数据挖掘 2024 试题回忆

神秘人

November 2024

1 形式

每年都是五道大题，每道题给一个小数据集，然后在小数据集上手动运行学过的几个数据处理算法，没有理解难度但是考试过程中需要大量敲击计算器。不同数据集只是结果不同，运行思路完全一致。

2 考试范围

考试范围可能会每年变换，但是刘老师都会提前通知。2024 年考察算法范围：

1. Apriori 关联规则挖掘（包括多层关联规则）
2. GSP、PrefixSpan
3. 决策树、贝叶斯分类、k-NN 分类
4. k-means、层次聚类、dbscan
5. pagerank

3 关联规则挖掘

给定五个项目集，两问：

(1) 找出其中的频繁项集。

(2) 写出由长度为 3 的频繁项集生成的关联规则。

4 朴素贝叶斯

给定连续 14 天的天气，湿度，气温等指标情况（均为离散变量）和当天是否打球（标签），现给定一个新的日子的天气，湿度，气温情况，判断当天是否会去打球。(与 PPT 上贝叶斯例题基本一致)

5 决策树分类

给定若干个实体以及他们各自的各类属性 a_1, a_2, a_3 和实际标签 y。其中属性 a_1, a_2 为各自只有两三个候选值的离散属性，a_3 为连续型变量，按照 a_3 划分时需要计算所有可能划分的信息增益。

(1) 计算数据集对于标签 y 的熵。

(2) 计算按照属性 a_1, a_2 划分的信息增益大小。

(3) 计算按照按照属性 a_3 划分的信息增益。

(4) 确定决策树第一次划分应该选择 a_1, a_2, a_3 中的哪个属性。

6 K-means 聚类

给定若干样本点 (x, y) 组成的数据集，约定 $K = 3$，初始点是其中三个点，写出 K-means 聚类过程和最终结果。(经过三次聚类之后收敛，计算器计算量较大)。

7 K-NN 分类

给定若干样本的属性值 x 和真实标签 y，使用距离加权算法计算新样本 $\hat{x}$ 在 K=1,3,5,9 时的标签。