实验：使用 K-means 模型进行聚类

2017211146 张静雅

一、 实验目的

1. 使用 K-means 模型对 cluster.dat 里的数据进行聚类分析，尝试使用不同的类别个数 K，分析聚类结果。
2. 按照 8:2 的比例将数据随机划分为训练集和测试集，至少尝试 3 个不同的 K 值，并画出 K 下的聚类结果，及不同模型在训练集和测试集上的损失，对结果进行讨论，发现能解释数据的最好的 K 值。

二、 算法原理

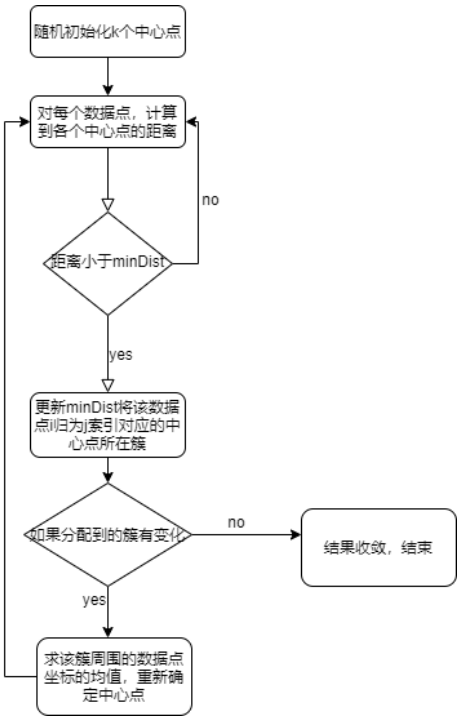
聚类分析是没有给定划分类别的情况下，根据样本相似度进行样本分组的一种方法，是一种非监督的学习算法。聚类的输入是一组未被标记的样本，聚类根据数据自身的距离或相似度划分为若干组，划分的原则是组内距离最小化而组间距离最大化。

各类簇内的样本越相似，其与该类均值间的误差平方越小，对所有类所得到的误差平方求和，即可验证分为 k 类时，各聚类是否是最优的。

K-Means：K-均值聚类也称为快速聚类法，在最小化误差函数的基础上将数据划分为预定的类数 K，该算法原理简单并便于处理大量数据。在分析结果时，采用肘部法则。

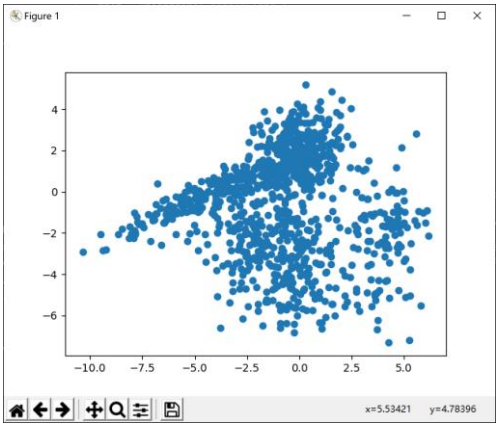
三、 算法流程图

1. 选择 K 个点作为初始质心
2. Repeat
3. 将每个点指派到最近的质心，形成 K 个簇
4. 重新计算每个簇的中心点
5. Until 簇不发生变化或达到最大的迭代次数。（本次实验设置簇不发生变化）

