二乘回归模型或判别模型，以估计缺失值

噪声数据（一个测量变量中的随机错误或偏差）

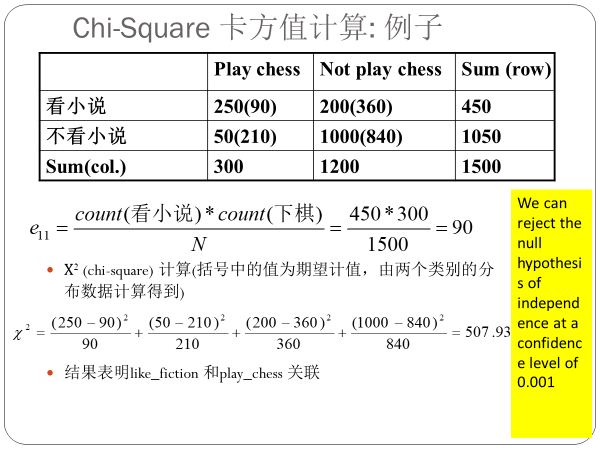
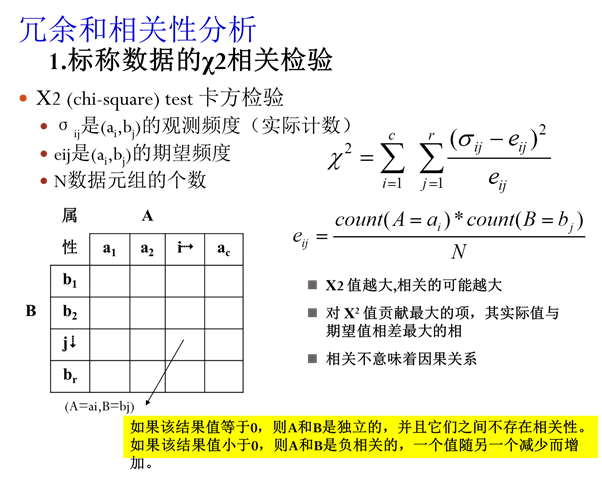
-分箱：通过考察数据的“近邻”（周围的值）来光滑有序数据的值,局部光滑。划分可以使用等频（同一区间内样本数目相同）或等宽（在同一个取值区间），用箱均值、用箱中位数、用箱边界（去替换箱中的每个数据）

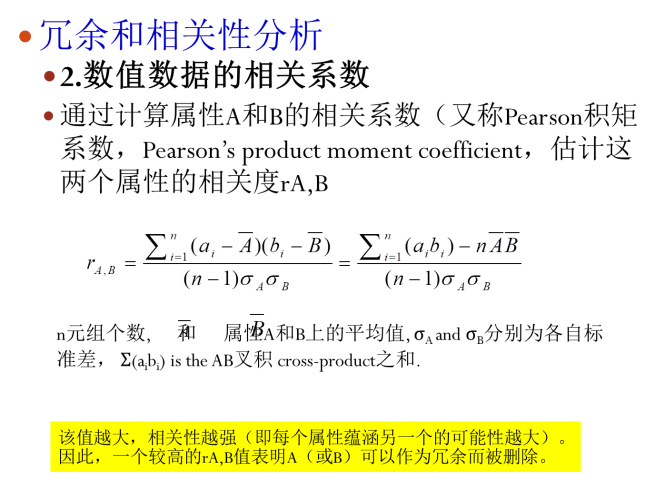
-回归：用一个函数（回归函数）拟合数据来光滑数据

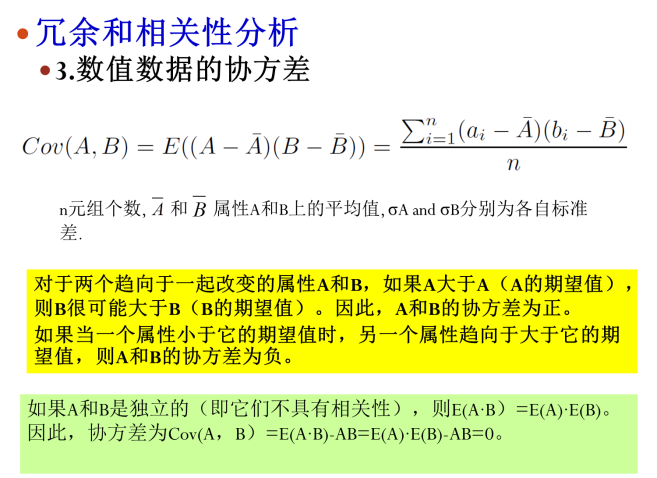
-聚类：将类似的值聚集为簇。检测离群点

1. 数据集成

合并多个数据源中的数据，存在一个一致的数据存储中，涉及3个主要问题：模式集成、冗余数据、冲突数据值







1. 数据归约（例子较多，有时间可以看看）

数据立方体聚集：

数据立方体存储多维聚集信息，提供对预计算的汇总数据进行快速访问。

属性子集选择：

通过删除不相关或冗余的属性（或维）减小数据集。逐步向前选择，逐步向后删除，逐步向前选择和逐步向后删除相结合，决策树归约

维度归约：

维度归约使用数据编码或变换，以便得到原数据的归约或“压缩”表示。串压缩、小波变换、主成分分析、奇异值分解

数值归约：

通过选择替代的、“较小的”数据表示形式来减少数据量

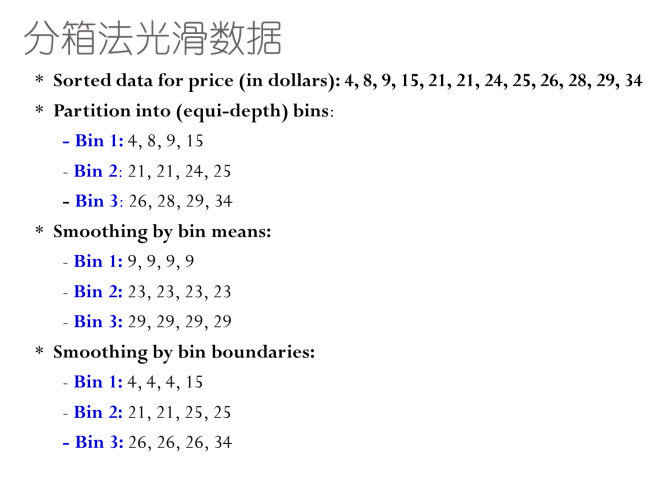
参数方法：假设数据适合某个模型，估计模型参数，仅存储的参数，并丢弃数据（孤立点除外)、回归（regression ）和对数线性模型

