哈尔滨工业大学计算机科学与技术学院

实验报告

课程名称：机器学习

课程类型：选修

实验题目：实现k-means聚类方法

学号：7203610316

姓名：符兴

1. 实验目的

理解K-means模型，实现一个k-means算法。

2. 实验要求及实验环境

2-1. 实验要求

1. 高斯分布产生k个高斯分布的数据（不同均值和方差）（其中参数自己设定）。
2. 使用k-means聚类，测试效果；

2-2. 实验环境

Ubuntu+VSCode+Python3.9

3. 设计思想（本程序中的用到的主要算法及数据结构）

3-1. 生成训练数据

在本次实验中，使用`np.random.multivariate_normal()`生成二维高斯分布。同时设定簇个数为5，它们的均值和方差分别为：

$$\mu = [0, -4], \sigma = \begin{bmatrix} 1.4, 0 \\ 0, 1.4 \end{bmatrix}$$

$$\mu = [3, 6], \sigma = \begin{bmatrix} 1.8, 0 \\ 0, 1.8 \end{bmatrix}$$

$$\mu = [7, -5], \sigma = \begin{bmatrix} 2.25, 0 \\ 0, 2.25 \end{bmatrix}$$

$$\mu = [-7, 8], \sigma = \begin{bmatrix} 1.7, 0 \\ 0, 1.7 \end{bmatrix}$$

$$\mu = [0, 15], \sigma = \begin{bmatrix} 2.55, 0 \\ 0, 2.55 \end{bmatrix}$$

3-2.K-means

相似度计算：

$$s(x,y)=\frac{\sum_{i=1}^{p}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\overline{x})^2\times\sum_{i=1}^{p}(y_i-\overline{y})}}$$

算法流程：

输入： 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
聚类簇数 k.

过程：

1: 从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_k\}$
2: **repeat**
3: 令 $C_i=\varnothing\ (1\leqslant i\leqslant k)$
4: **for** $j=1,2,\ldots,m$ **do**
5: 计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i\ (1\leqslant i\leqslant k)$ 的距离: $d_{ji}=||\boldsymbol{x}_j-\boldsymbol{\mu}_i||_2$;
6: 根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j=\arg\min_{i\in\{1,2,\ldots,k\}}d_{ji}$;
7: 将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j}=C_{\lambda_j}\bigcup\{\boldsymbol{x}_j\}$;
8: **end for**
9: **for** $i=1,2,\ldots,k$ **do**
10: 计算新均值向量: $\boldsymbol{\mu}_i'=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$;
11: **if** $\boldsymbol{\mu}_i'\neq\boldsymbol{\mu}_i$ **then**
12: 将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}_i'$
13: **else**
14: 保持当前均值向量不变
15: **end if**
16: **end for**
17: **until** 当前均值向量均未更新

输出： 簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$

K-means是通过随机选择k个点作为初始的聚类中心，然后计算样本点距离哪个中心最近，然后就把这个样本点标记为这个中心代表的这个类中；全部标记结束后，就可以得到k个簇，通过求平均值的方式得到每个簇新的中心；如果新的中心相对于上一次的中心变化幅度小于设定的误差，程序则结束并输出结果；否则需要重复之前的划分过程，如此往复。

所以K-Means基于以下两个公式，每次最小化E然后重新求均值：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$$