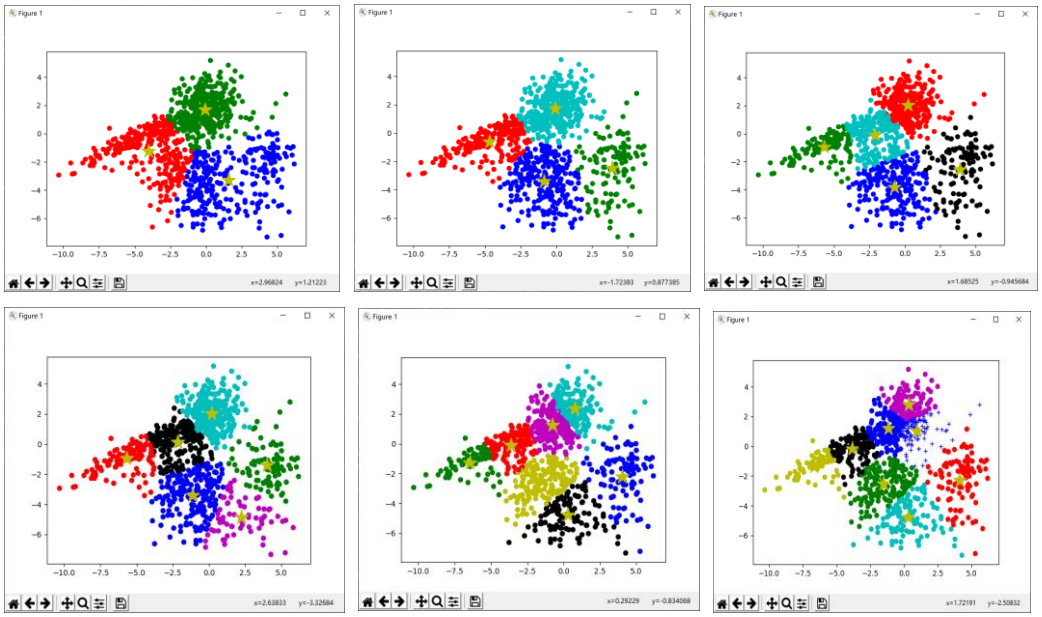


四、 实验结果分析

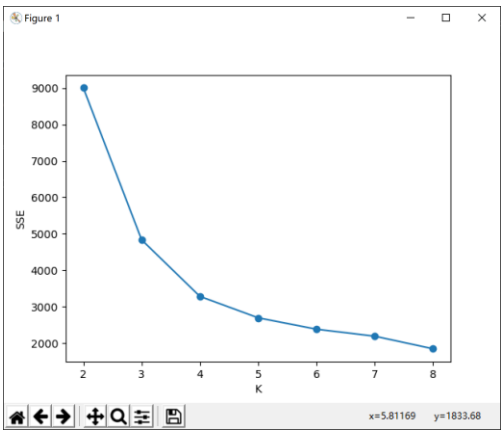
1. 原始数据集：



2. K=3、4、5、6、7、8 时的聚类结果



测试的 SSE 随 K 值的变化如下

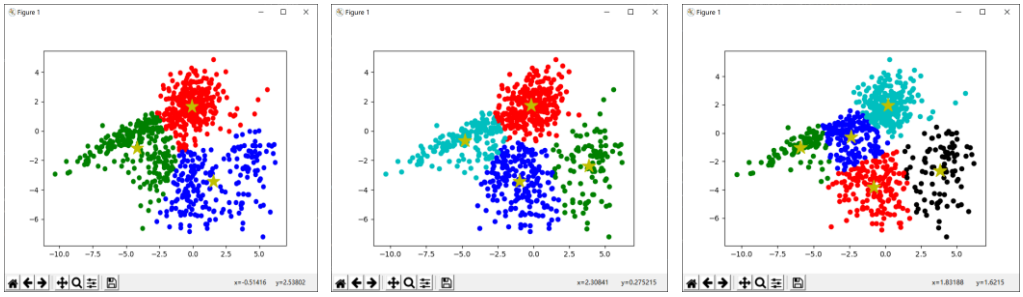


根据肘部法则：我们知道 k-means 是以最小化样本与质点平方误差作为目标函数，

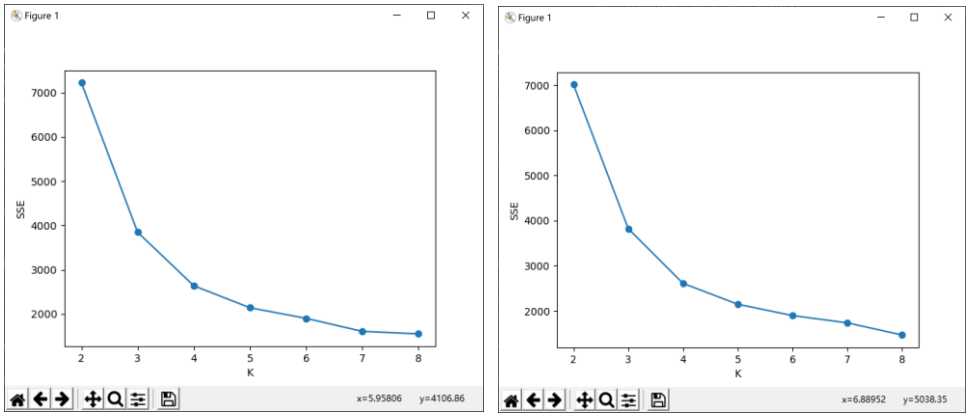
将每个簇的质点与簇内样本点的平方距离误差和称为畸变程度，那么，对于一个簇，它的畸变程度越低，代表簇内成员越紧密，畸变程度越高，代表簇内结构越松散。 畸变程度会随着类别的增加而降低，但对于有一定区分度的数据，在达到某个临界点时畸变程度会得到极大改善，之后缓慢下降，这个临界点就可以考虑为聚类性能较好的点。由上图可知：在 K=4 之后，SSE 下降幅度减缓，此时 SSE 即为聚类性能较好的点。

