第6章 聚类和降维

Exercise 6.1

假设有如下八个点：（3，1）（3，2）（4，1）（4，2）（1，3）（1，4）（2，3）（2，4），使用kmeans算法对其进行聚类。假设初始聚类中心点分别为（0，4）和（3，3），则最终的聚类中心为?

Exercise 6.2

K-means是否会一直陷入选择质心的循环停不下来?

（1）迭代次数设置
（2）设定收敛判断距离

Exercise 6.3

对以下样本数据进行主成分分析，可选择$d = 2$

x = {[2,3,3,4,5,7],[2,4,5,5,6,8]}

Exercise 6.4

k-means是一种迭代算法，在其内部循环中重复执行以下两个步骤，哪两个?

A、移动簇中心，更新簇中心u_k

B、分配簇，其中参数$c^{(i)}$被更新

C、移动簇中心u_k,将其设置为等于最近的训练示例$c^{(i)}$

D、簇中心分配步骤，其中每个簇质心u_i被分配（通过设置$c^{(i)}$）到最近的训练示例$x^{(i)}$

Exercise 6.5

最常用的降维算法是PCA，以下哪项是关于PCA的?

1、PCA是一种无监督的方法

2、它搜索数据具有最大差异的方向

3、主成分的最大数量<= 特征能数量

4、所有主成分彼此正交

A、2，3和4

B、1，2和3

C、1，2和4

D，以上都有

Exercise 6.6

主成分分析（PCA）是一种重要的降维技术，以下对于PCA的描述不正确的是（）：

A、主成分分析是一种无监督方法

B、主成分数量一定小于等于特征的数量

C、各个主成分之间相互正交

D、原始数据在第一主成分上的投影方差最小

Exercise 6.7

应用PCA后，以下哪项可以是前两个主成分？

1、（0.5，0.5，0.5，0.5）和（0.71，0.71，0.0）

2、（0.5，0.5，0.5，0.5）和（0.0，-0.71，0.71）

3、（0.5，0.5，0.5，0.5）和（0.5，0.5，-0.5，-0.5）

4、（0.5，0.5，0.5，0.5）和（-0.5，-0.5，0.5，0.5）

A、1和2

B、1和3

C、2和4

D、3和4

Exercise 6.8

给定含有5个样本的集合

$$X = \begin{bmatrix} 0 & 0 & 1 & 5 & 5 \\ 2 & 0 & 0 & 0 & 2 \end{bmatrix}$$

请用k均值聚类算法将样本聚到两个类中。

实践题

Exercise 6.1

聚类是无监督学习算法中的一种。聚类背后的主要思想相对简单，往往可以根据数据点之间的距离来将同类样本聚合在一起。

现有数据（一行代表一个样本）：

2 5

4 6

3 1

6 4

7 2

8 4

2 3

3 1

5 7

6 9

12 16

10 11

15 19

16 12

11 15

10 14

19 11

17 14

16 11

13 19

针对以上数据，实现利用K-means算法，完成以下要求：

1、当k=2（分为两个簇），输出聚类中心的坐标，各点所聚类中心位置以及SSE误差平方和。

2、画出样本数据及聚类中心的位置

3、画出，在不断增大k,(2<k<=3)，聚类代价（SSE误差平方和）的变化。

注：

距离度量 可以选用欧式距离度量 也可以其他。

SSE误差平方和介绍：

sse参数计算的内容是当前迭代得到的中心位置到各自中心点簇的欧式距离总和。值越小，分类效果越好。

提交要求：

1、提交上述要求得出的数据及图像

2、程序代码