数据挖掘第二次作业

姓名： 黄 姗

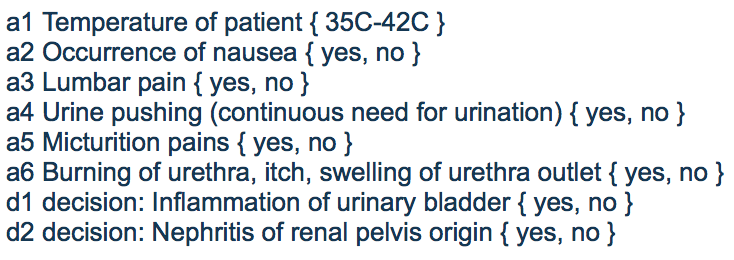
学号：2120150995

# 数据介绍

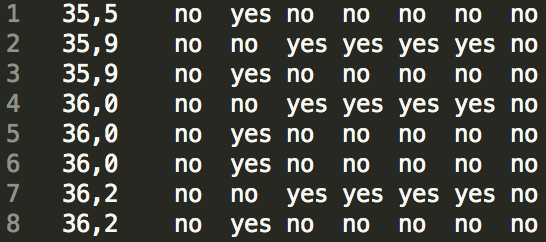
UCI “急性炎症”数据集，网页链接为<http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations>

简单描述：

数据共有120条，每行包括8个字段，以制表符(\t)间隔，分别记作a1-a6和d1-d2，其中前六个为症状检查属性，后两个为炎症检查结果。A1为温度属性，取值为35-42摄氏度，其他属性取值均为yes和no。详细介绍如下：



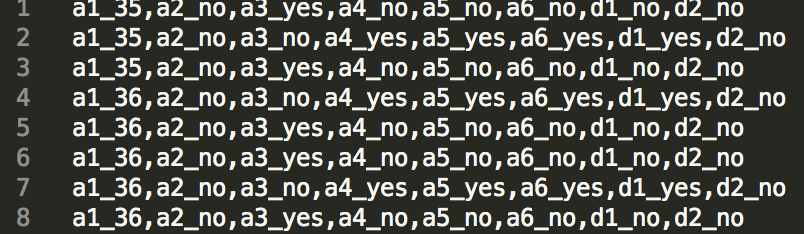
举例如下



# 实验过程

1、数据预处理

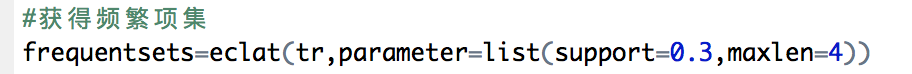
首先，需要把连续的温度属性处理为离散属性值，这里，我们采取的方法为取温度值的整数部分，离散为{35，36，37，38，39，40，41，42}，其他属性值均已是离散值，但为了接下来的处理，我们加上列名，处理为a1\_35,a2\_yes,a3\_no类似的形式。处理程序为presolve.py，处理后的数据举例为：



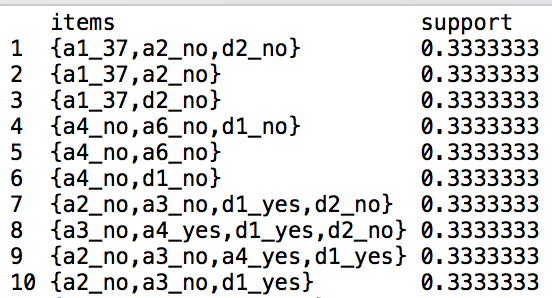
数据预处理采用python处理。结果保存为preDiagnosis.data中。

# 获取频繁项集

接下来的数据处理，我们使用R语言，使用如下函数可以获得频繁项集。

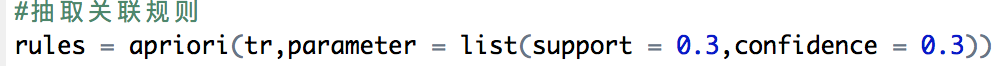


可以设置支持度、最大项长度，我们将结果保存到frequencySet.txt中，如下图所示：

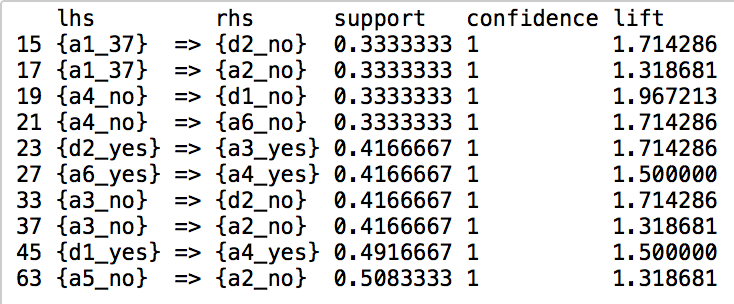


# 关联规则

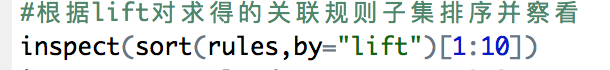
接下来采用Aprioris算法来的到关联规则，可以设置支持度和置信度参数。S



计算的结果如下：



结果保存到rules.txt中，此外，还可以对结果按照置信度、支持度等进行排序选择。这里也选择Lift指标进行评价，并保留了前10个规则。



# 可视化

1. 关联规则可视化如下图1所示，横坐标表示支持度，纵坐标表示置信度，颜色表示lift值：
2. 关联规则可视化如下图2所示，横坐标表示支持度，纵坐标表示lift值，颜色表示置信度：
3. 泡泡图如图3所示：
4. 平行坐标图如图4所示：