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x_i$$

在K-means中，种子的选取非常重要，随机选择种子的结果会有所不同，有些种子会导致收敛速度较差，或收敛到次优聚类。

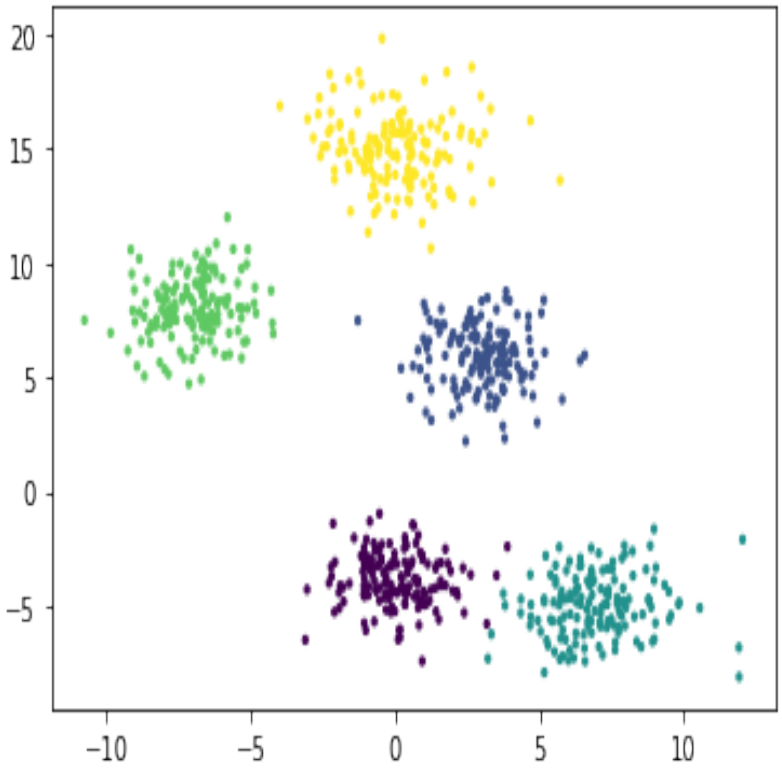
此外，本次实验使用轮廓系数（Silhouette Coefficient）来判断每类的聚类效果。