非参数方法：不假定模型、直方图、聚类、抽样

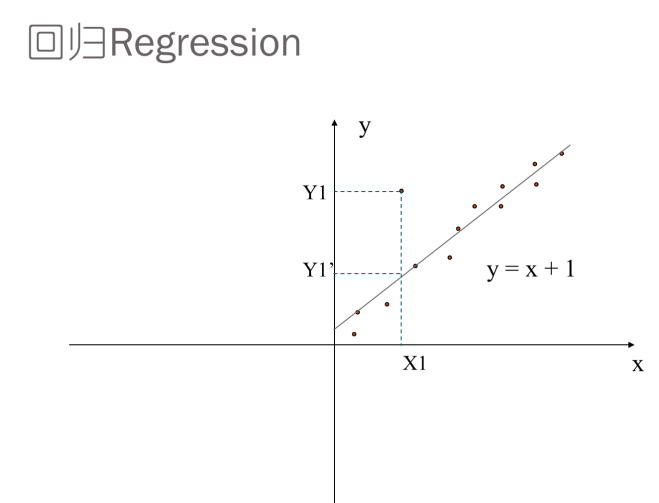
1. 数据变换

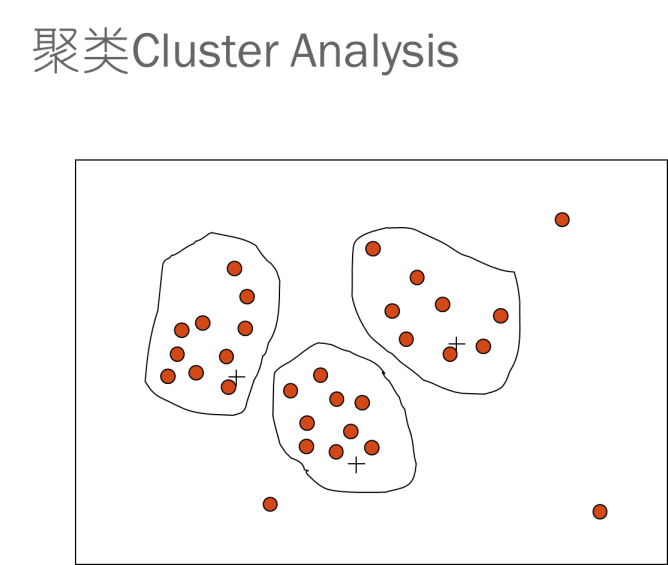
与考点2数据规范化重叠

1. **相关例题**
2. 缺失值处理

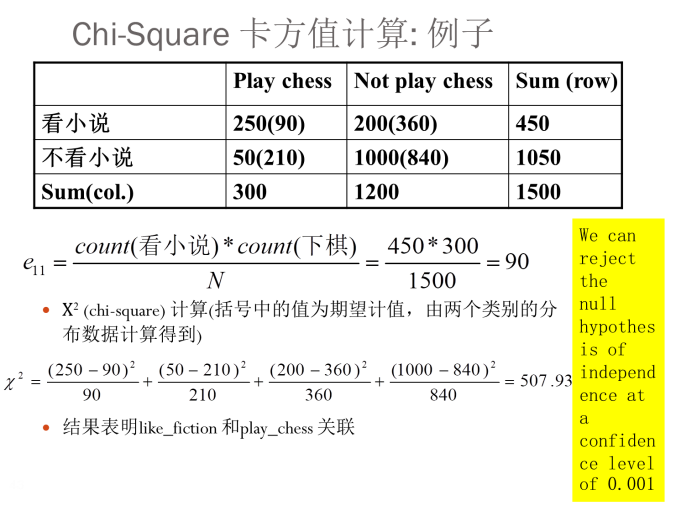


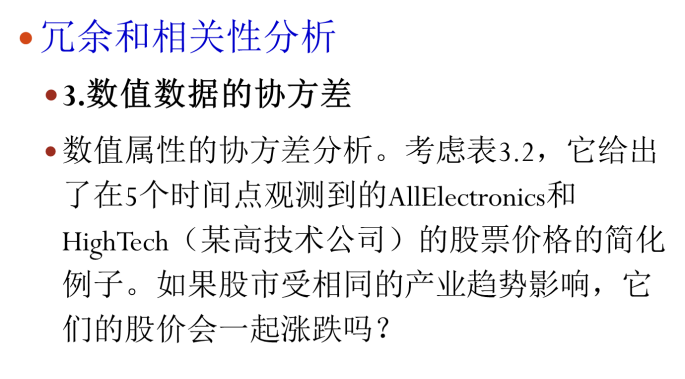
1. 噪声数据

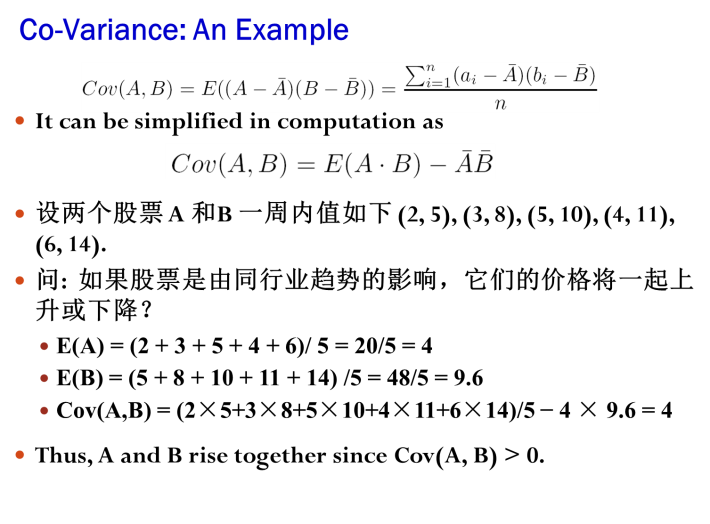




1. 数据集成-冗余和相关性分析



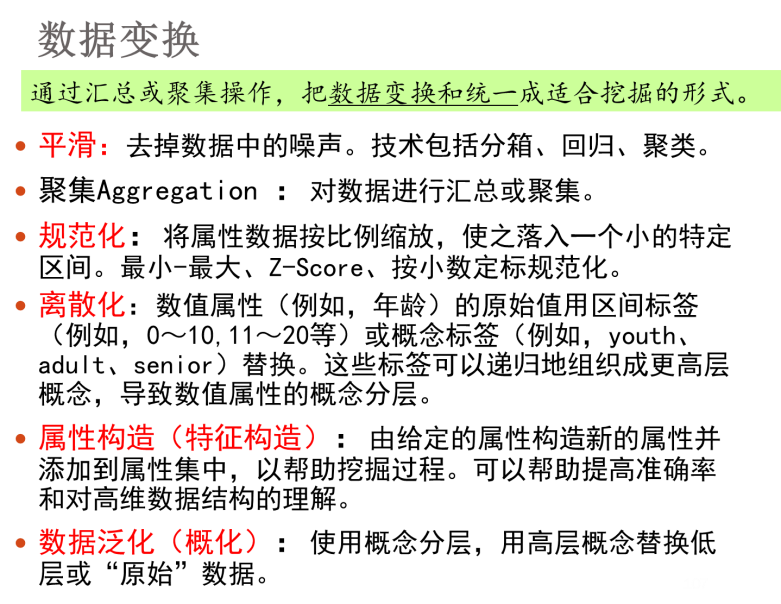


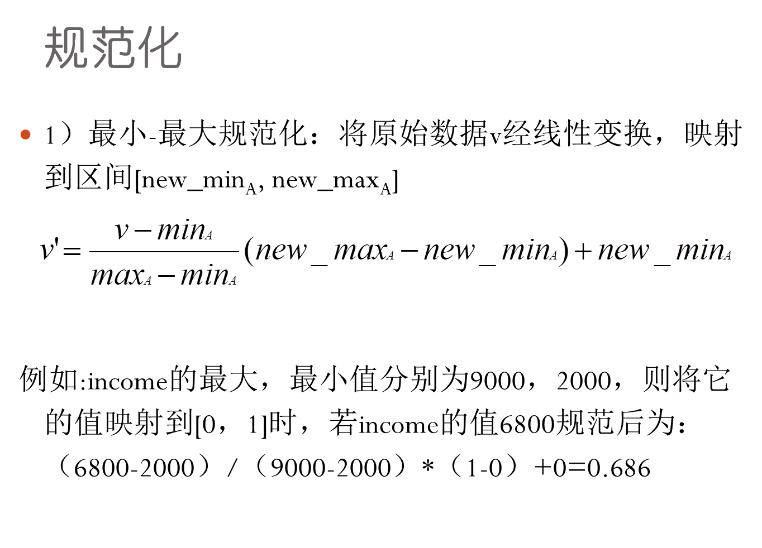


