数据挖掘上机报告（实验2）

1. **实验内容**

本实验的目的是利用Apriori算法进行关联规则分析。通过对美国众议院投票数据集进行分析，挖掘出投票选项之间的关联关系。具体来说，数据集包含了每一位议员在多个议题上的投票情况（“y”表示赞成，“n”表示反对），以及每一议题的名称。通过Apriori算法，可以从这些投票数据中提取出频繁项集，并根据这些项集生成关联规则，帮助理解哪些议题的投票可能具有某种关联性。

1. **分析及设计**

实验使用的数据集是“house-votes-84.data”，该数据集包含了84个议题，每个议题的投票记录包含在多个数据项中。每个记录有一列表示投票者的党派，后面的列表示每个议题的投票结果。投票结果为“y”、“n”或者“?”。我们需要先处理缺失值，并将“y”与“n”转换为数字1和-1，以便后续的分析。

为了实现关联规则分析，设计方案包括以下几个步骤：

数据预处理

· 读取数据集并填充缺失值（“?”替换为0）。

· 将每个投票结果中的“y”转换为1，“n”转换为-1。

· 确定哪些列代表各议题，以及如何将这些列转换为适合进行关联分析的数据格式。

**生成频繁项集**：

使用Apriori算法从投票数据中提取频繁项集。Apriori算法是一种经典的关联规则挖掘算法，核心思想是通过逐层生成候选项集并通过支持度进行筛选，从而找出频繁项集。

**生成关联规则**：

从频繁项集中生成关联规则，通过设定最小置信度值来确保关联规则的可靠性。

**输出结果并分析**：

输出生成的频繁项集及其关联规则，并对结果进行分析，寻找有趣的模式和关系。

1. **详细实现**

首先，我们使用pandas库读取数据集，并进行预处理。数据中的缺失值被替换为0，投票结果被替换为数字1和-1，方便后续的处理。

import pandas as pd

from efficient\_apriori import apriori

def read\_voting\_data(file\_path):

raw\_data = pd.read\_table(file\_path, sep=",", header=None, na\_values="?")

print(raw\_data)

column\_names = [

"party", "handicapped-infants", "water-project-cost-sharing",

"adoption-of-the-budget-resolution", "physician-fee-freeze",

"el-salvador-aid", "religious-groups-in-schools",

"anti-satellite-test-ban", "aid-to-nicaraguan-contras", "mx-missile",

"immigration", "synfuels-corporation-cutback", "education-spending",

"superfund-right-to-sue", "crime", "duty-free-exports",

"export-administration-act-south-africa"

]

raw\_data = raw\_data.fillna(0) # 填充缺失值

raw\_data.columns = column\_names

processed\_data = raw\_data.replace({"y": 1, "n": -1}) # 转换投票结果

return processed\_data

接着，我们将每一行投票数据作为一个事务，每个事务是一个包含议题投票结果的元组。然后使用efficient\_apriori库的apriori函数来提取频繁项集和生成关联规则。

voting\_data = read\_voting\_data(r"./data/house-votes-84.data")

transactions = []

for i in range(voting\_data.shape[0]):

transactions.append(tuple([str(voting\_data.values[i][j]) for j in range(voting\_data.shape[1])]))

item\_sets, association\_rules = apriori(transactions, min\_support=0.3, min\_confidence=0.9)

# 输出关联规则

for rule in association\_rules:

