聚类实验要求和内容

一、 实验目的

1. 理解 k-means 算法的算法过程，并能够使用 Python 3 中 sklearn 包所提供的 KMeans 聚类包对数据集进行聚类
2. 理解层次聚类算法的算法过程，并能够使用 Python 3 中 scipy 包提供的 hierarchy 方法对 sklearn 中一个特定大的数据集 iris 进行聚类

二、 算法过程

因为是直接调用 Python 的包，所以此部分略去，感兴趣同学可以仔细阅读其文档

三、 实验步骤

(一) K-means 聚类

1. 由于需要使用 sklearn 包，所以需要先进行安装，可直接在 Python 3 运行终端输入语句：pip install sklearn
2. 需要调用的包有两个，第一个是调用其聚类包 cluster，第二个是一个可视化工具，以完成如图 7-5 的工作

```
from sklearn import cluster
import matplotlib.pyplot as plt
```

3. 该部分使用的数据集为书中例 7-2 所提供的数据集

```
X = [[4, 10],
     [7, 10],
     [4, 8],
     [6, 8],
     [3, 4],
     [2, 2],
     [5, 2],
     [10, 5],
     [12, 6],
     [11, 4],
     [9, 3],
     [12, 3]]
```

4. 调用 KMeans 包中的函数，并且可以设置类数以及对输入的数据集进行预测。这个调用方法可以自行上网搜索查询。
5. 完成可视化工作，利用 matplotlib 中的 pyplot 包进行数据可视化

显示出结果。

（二） 层次聚类

1. 需要调用两个包，第一个是层次聚类的包，第二个是数据集所在的包

```
import scipy.cluster.hierarchy as hcluster
import sklearn.datasets as datasets
```

2. 调用数据集的方法如下：

```
iris = datasets.load_iris()  # iris 是一个有标签的数据集
iris_data = iris.data  # 直接可以得到除去标签和其他内容而得到纯数据
```

3. 层次聚类使用方法为：

```
result = hcluster.fclusterdata(iris_data, criterion="maxclust", t=6)
```