DSCI 6001P 数据科学基础

作业 4 - 聚类、哈希、数据流

提交截止日期：12.12 号晚上 24 点之前

提交方式：电子版发送至 汪远 (wy1001@mail.ustc.edu.cn)

邮件主题和 PDF 命名格式：HW4-姓名-学号，如 HW4-张三-SA24123123

1. K-mediods 算法描述：

a) 首先随机选取一组聚类样本作为中心点集

b) 每个中心点对应一个簇

c) 计算各样本点到各个中心点的距离（如欧几里德距离），将样本点放入距离中心点最短的那个簇中

d) 计算各簇中，距簇内各样本点距离的绝度误差最小的点，作为新的中心点

e) 如果新的中心点集与原中心点集相同，算法终止；如果新的中心点集与原中心点集不完全相同，返回 b)

问题：

a) 阐述 K-mediods 算法和 K-means 算法相同的缺陷

b) 阐述 K-mediods 算法相比于 K-means 算法的优势

c) 阐述 K-mediods 算法相比于 K-means 算法的不足

d) 思考一个自动确定聚类个数的改进 kmeans 算法，或者说如何确定 kmeans 聚类个数（伪代码或者算法描述）

2. 假设数据挖掘的任务是将如下的 8 个点(用(x，y)代表位置)聚类为 3 个簇。

$$A_1(2,10), A_1(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9)$$

距离函数是欧氏距离。假设初始我们选择$A_1, B_1$和$C_1$分别为每个簇的中心，用 K-均值算法给出：

(a)在第一轮执行后的 3 个簇中心。

(b)最后的 3 个簇。

3. 假设你打算在一个给定的区域分配一些自动取款机(ATM)，使得满足大量约束条件。住宅或工作场所可以被聚类以便每个簇被分配一个 ATM。然而.该聚类可能被两个因素所约束：(1)障碍物对象,即有一些可能影响 ATM 可达性的桥梁、河流和公路。(2)用户指定的其他约束，如每个 ATM 应该能为 10000 户家庭服务。在这两个约束眼树下，怎样修改聚类算法(如 K-均值)来实现高质量的聚类？

4. 使用如下表中的相似度矩阵进行单链和全链层次聚类。绘制树状图显示结果，树状图应清楚地显示合并的次序。

| | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| P1 | 1.00 | 0.10 | 0.41 | 0.25 | 0.35 |
| P2 | 0.10 | 1.00 | 0.64 | 0.47 | 0.98 |
| P3 | 0.41 | 0.64 | 1.00 | 0.44 | 0.85 |
| P4 | 0.25 | 0.47 | 0.44 | 1.00 | 0.76 |
| P5 | 0.35 | 0.98 | 0.85 | 0.76 | 1.00 |

5. 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？

6. 假设一个布隆过滤器的容量为 $8x10^9$ 位，集合中有 $1 \times 10^9$ 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？

7. 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。

8. 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。

9. 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。