第四问

第四问让我们研究不同类比成分之间的关联关系，这里的类别我理解为表一中的四种属性，即

(1) 按类型，成分之间的关联关系

(2) 按纹饰，成分之间的关联关系

(3) 按颜色，成分之间的关联关系

(4) 按风化程度，成分之间的关联关系

既然是研究不同成分之间的关联关系，因为用关联分析法最为适合，关联分析法的原理和介绍如下：

关联规则简介

关联规则挖掘可以让我们从数据集中发现项与项之间的关系，它在我们的生活中有很多应用场景，“购物篮分析”就是一个常见的场景，这个场景可以从消费者交易记录中发掘商品与商品之间的关联关系，进而通过商品捆绑销售或者相关推荐的方式带来更多的销售量。

1. 搞懂关联规则中的几个重要概念：支持度、置信度、提升度
2. Apriori 算法的工作原理
3. 在实际工作中，我们该如何进行关联规则挖掘

关联规则中重要的概念

举一个超市购物的例子，下面是几名客户购买的商品列表

订单编号	购买商品
1	牛奶、面包、尿布
2	可乐、面包、尿布、啤酒
3	牛奶、尿布、啤酒、鸡蛋
4	面包、牛奶、尿布、啤酒
5	面包、牛奶、尿布、可乐

支持度

支持度是个百分比，它指的是某个商品组合出现的次数与总次数之间的比例。支持度越高，代表这个组合出现的频率越大。

我们看啤酒出现了 3 次，那么 5 笔订单中啤酒的支持度是 3/5=0.6。同理，尿布出现了 5 次，那么尿布的支持度是 5/5=1。尿布和啤酒同时出现的支持度是 3/6=0.6。

置信度

它指的就是当你购买了商品 A，会有多大的概率购买商品 B。

我们可以看上面的商品，购买尿布的同时又购买啤酒的订单数是 3，购买啤酒的订单数是 3，那么（尿布->啤酒）置信度= 3/3=1。

再看购买了啤酒同时购买尿布的订单数是 3，购买尿布的订单数是 5，那么（啤酒->尿布）置信度=3/5=0.6。

提升度(Lift):

表示含有商品 A 的条件下，同时含有商品 B 的概率，与商品 B 总体发生的概率之比。
Lift(A→B)= P(B|A) / P(B)

Apriori 的工作原理

Apriori 算法其实就是查找频繁项集 (frequent itemset) 的过程，所以我们需要先了解是频繁项集。

频繁项集就是支持度大于等于最小支持度阈值的项集，所以小于最小值支持度的项目就是非频繁项集，而大于等于最小支持度的项集就是频繁项集。

下面我们来举个栗子：

假设我随机指定最小支持度是 0.2。首先，我们先计算单个商品的支持度：

购买商品	支持度
牛奶	4/5
面包	4/5
尿布	5/5
可乐	2/5
啤酒	3/5
鸡蛋	1/5

因为最小支持度是 0.2，所以你能看到商品 鸡蛋 是不符合最小支持度的，不属于频繁项集，于是经过筛选商品的频繁项集如下：

购买商品	支持度
牛奶	4/5
面包	4/5
尿布	5/5
可乐	2/5
啤酒	3/5

在这个基础上，我们将商品两两组合，得到两个商品的支持度：

购买商品	支持度
牛奶、面包	3/5
牛奶、尿布	4/5
牛奶、可乐	1/5
牛奶、啤酒	2/5
面包、尿布	4/5
面包、可乐	2/5
面包、啤酒	2/5
尿布、可乐	2/5
尿布、啤酒	3/5
可乐、啤酒	1/5

筛选大于最小支持度（0.2）的数据后

购买商品	支持度
牛奶、面包	3/5
牛奶、尿布	4/5
牛奶、啤酒	2/5
面包、尿布	4/5
面包、可乐	2/5
面包、啤酒	2/5
尿布、可乐	2/5
尿布、啤酒	3/5

在这个基础上，我们再将商品三个组合，得到三个商品的支持度：

购买商品	支持度
牛奶、面包、尿布	3/5
牛奶、面包、可乐	1/5
牛奶、面包、啤酒	1/5
面包、尿布、可乐	1/5
面包、尿布、啤酒	2/5
尿布、可乐、啤酒	1/5

筛选大于最小支持度（0.2）的数据后

购买商品	支持度
牛奶、面包、尿布	3/5
面包、尿布、啤酒	2/5

在这个基础上，我们将商品四个组合，得到四个商品的支持度：

购买商品	支持度
牛奶、面包、尿布、可乐	1/5
牛奶、面包、尿布、啤酒	1/5
面包、尿布、可乐、啤酒	1/5

再次筛选大于最小支持度（0.2）数据的话，就全删除了，那么，此时算法结束，上一次的结果就是我们要找的频繁项，也就是{牛奶、面包、尿布}、{面包、尿布、啤酒}。

总结一下上述 Apriori 算法过程：

1. K=1，计算 K 项集的支持度
2. 筛选掉小于最小支持度的项集
3. 如果项集为空，则对应 K-1 项集的结果为最终结果
4. 否则 K=K+1，重复 1-3 步
5. 我们可以看到 Apriori 在计算的过程中有以下几个缺点：
6. 可能产生大量的候选集。因为采用排列组合的方式，把可能的项集都组合出来了
7. 每次计算都需要重新扫描数据集，来计算每个项集的支持度

化学成分分析

因此，利用上述关联分析法，基于某种类型的分布下，对组成成分，例如二氧化硅，氧化钠等等一步一步的进行求其支持度，置信度等，根据这些概率的大小，我们可以得到在某种类别下，那些成分之间的关联性较强，哪些较弱，比较其差异性。

这里只是提供一个思路，具体解题方式根据算法一步一步求解，利用 python 、pandas 数据处理库等工具会较为方便。