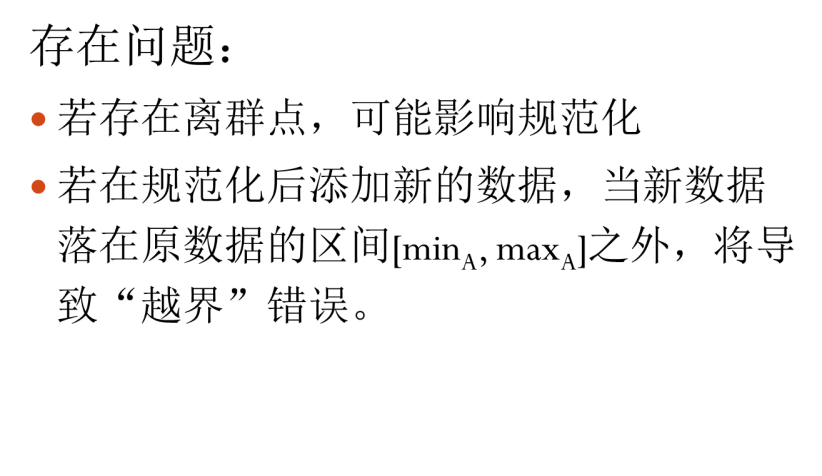
1. **数据规范化**
2. **考点**

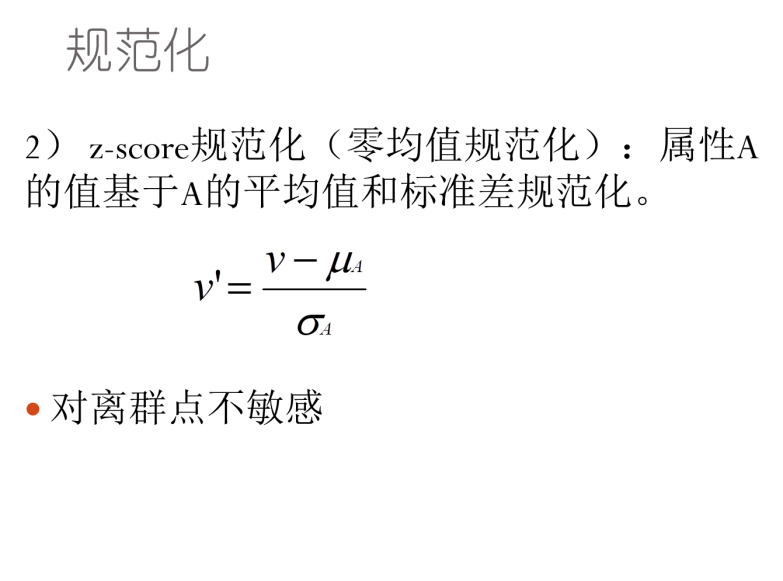
通过最大值和最小值做规范化，Z-Score

1. **相关概念**

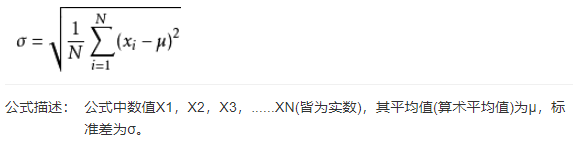




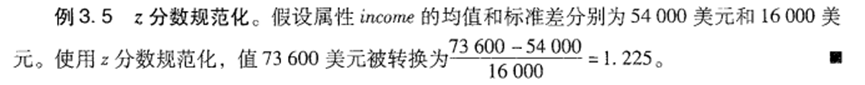




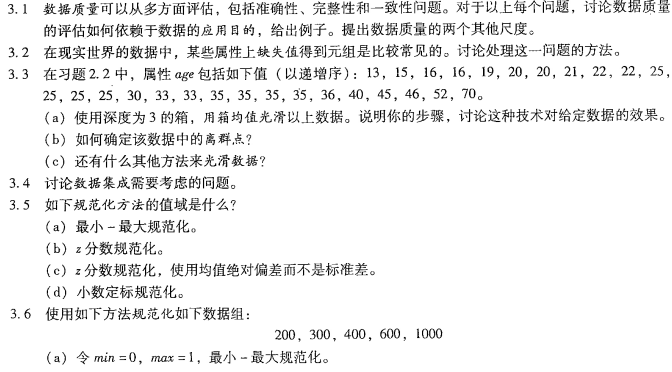
标准差公式：

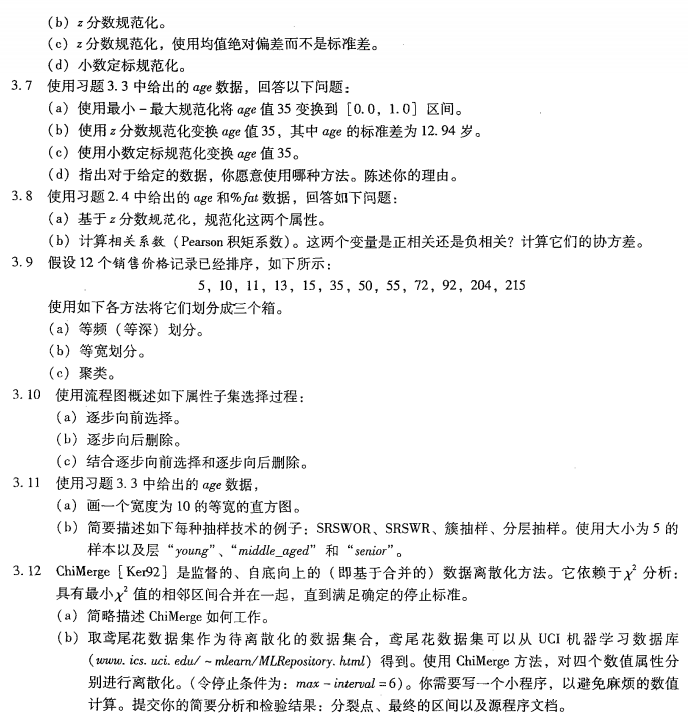


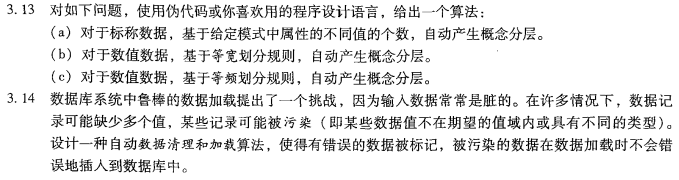
1. **相关例题**



1. **对应课本习题**



