作业三

关联规则

关联规则通常被零售商用于市场篮子分析（MBA），以了解其顾客的购买行为。这些信息随后可以用于许多不同的目的，如产品交叉销售和提升销售、促销活动、忠诚计划、店铺设计、折扣计划等。

项集的评估：一旦找到数据集的频繁项集，您需要从中选择一部分作为您的推荐。用于衡量选择推荐规则的显著性和兴趣的常见度量标准有：

1. **置信度**（表示为 $\mathit{conf}(A \to B)$）：置信度被定义为在购物篮中已经包含 A 的情况下，B 发生的概率：

$$\mathit{conf}(A \to B) = Pr(B|A),$$

   其中，$Pr(B|A)$ 是在已经存在项集 A 的情况下找到项集 B 的条件概率。

2. **提升度**（表示为 $\mathit{lift}(A \to B)$）：提升度度量的是相对于“A 和 B 相互独立”的情况，给定“A 发生”，能让 B 同时发生的概率提高多少：

$$\mathit{lift}(A \to B) = \frac{\mathit{conf}(A \to B)}{S(B)},$$

   其中，$S(B) = \frac{B\text{ 发生的次数}}{\text{数据集中所有事件（篮子）的总数}}$。

3. **确信度**（表示为 $\mathit{conv}(A \to B)$）：确信度比较了“理论上，当它们相互独立时，A 出现而 B 没有出现的概率”与“实际上，A 出现而 B 没有出现的频率”：

$$\mathit{conv}(A \to B) = \frac{1 - S(B)}{1 - \mathit{conf}(A \to B)}.$$

(a) [15 分]

使用置信度的一个缺点是它忽略了 $Pr(B)$。为什么这是一个缺点？请解释一下提升度和确信度为什么不受这个缺点的影响。

(b) [15 分]

如果 度量$(A \to B)$ = 度量$(B \to A)$，则称该度量是对称的。在前文介绍的几种度量中，有哪些是对称的？对于每个度量，请证明该度量是对称的，或者提供一个反例表明该度量不是对称的。

(c) [20 分]

完美蕴含是指相关的条件概率为 1。例如，$P(B|A) = 1$，则称 $A \to B$ 为完美蕴含。如果一个度量对于所有完美蕴含的情况，都能达到其最大可实现值，则称该度量是可取的。这使得我们很容易识别出最佳规则。在上述度量中，哪些是可取的？您可以忽略 0/0 的情况，但不能忽略其他无穷大的情况（即，最大值也可以为无穷大）。同时，您可以通过一个例子来简单解释。

在产品推荐中的应用：向现有客户销售额外产品或服务的行为被称为交叉销售。提供产品推荐是在线零售商经常使用的交叉销售的一个例子。一个简单的方法是通过推荐那些经常被客户一同浏览的产品来进行产品推荐。

假设我们想基于客户已经在线浏览过的产品向其推荐新产品。使用 A-priori 算法编写一个程序，以找到经常被同时浏览的产品。将支持度设定为 $s = 100$（即，一个产品对至少需要被同时浏览 100 次才被视为频繁），并找到大小为 2 和 3 的项集。

使用 *browsing.txt* 中的在线浏览行为数据集。每一行代表一个客户的浏览会话。在每一行中，每个 8 个字符的字符串表示在该会话期间浏览的一个产品的 ID。这些产品 ID 由空格分隔。有些行包含重复的产品。删除或忽略重复产品不应影响您的结果。

(d) [20 分]

识别那些支持度至少为 100 的产品对 (X, Y)。对于所有满足条件的产品对，计算相应关联规则（$X \Rightarrow Y, Y \Rightarrow X$）的置信度分数。按照置信度分数的降序对规则进行排序，并列出排序后的前 5 条规则。

(e) [30 分]

识别那些支持度至少为 100 的产品三元组 (X, Y, Z)。对于所有满足条件的产品三元组，计算相应关联规则（$(X, Y) \Rightarrow Z, (X, Z) \Rightarrow Y, (Y, Z) \Rightarrow X$）的置

信度分数。按照置信度分数的降序对规则进行排序，并列出排序后的前 5 条规则。