DSCI 6001P 数据科学基础

作业 4

提交截止日期：12.4 号晚上 24 点之前

提交方式：电子版发送至 毕书显 (stanbi@mail.ustc.edu.cn)

1. 考虑下表的购物篮事务：

事务 ID	购买项
1	{牛奶，啤酒，尿布}
2	{面包，黄油，牛奶}
3	{牛奶，尿布，饼干}
4	{面包，黄油，啤酒}
5	{啤酒，饼干，尿布}
6	{牛奶，尿布，面包，黄油}
7	{面包，黄油，尿布}
8	{啤酒，尿布}
9	{牛奶，尿布，面包，啤酒}
10	{啤酒，饼干}

(1) 从这些数据中，能够提取出的关联规则的最大数量是多少（包括零支持度的规则）？

(2) 能够提取的频繁项集的最大长度是多少（假定最小支持度>0）？

(3) 写出从该数据及中能够提取的 3-项集的最大数量的表达式。

(4) 找出具有最大支持度的项集（长度为 2 或更大）。

2. 数据库有 5 个事务。设 min_sup =60% ， min_conf = 80%。

TID	购买的商品
T100	|M, O, N, K, E, Y|
T200	|D, O, N, K, E, Y|
T300	|M, A, K, E|
T400	|M, U, C, K, Y|
T500	|C, O, O, K, I, E|

(a)分别使用 Apriori 算法和 FP-growth 算法找出频繁项集。比较两种挖掘过程的有效性。

(b)列举所有与下面的元规则匹配的**强**关联规则(给出支持度 s 和置信度 c)，其中，X 是代表顾客的变量，$item_i$是表示项的变量(如“A”，“B”等)：

$$\forall x \in \text{transaction}, \text{buys}(X, \text{item}_1) \wedge \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3) \ [s, c]$$

3. 设计一种方法，对**无限的**数据流进行有效的朴素贝叶斯分类（即只能扫描数据流一次）。如果想发现这种分类模式的演变（例如，将当前的分类模式与较早的模式进行比较，如与一周以前的模式相比），你有何修改建议？

4. 假设一个布隆过滤器的容量为 8×10^9 位，集合中有 1×10^9 个元素。如果使用 3 个哈希函数，试计算误判率。如果使用 4 个哈希函数呢？

5. 假定全集 A 有 n 个元素，随机从中抽取出两个子集 A_1 和 A_2，且每个子集都有 m 个元素，求 A_1 和 A_2 两个集合的期望相似度。

6. 给定输入流< b, a, c, a, d, e, a, f, a, d>, 计数器个数 k = 3。请逐步写出 Misra- Gries 算法执行的结果。

7. 给定数据流< 4, 1, 3, 5, 1, 3, 2, 6, 7, 0, 9 >, 若哈希函数形如 h(x)= (ax + b) mod 8, 其中 a 和 b 是任意给定的常数。假设给定如下哈希函数：
   (1) h(x) = (3x + 2) mod 8；
   (2) h(x) = (7x + 5) mod 8；
   (3) h(x) = (5x + 3) mod 8。
   请利用 Count-Min sketch 算法估计频繁项。

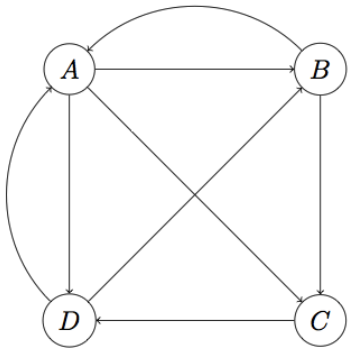
8. 在不考虑 damping factor 的情况下，对于连通图 PageRank 的定义如下：

   设 u 是有向图 G 中的一个顶点，$N(u)$ 表示图 G 中指向顶点 u 的顶点集合。则图顶点 u 的 **PageRank** 值可以计算为

   $$PR(u) = \sum_{v \in N(u)} \frac{PR(v)}{N(v)}, \tag{7.30}$$

   其中 $PR(u)$ 表示顶点 u 的 **PageRank** 值。

   若存在 A、B、C、D 四个网站，其链接结构如下图所示，计算器 PageRank 值（有向图，注意箭头）。



9. 给定一个转移矩阵 P 和状态向量 π：

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix} \text{和} \pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T$$

   a. 计算 πP, πP^2 和 πP^3 。
   b. 证明 πP^n 的结果接近一个常数向量。
   c. 给定任一个转移矩阵 P， 满足矩阵中每个元素为概率值，且每一行的元素和为 1。证明 P 和 $\frac{1}{n}((n-1)I + P)$ 有相同的平稳分布，其中 I 表示单位矩阵。