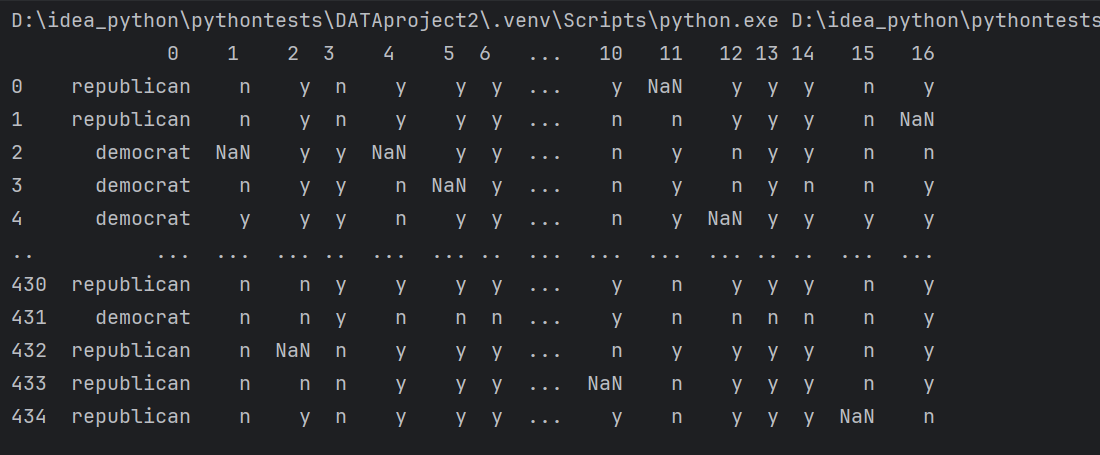
print(rule)

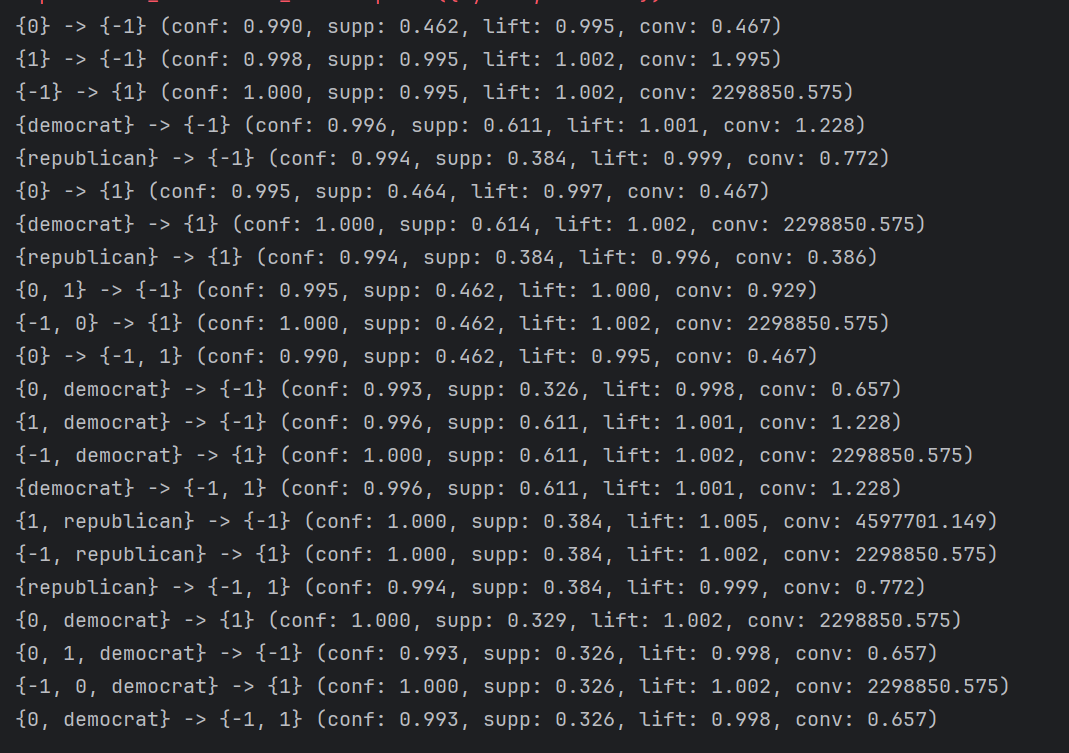
关联规则的输出包括两个部分：

**频繁项集**：表示某些议题的投票结果经常一起出现。

**关联规则**：表示如果某些议题的投票结果为特定值，那么其他议题的投票结果也有很高的概率为某个值。

1. **实验结果**





1. **心得体会**

通过本次实验，我对Apriori算法的实现和应用有了更深的理解。在实验过程中，最让我欣赏的是如何通过设置支持度和置信度来控制生成规则的质量，使得结果既不冗余也不失效。在调试过程中，我发现数据预处理非常关键，只有准确地处理缺失值和数据转换，才能得到正确的关联规则。