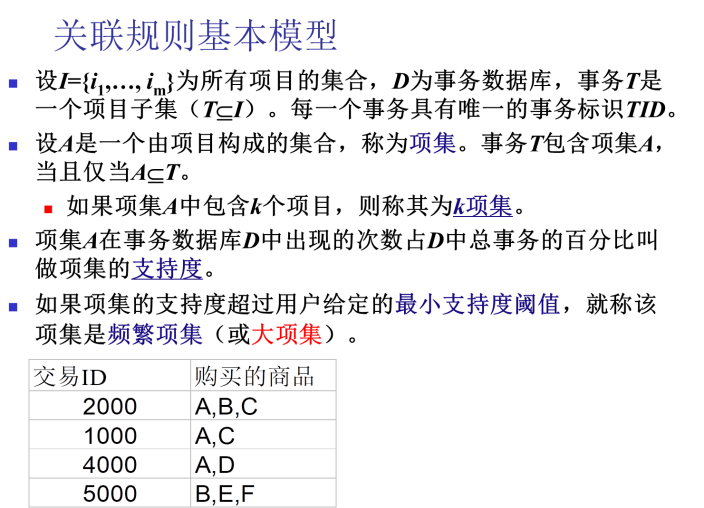


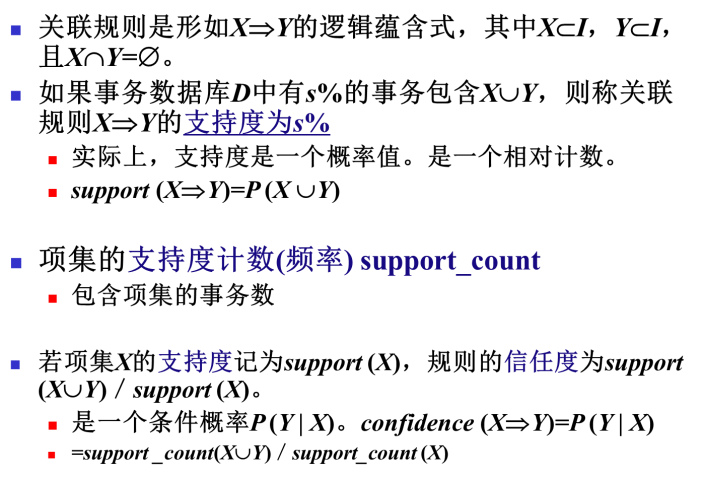


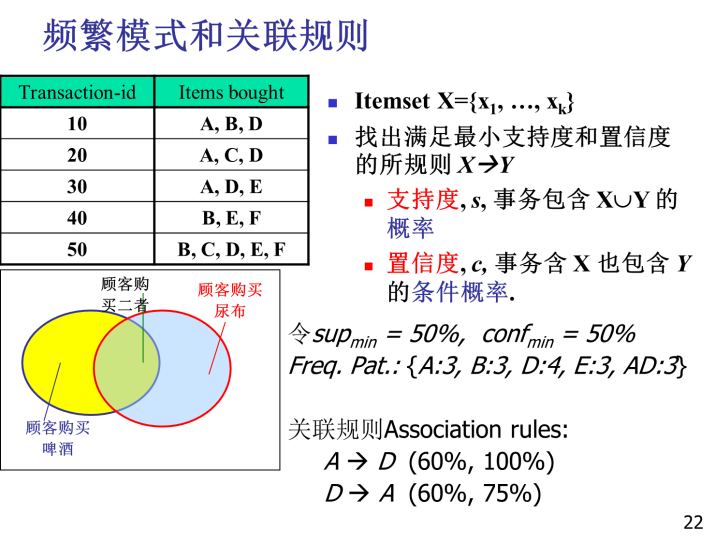
1. **关联规则**
2. **考点**

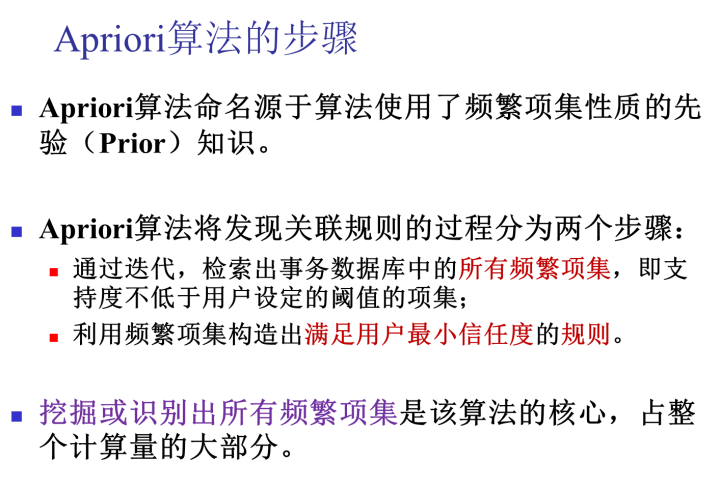
Apriori算法

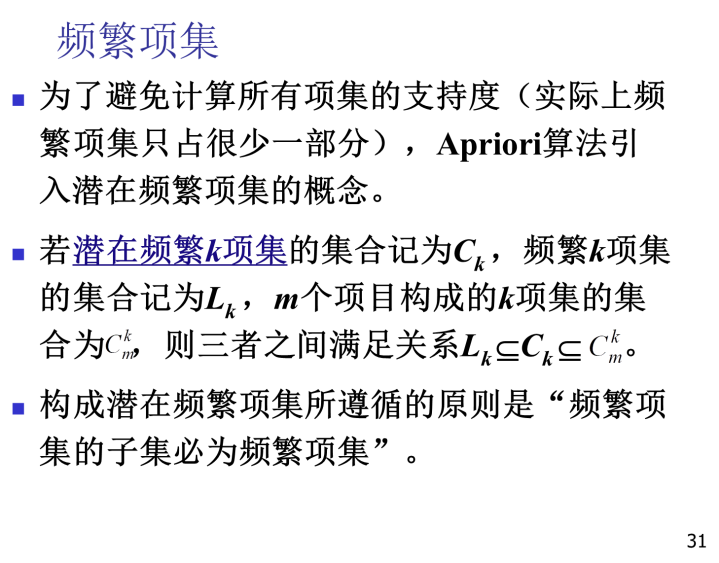
1. **相关概念**



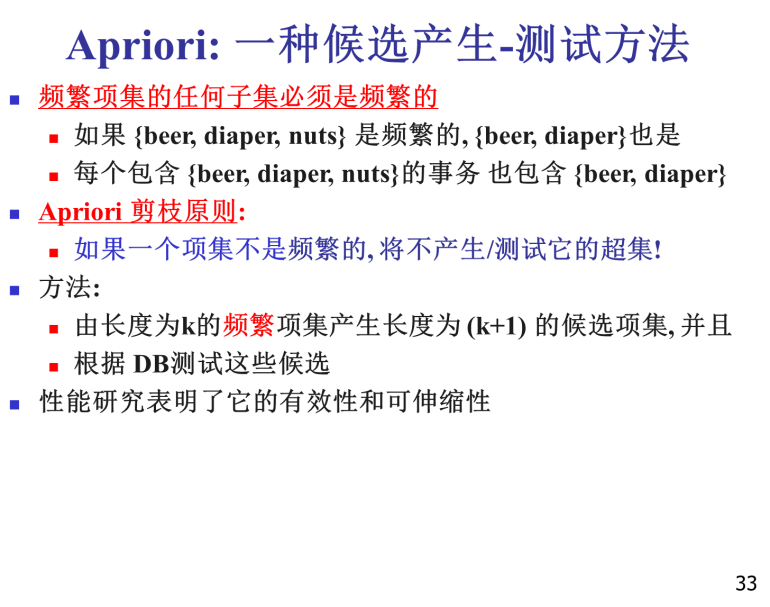


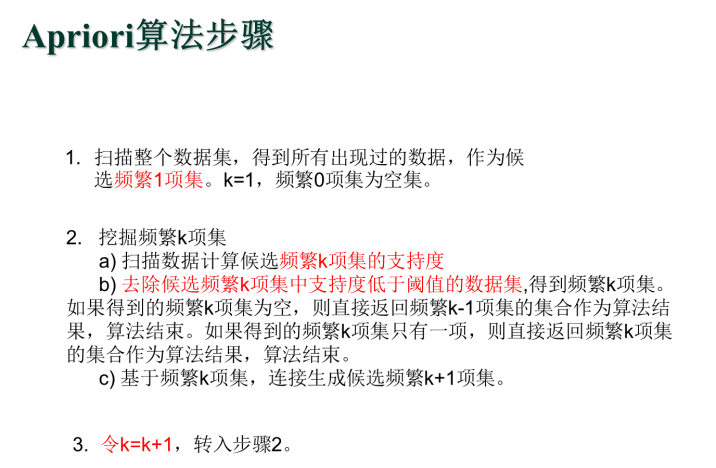


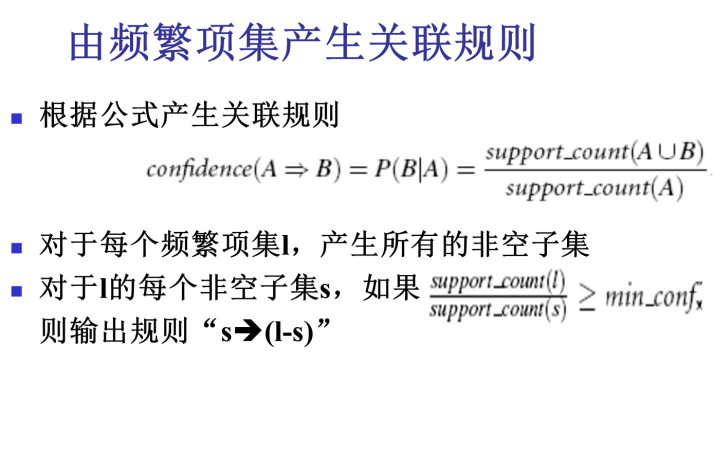




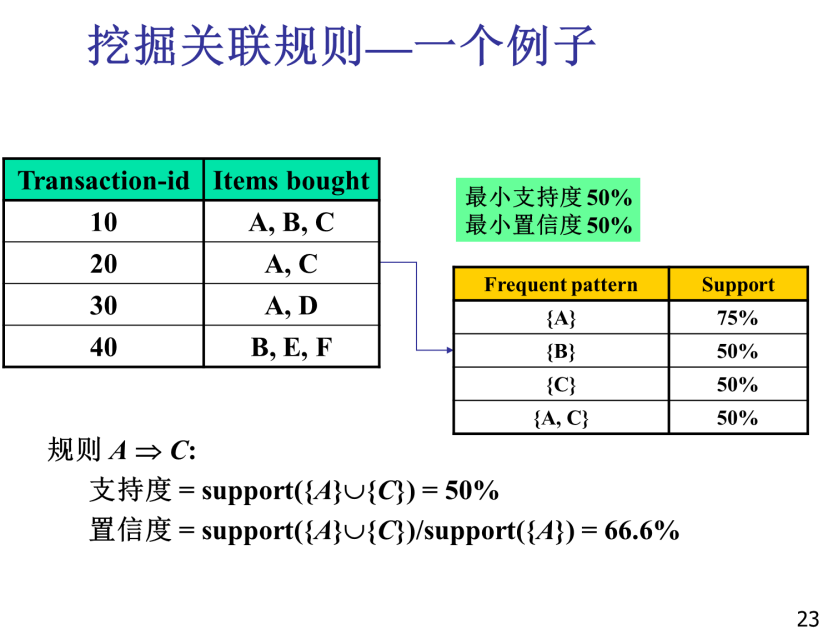


