DSCI 6001P 数据科学基础

作业 3 集成、聚类、贝叶斯

提交截止日期：11.13 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. 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. 集成学习：

(1) 试析随机森林为何比决策树 Bagging 集成的训练速度更快？

(2) 集成学习中多样性增强的方法有哪些？分别阐述这些方法适用的前提。

(3) Bagging 能否提升朴素贝叶斯分类的性能？为什么？

(4) 分析 GradientBoosting [Friedman, 2001]和 AdaBoost 的异同？

3. 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。

4. 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？

5. 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。

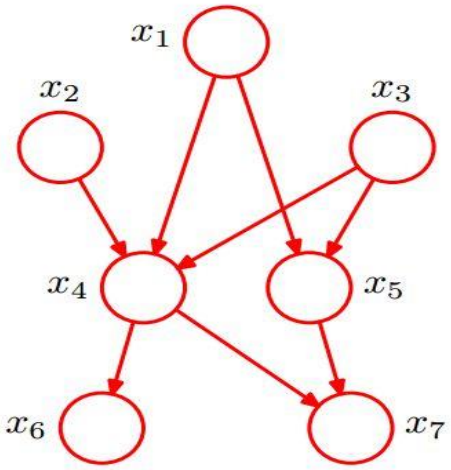
	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.25	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.25	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

6. 给定一个贝叶斯网络如下图所示：

（1） 在给定x_1，x_3的情况下，x_5，x_6是条件独立的吗？

（2） 在给定x_2，x_3的情况下，x_5，x_6是条件独立的吗？

（3） 写出x_1，x_2,…x_7的联合概率分布



7. 试由下表的训练数据学习一个朴素贝叶斯分类器并确定 x = （2，S，T）的类判别结果 y。表中 X（1）， X（2） X（3）为特征，Y 为类标记。

	1	2	3	4	5	6	7	8	9	10
$X^{(1)}$	1	1	1	2	2	1	2	2	3	3
$X^{(2)}$	S	M	M	S	S	L	M	M	L	S
$X^{(3)}$	T	T	F	F	F	T	F	T	T	F
Y	-1	-1	1	1	-1	-1	-1	1	1	1