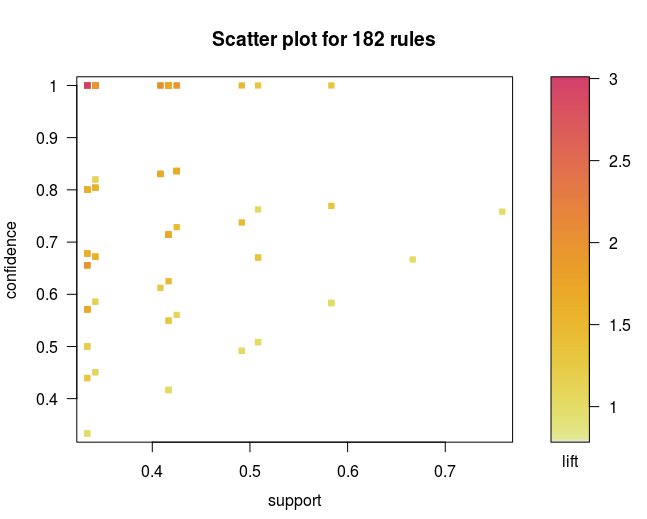


图1

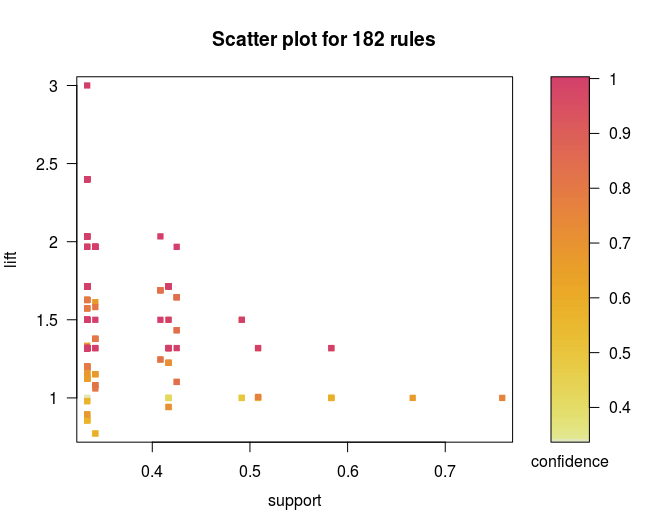


图2

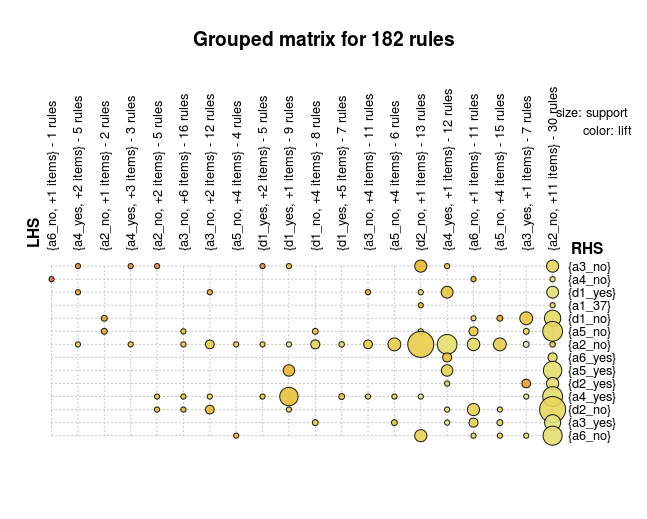


图3

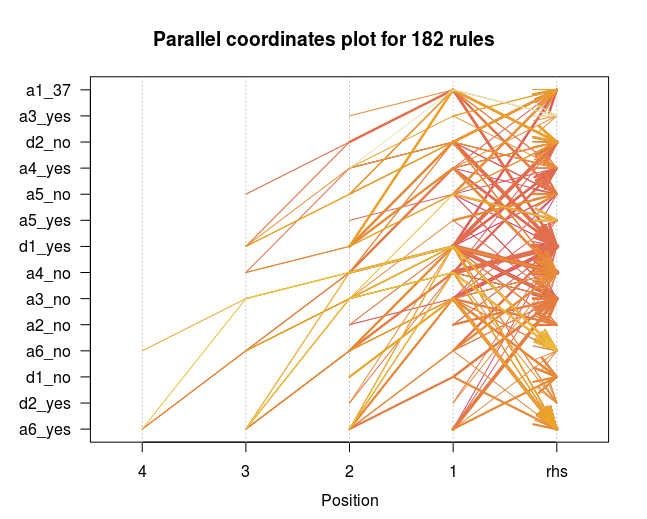


图4