# 课堂练习

2021.5.14

杜逆索

| TID | 项目集 |
| --- | --- |
| 1 | a,c,d,f,g |
| 2 | a,b,d,e,g |
| 3 | a,d,f,g |
| 4 | b,d,f |
| 5 | e,f,g |
| 6 | a,b,c,d,g |
| 7 | a,b,e,g |

1. 给定上表所示的一个事物数据库，写出Apriori算法生成频繁项目集的过程(假定最小支持度=0.5)。

| 样本 | Ca+浓度 | Mg+浓度 | Na+浓度 | Cl−浓度 | 类型 |
|---|---|---|---|---|---|
| A | 0.2 | 0.5 | 0.1 | 0.1 | 冰川水 |
| B | 0.4 | 0.3 | 0.4 | 0.3 | 湖泊水 |
| C | 0.3 | 0.4 | 0.6 | 0.3 | 冰川水 |
| D | 0.2 | 0.6 | 0.2 | 0.1 | 冰川水 |
| E | 0.5 | 0.5 | 0.1 | 0 | 湖泊水 |
| F | 0.3 | 0.3 | 0.4 | 0.4 | 湖泊水 |
| G | 0.3 | 0.3 | 0.3 | 0.2 | ? |
| H | 0.1 | 0.5 | 0.2 | 0.2 | ? |

2. 使用K-邻近法对两个未知类型的样本进行分类(冰川水或者湖泊水)，本例我们使用K=3，即选择最近的3个邻居。

| Ca+浓度 | Mg+浓度 | Na+浓度 | Cl-浓度 | 类型 |
|---|---|---|---|---|
| 低 | 高 | 高 | 高 | 冰川水 |
| 高 | 低 | 高 | 高 | 冰川水 |
| 低 | 高 | 低 | 低 | 冰川水 |
| 高 | 高 | 低 | 低 | 冰川水 |
| 低 | 低 | 低 | 低 | 湖泊水 |
| 高 | 低 | 低 | 低 | 湖泊水 |
| 低 | 高 | 高 | 低 | 湖泊水 |
| 高 | 低 | 高 | 低 | 湖泊水 |
| 低 | 高 | 高 | 低 | ? |
| 高 | 高 | 低 | 高 | ? |

3. 使用ID3决策树算法对两个未知类型的样本进行分类。

（4）请首先写出D1和D3进行交集合并的语句，然后写出各语句执行之后的结果。

1）D1 = pd.DataFrame({'id':[801, 802, 803,804, 805, 806, 807, 808, 809, 810], 'name':['Ansel', 'Wang', 'Jessica', 'Sak','Liu', 'John', 'TT','Walter','Andrew','Song']})

2）D2 = pd.DataFrame({'id':[803, 804, 808,901], 'save': [3000, 500, 1200, 8800]})

3）D3 = pd.DataFrame({‘id2':[803, 804, 808,901], 'save': [3000, 500, 1200, 8800]})

4）a = pd.merge(D1, D2, on='id')

5）b = pd.merge(D1, D2, on='id', how='outer')

（5）要求合并D1和D2的数据并且为并集，请首先写出合并执行代码，然后再写出各语句执行之后的结果。

1）D1 = pd.DataFrame({'id':[801, 802, 803,804, 805, 806, 807, 808, 809, 810], 'name':['Ansel', 'Wang', 'Jessica', 'Sak','Liu', 'John', 'TT','Walter','Andrew','Song']})

2）D2 = pd.DataFrame({'save': [3000, 500, 1200, 8800]})

（6）概述k均值和k中心点算法的优缺点。

（7）简述数据清理、数据变换、刷新的概念。