3. 按照 8:2 的比例随机将数据划分为训练集和测试集(800 条和 200 条)

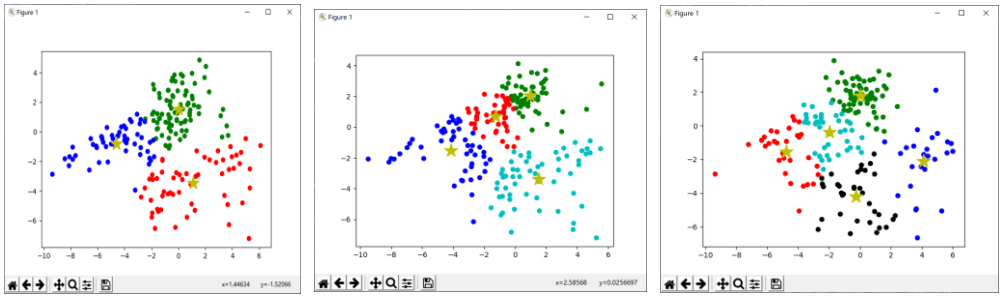
（1）在训练集上 K=3、4、5 时的聚类结果如下



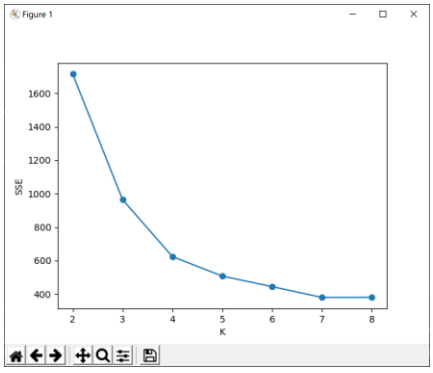
在训练集上多次测试的损失结果为：



（2）在测试集上，k=3、4、5 时的聚类结果：



在测试集上的损失：



由上述图示可以基本确定，K=4 时是性能最好的点，因为此时 SSE 下降最快，之后下降

速度减缓。

五、 实验总结及遇到的问题

1. 通过本次实验，我对 K-means 算法的原理、实现过程、结果分析有了一个完整的学习和认识，自己的动手能力和编程能力也得到了提高。
2. 在进行初始化中心点的时候，由于对数据不太熟悉只能先采用随机确定的方法，之后考虑了使用课件中的：将样本随机排序后使用前 c 个点作为代表点，得到的结果和随机取数字类似。

   • 目标函数：定义为每个样本与其簇中心点的距离的平方和（the Sum of Squared Error, SSE）

3. 由于目标函数的 SSE：$SSE = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\mathrm{dist}(x_n - \mu_k)$ e.g., dist(x_n- μ_k)=||x_n-μ_k||^2 是一个凸函数，只能找到局部最优值，所以只能随机初始化多次，再看结果的相似性。
4. 根据查阅资料，还有二分 K-means 算法，以及利用轮廓系数来评价最优 K 值的方法，轮廓系数主要是利用了簇内紧密，簇间远离的评价指标，s=b-amax(a,b) a 代表同簇样本到彼此间距离的均值，b 代表样本到除了自身所在簇外的最近簇的样本的均值，s 取值在[-1,1]之间，如果 s 接近 1，则代表样本所在簇合理，s 接近-1,则代表 s 更应该分在其他簇中。通过使用 sklearn 中的 kmeans 方法进行聚类，用 calinski_harabaz_score 方法评价聚类效果的好坏，结果：

```
(py3.7) PS E:\代码> D:\anaconda\envs\py3.7\python.exe "e:\代码\hello.py"
这个是k=2次时的轮廓系数：  0.3970081227587967
这个是k=3次时的轮廓系数：  0.4620140097494282
这个是k=4次时的轮廓系数：  0.5007699836107922
这个是k=5次时的轮廓系数：  0.41501902844414024
这个是k=6次时的轮廓系数：  0.401356434676943
这个是k=7次时的轮廓系数：  0.35876073168380906
这个是k=8次时的轮廓系数：  0.35833904815359446
```

也可以知道，k=4 的时候聚类性能比较好。