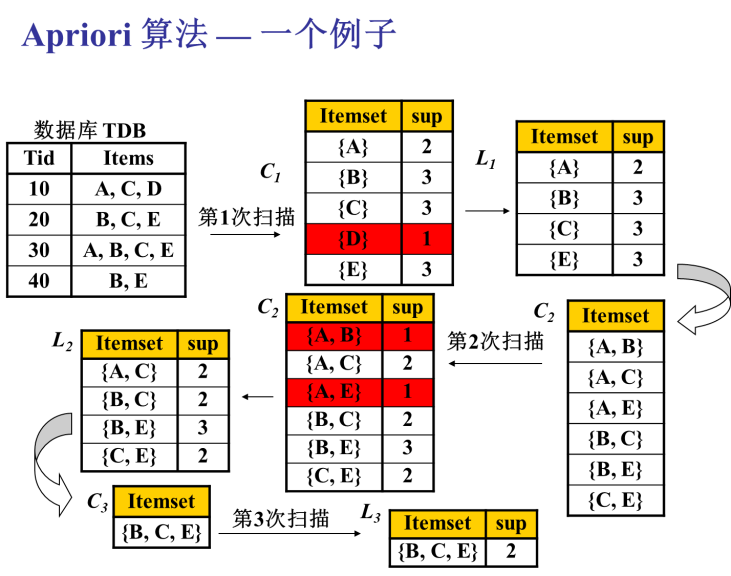






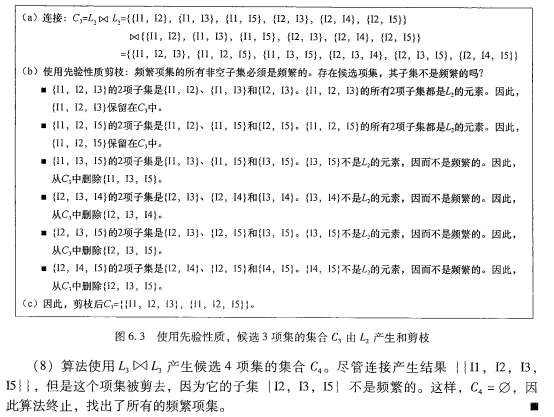
1. **相关例题**

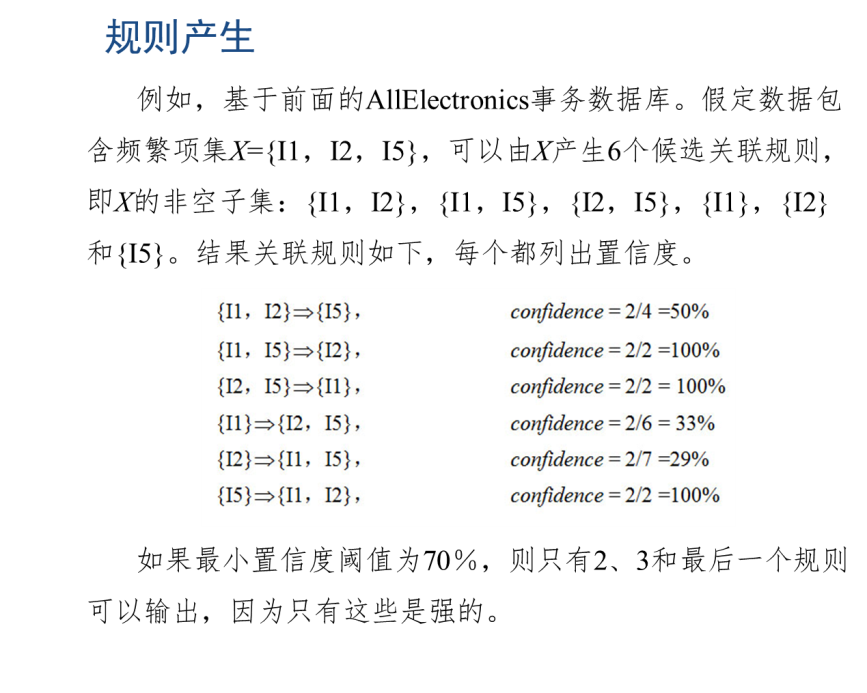


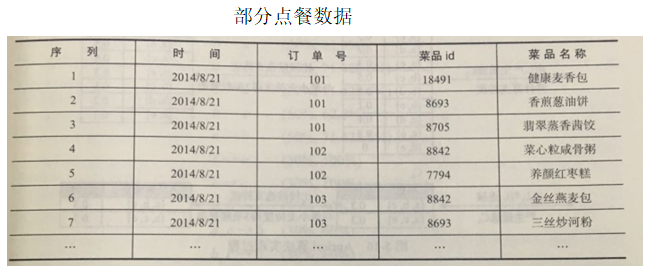


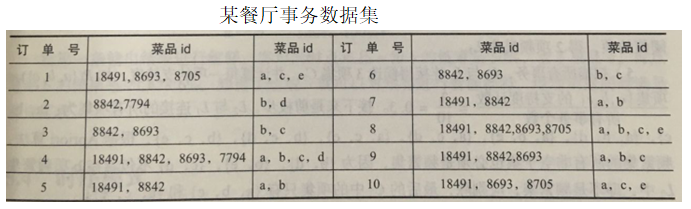


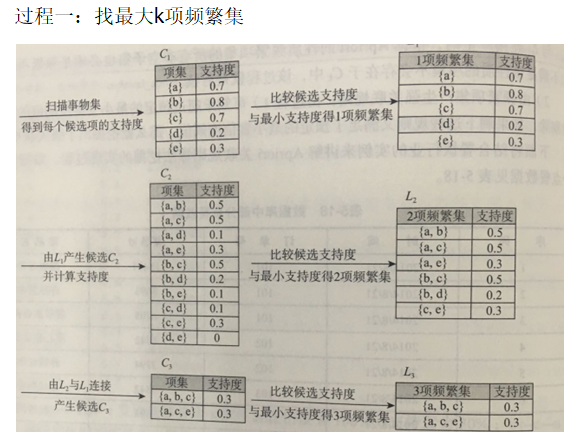


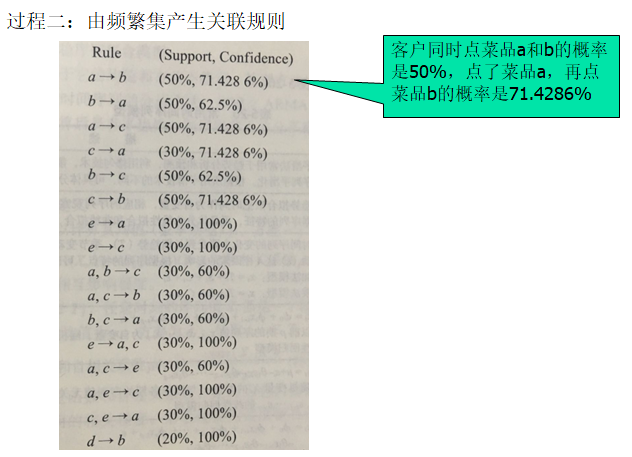




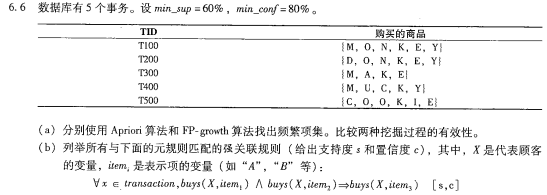




