2013 教育部-IBM 产学合作专业综合改革项目
厦门大学《数据挖掘原理及实践》课程习题

第 8 章 聚类分析

1. 使用基于中心，邻近性和密度的方法，识别下图中的簇。对于每种情况指出簇的个数，并简单给出理由。注意，明暗度或点数表示密度，其中假定基于中心即 *K* 均值，基于邻近性即单链，基于密度即 DBSCAN。

2. 对于下面的二维点集，(1) 简略描述对于给定的簇个数，如何使用 *K* 均值将它们划分成簇。(2) 指出结果质心大约在何处。假定使用平方误差目标函数，如果你认为存在多于一个解，则指出每个解是全局最小还是局部最小。注意，图中每个图的标记与本题的对应部分匹配。

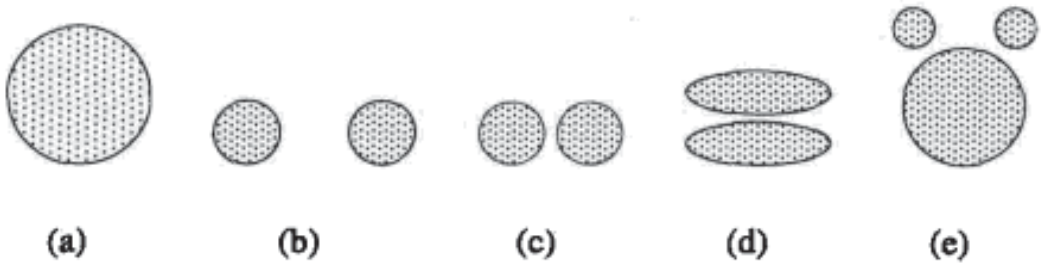
(a) *K*=2，假定点均匀分布在圆中，理论上有多少种方法能将这些点划分成两个簇？两个质心在何处？(不必提供质心的准确位置，定性描述即可) 。

(b) *K*=3。两个圆的边之间的距离略大于圆的半径。

(c) *K*=3。两个圆的边之间的距离比圆的半径小得多。

(d) *K*=2。

(e) *K*=3。提示：利用对称性，粗略找出结果。



3. 对于使用 K 均值对时间序列数据聚类，余弦度量是合适的相似性度量么？为什么？如果不是，哪种相似性度量更合适?

4. 使用下表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。

	P1	P2	P3	P4	P5
P1	1	0.1	0.41	0.55	0.35
P2	0.1	1	0.64	0.47	0.98
P3	0.41	0.64	1	0.44	0.85
P4	0.55	0.47	0.44	1	0.76
P5	0.35	0.98	0.85	0.76	1

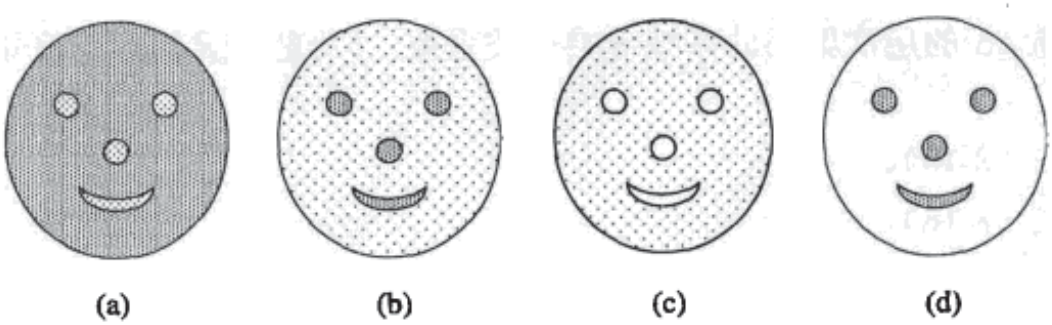
5. 假定使用 Ward 方法，K 均值找到 K 个簇，这些解中的哪些解代表局部或全局最小？并解释。

6. 考虑下图中显示的 4 个脸。明暗度或点数代表密度。线用来划分区域。

(1) 对于每个图，可以使用单链找出鼻子. 眼睛和嘴巴所代表的模式么？解释原因。

(2) 对于每个图，可以用 K 均值找出鼻子. 眼睛和嘴巴所代表的模式么？解释原因。

(3) 对于每个图，可以用 DBSCAN 找出鼻子. 眼睛和嘴巴所代表的模式么? 解释原因。



7. 请阐述 *K* 均值聚类算法的基本步骤和该算法的优缺点。

8. 层次聚类的基本方法有哪些? 层次聚类的优缺点有哪些?

9. 根据下表相似度数据矩阵进行单链和全链层次聚类。

	p_1	p_2	p_3	p_4	p_5
p_1	1.00	0.10	0.41	0.55	0.35
p_2	0.10	1.00	0.64	0.47	0.98
p_3	0.41	0.64	1.00	0.44	0.85
p_4	0.55	0.47	0.44	1.00	0.76
p_5	0.35	0.98	0.85	0.76	1.00

10. 生物学家 W. L. Grogan 和 W. W. Wirth 在 1981 年对两种蠓虫 Af 和 Apf 根据它们的触角长度和翼长加以区分。先测得 6 只 Apf 和 9 只 Af 的触角长度和翼长数据，请分别采用分类和聚类的方法对它们进行识别。

Af	1	2	3	4	5	6	7	8	9
触角长度	1.24	1.36	1.38	1.38	1.38	1.4	1.48	1.54	1.56
翼长	1.72	1.74	1.64	1.82	1.9	1.7	1.82	1.82	2.08

Apf	1	2	3	4	5	6
触角长度	1.14	1.18	1.20	1.26	1.28	1.30
翼长	1.78	1.96	1.86	2.00	2.00	1.96