$$S_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

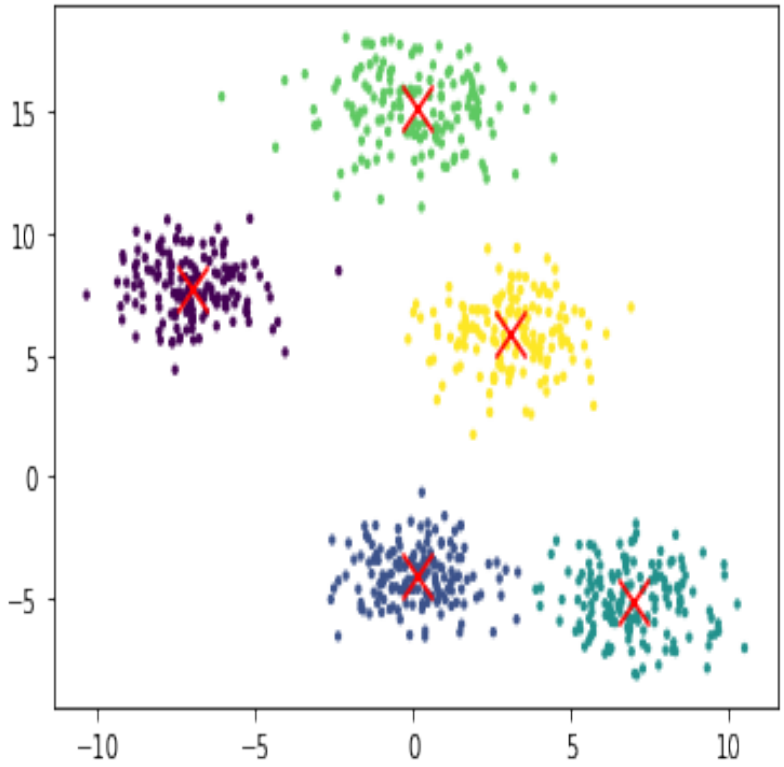
4. 实验结果分析

类别个数设定为5，任意初始化中心坐标，通过不同的初始中心点观察模型聚类效果的差异。

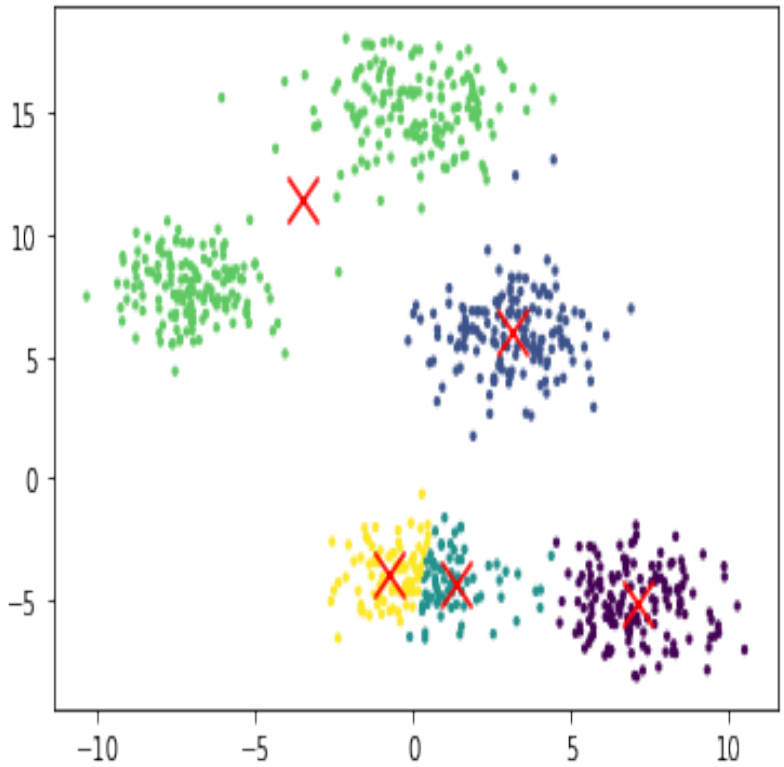
表1 不同的初始中心点的实验数据



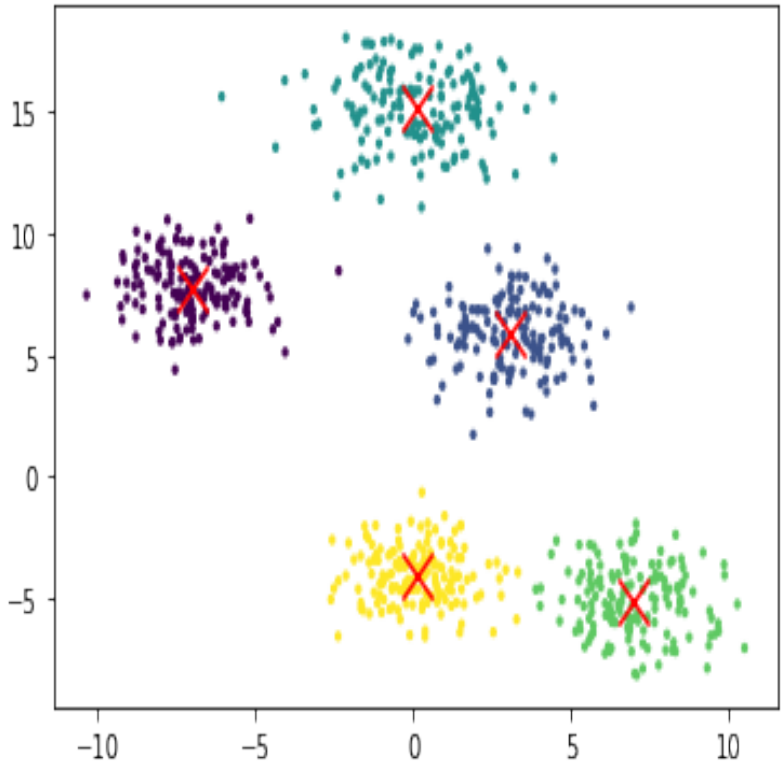
生成的数据分布



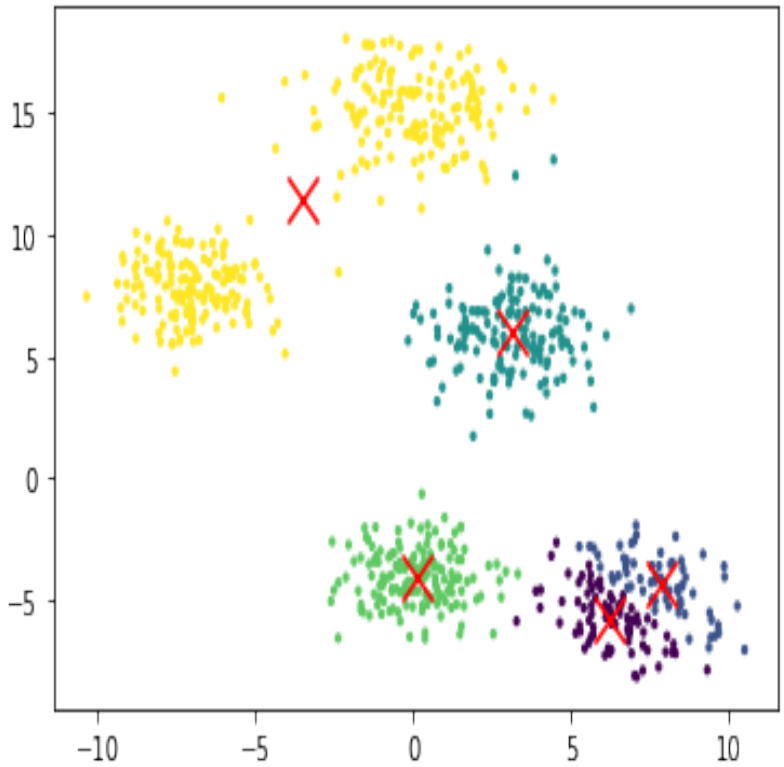
第一次聚类结果 轮廓系数：0.703



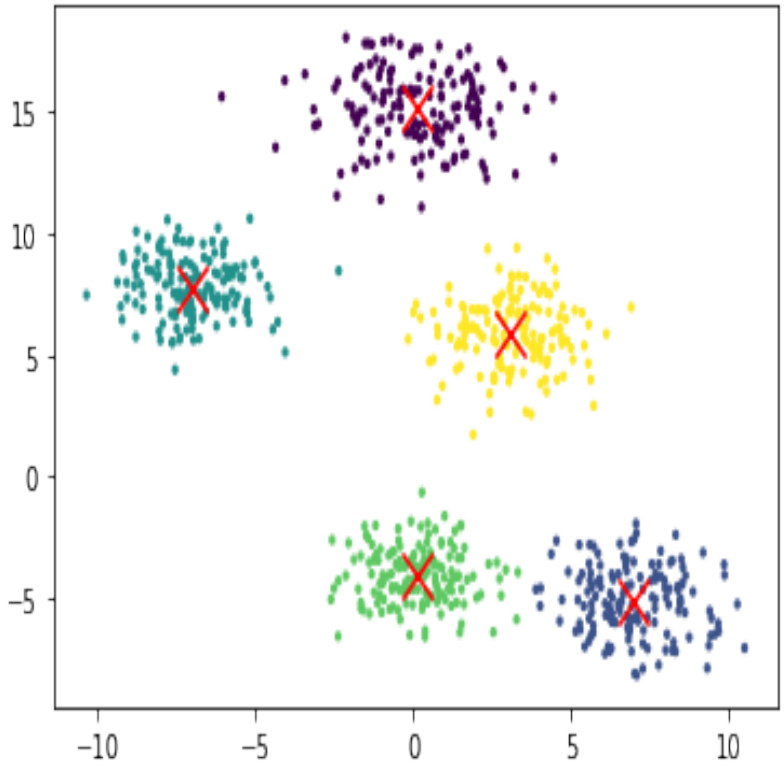
第二次聚类结果 轮廓系数：0.466



第三次聚类结果 轮廓系数：0.703



第四次聚类结果 轮廓系数：0.480



第五次聚类结果 轮廓系数：0.703

从上面五次的实验结果可以发现，K-means对初始中心点的设定极其敏感。如实验结果图2、4所示，如果有两个初始中心点距离较近，最后聚类结果可能会把原来是一类的划分为两类，原来不是一类的数据划分为一类；同时划分不正确的聚类结果其轮廓系数也比较低。

5. 结论

1. K-means需要提前设定簇的个数
2. K-means对初始中心点的设定极其敏感，会直接影响聚类效果。
3. K-means聚类效果可以通过轮廓系数进行评价，轮廓系数越高聚类效果越好。

6. 参考文献

[1]周志华. 机器学习