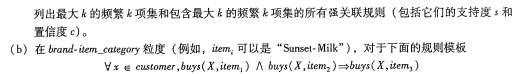


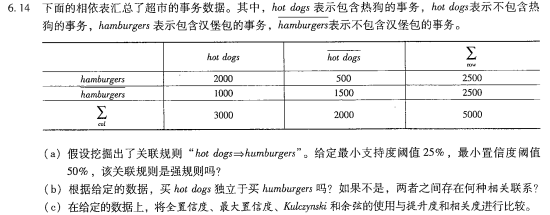


1. **书本对应习题**





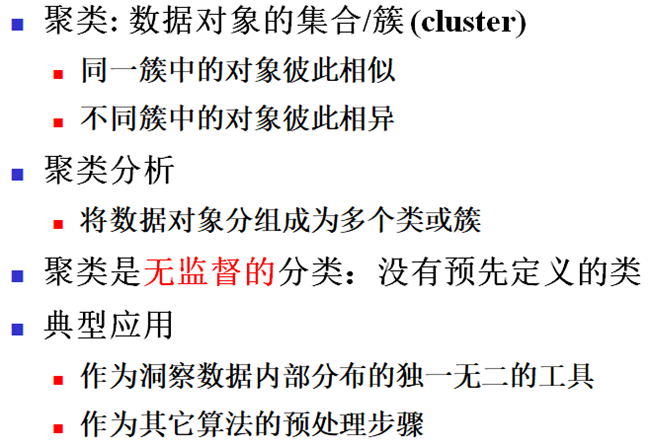


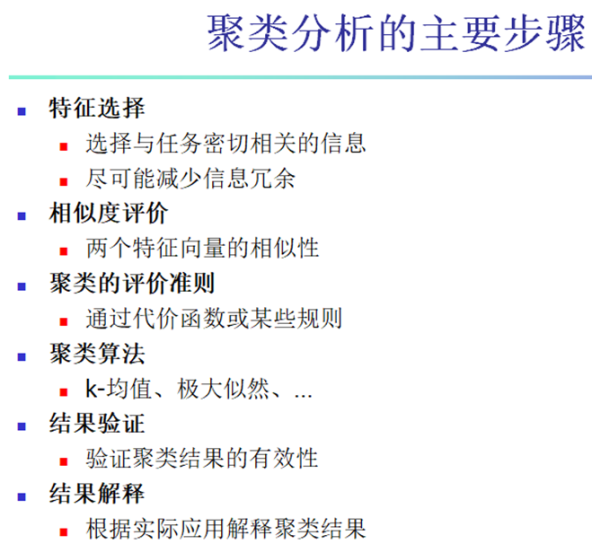


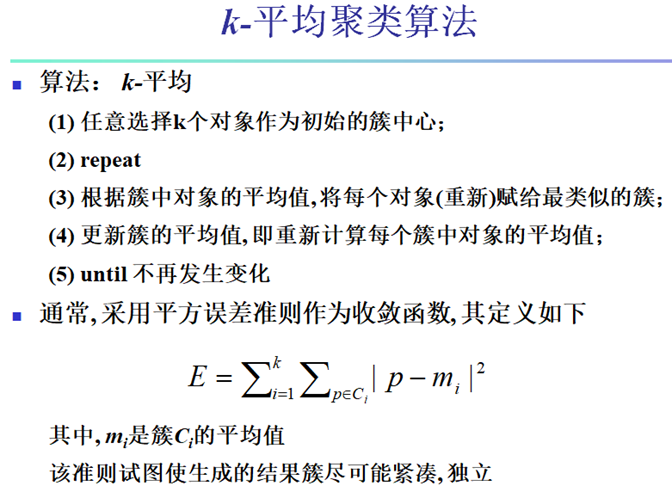
1. **聚类**
2. **考点**

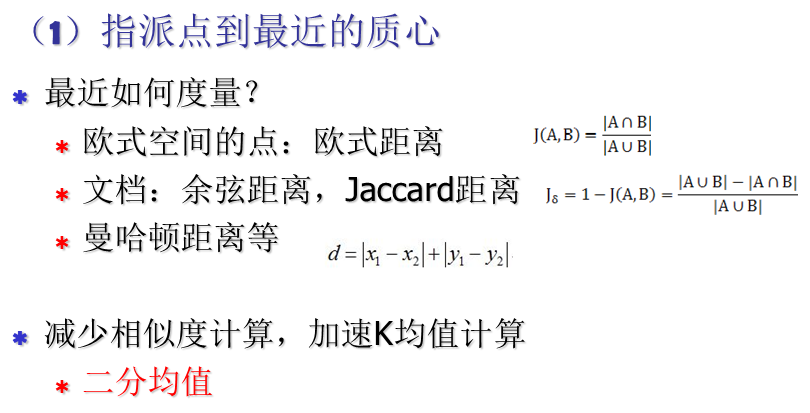
K均值算法

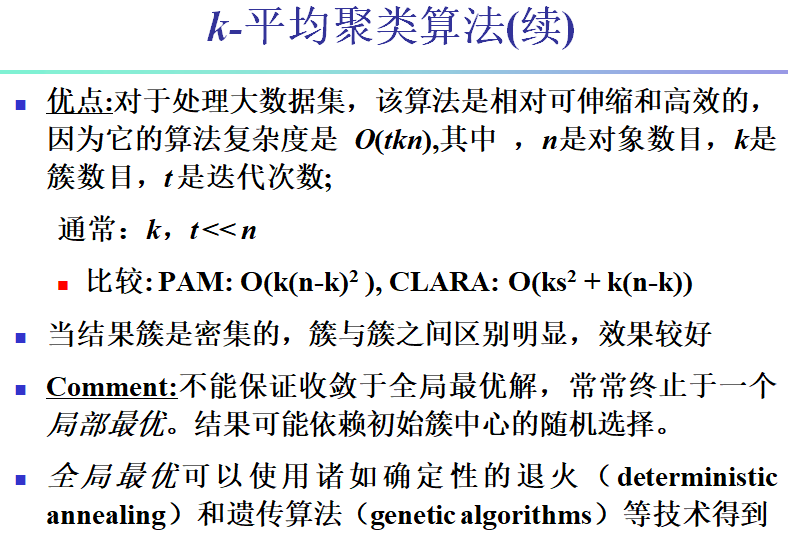
1. **相关概念**

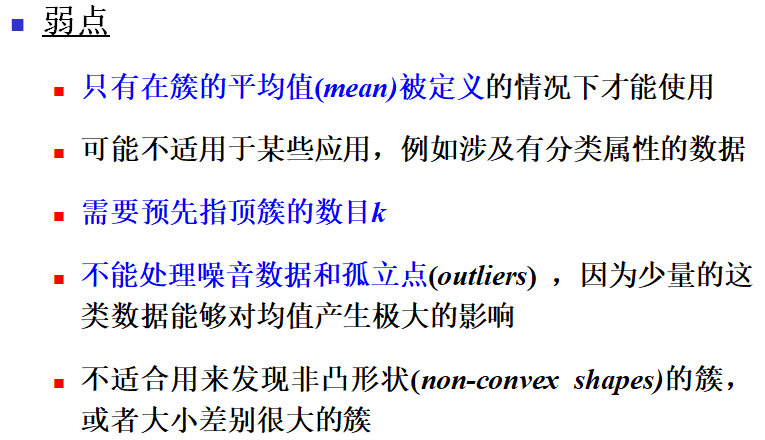




