机器学习与数据挖掘

Homework 4: Clustering Techniques

中山大学计算机学院　计算机科学与技术

19335174　施天予

一、Implement K-Means Manually

(a). What' s the center of the first cluster (red) after one iteration?

$$\mu_1 = [5.171, 3.171]$$

(b). What' s the center of the second cluster (green) after two iterations?

$$\mu_2 = [5.300, 4.000]$$

(c). What' s the center of the third cluster (blue) when the clustering converges?

$$\mu_3 = [6.200, 3.025]$$

(d). How many iterations are required for the clusters to converge?

经过 2 次迭代后收敛。如图1所示，第 2 次和第 3 次迭代的结果完全相同。

```
初始值
red center:  [6.2, 3.2]
blue center:  [6.5, 3.0]
green center:  [6.6, 3.7]
迭代次数 1
red:  [[5.9, 3.2], [4.6, 2.9], [4.7, 3.2], [5.0, 3.0], [4.9, 3.1], [5.1, 3.8], [6.0, 3.0]]
blue:  [[6.2, 2.8], [6.7, 3.1]]
green:  [[5.5, 4.2]]
red center:  [5.171428571428572, 3.1714285714285713]
blue center:  [6.45, 2.95]
green center:  [5.5, 4.2]
迭代次数 2
red:  [[4.6, 2.9], [4.7, 3.2], [5.0, 3.0], [4.9, 3.1]]
blue:  [[5.9, 3.2], [6.2, 2.8], [6.7, 3.1], [6.0, 3.0]]
green:  [[5.5, 4.2], [5.1, 3.8]]
red center:  [4.800000000000001, 3.05]
blue center:  [6.2, 3.025]
green center:  [5.3, 4.0]
迭代次数 3
red:  [[4.6, 2.9], [4.7, 3.2], [5.0, 3.0], [4.9, 3.1]]
blue:  [[5.9, 3.2], [6.2, 2.8], [6.7, 3.1], [6.0, 3.0]]
green:  [[5.5, 4.2], [5.1, 3.8]]
red center:  [4.800000000000001, 3.05]
blue center:  [6.2, 3.025]
green center:  [5.3, 4.0]
```

图 1: 聚类过程

二、Application of K-Means

(a). For dataset A, which result is more likely to be generated by K-means method?

A2

(b). Dataset B (B1 or B2?)

B2

(c). Dataset C (C1 or C2?)

C2

(d). Dataset D (D1 or D2?)

D1

(e). Dataset E (E1 or E2?)

E2

(f). Dataset F (F1 or F2?)

F2

(g). Provide the reasons/principles that draw your answers to the questions (a) to (f).

K-Means 算法中，对于每个簇中的点，该点距离簇心的距离比距离其他簇心的距离都要近

(h). For dataset F, do you think k-means perform well? Why? Are there other better clustering algorithms to be used to cluster data distributing like the data in the dataset F?

数据集 F 用 K-Means 算法效果并不好，因为很明显数据可以直接分成左右两簇，可以用层次聚类或密度聚类进行划分。

三、Applications of Clustering Techniques in IR and DM

信息检索：

- 对搜索结果进行聚类，使相似的文档一起显示。扫描几个连贯的组通常比许多单个文档更容易，如果搜索词具有不同的词义，此功能十分有用。
- 获取更好的用户界面。根据用户选择或聚集的文档组进行聚类，以获取用户所选择文档组。不断重复，合并选定的组，再次对结果集进行聚类，直到找到感兴趣的簇。

数据挖掘：

- 商场的客户群可进行了聚类分析。将客户特征与所购商品类别进行联合聚类，分析顾客特征与购买商品类别之间的联系，从而更好地排布商品