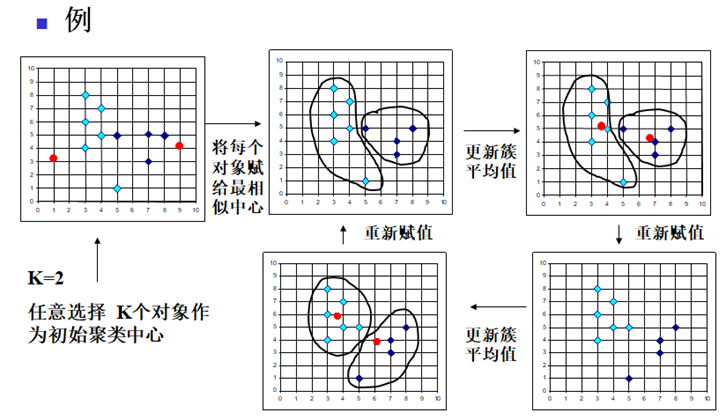




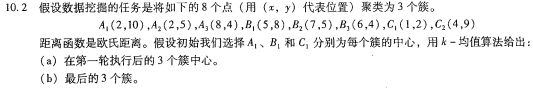


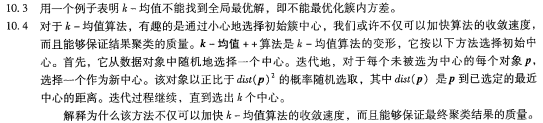


1. **相关例题**



1. **课本对应习题**

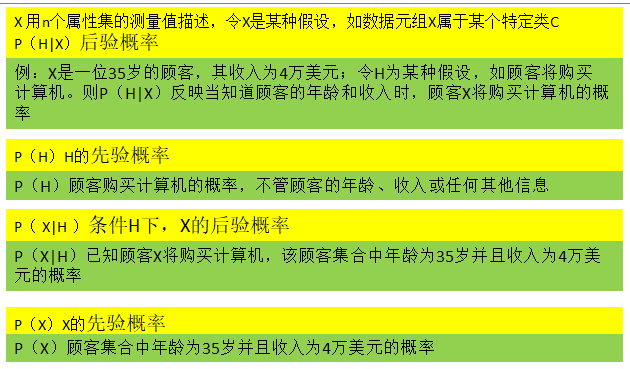


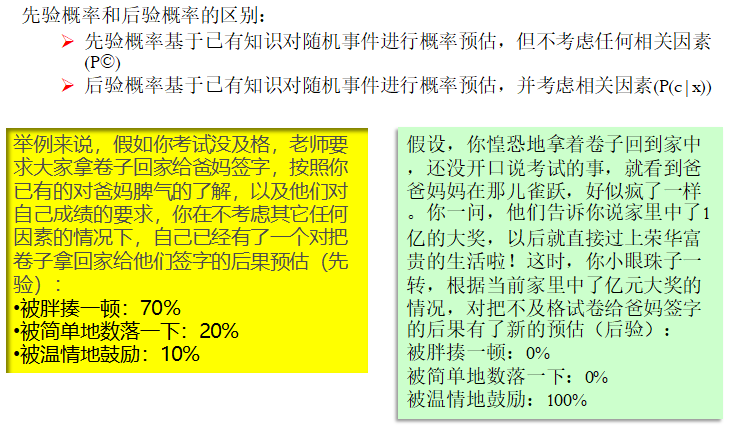


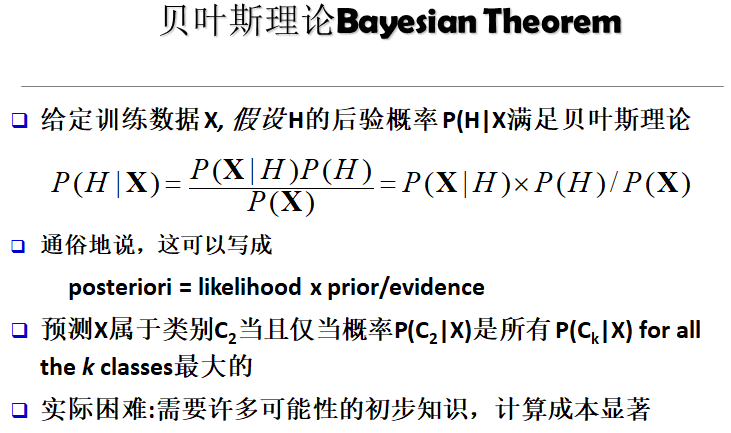
1. **分类**
2. **考点**

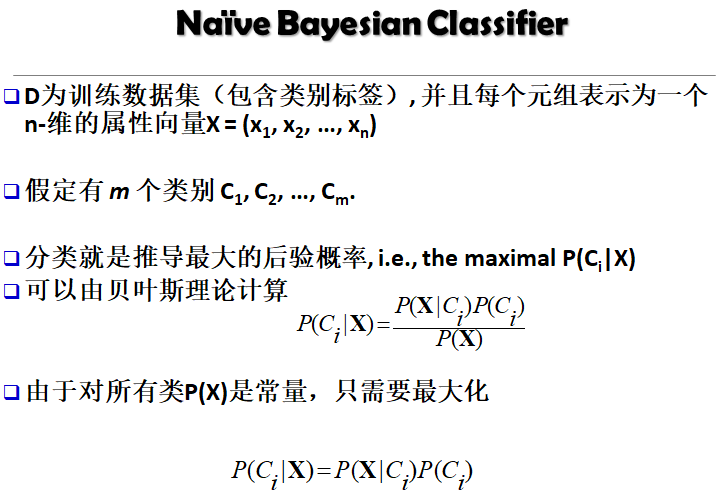
朴素贝叶斯算法

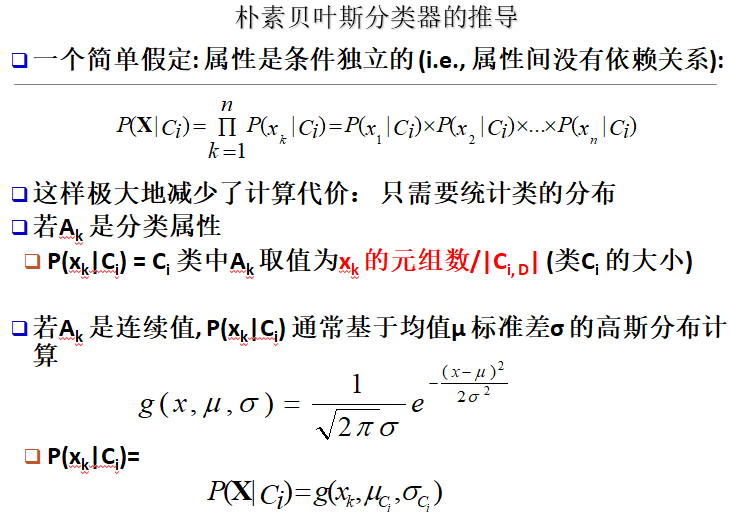
1. **相关概念**



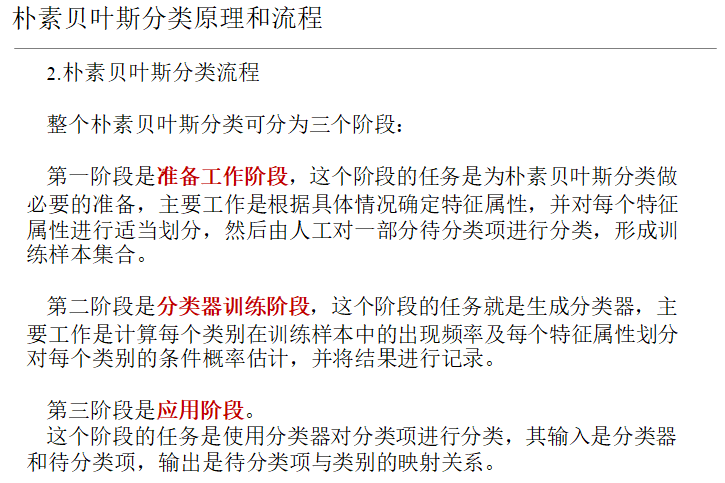


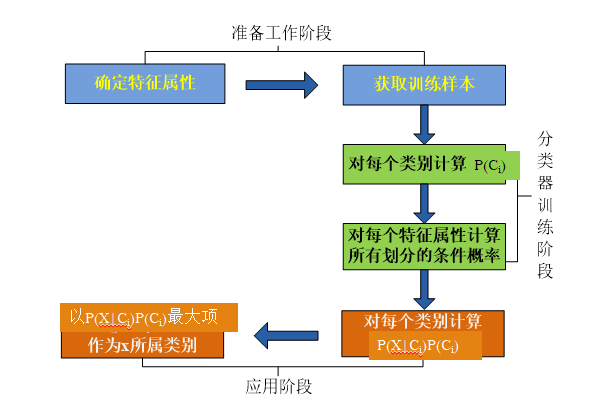






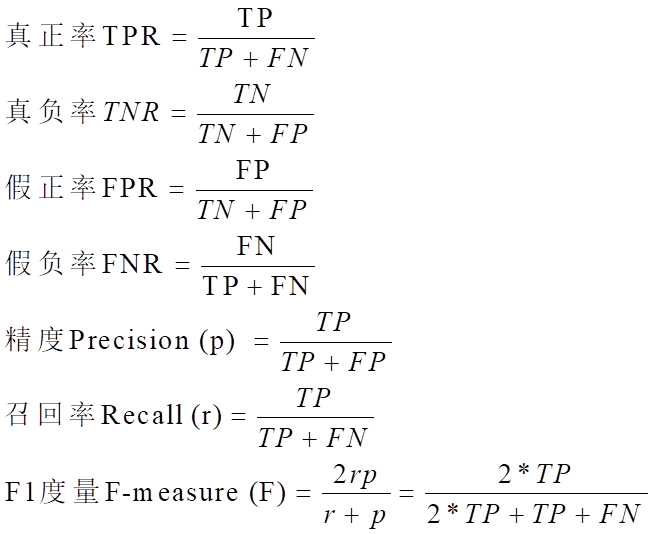






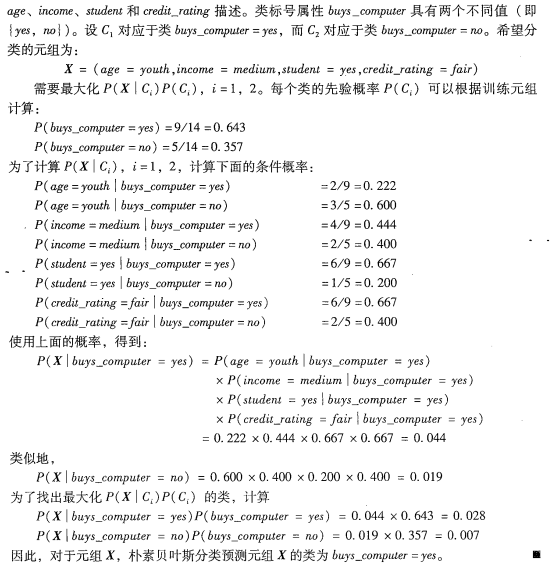
分类器评价指标：（使用混淆矩阵）



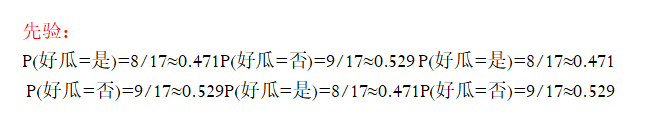


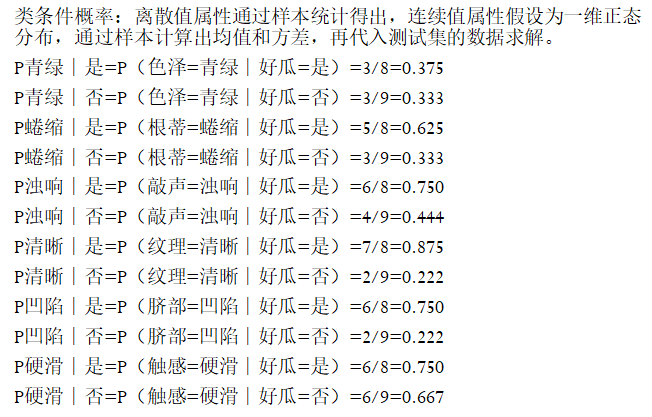
1. **相关例题**

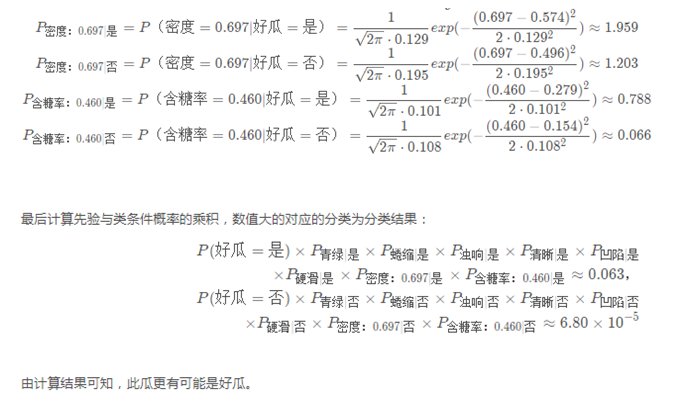


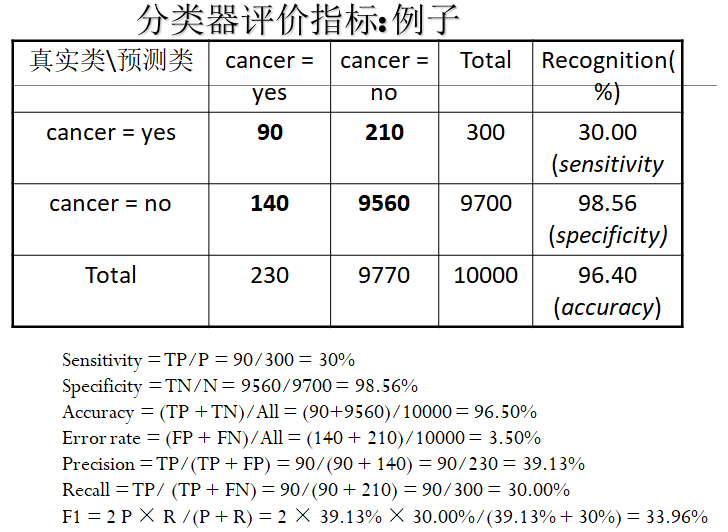












1. **书本对应习题**

