基于K-means的鸢尾花分类

主讲人：王利猛 王瑾

目录

CONTENT

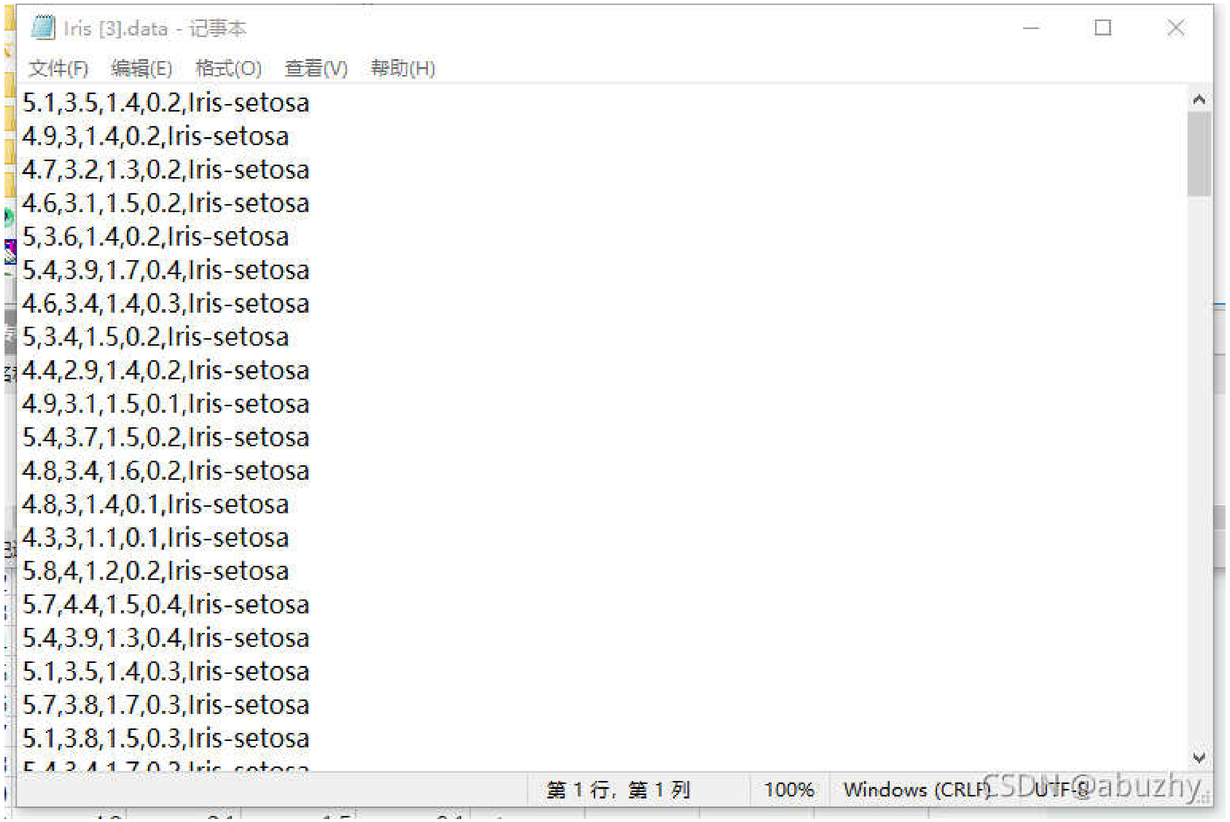
01

PART ONE

数据集介绍

数据集介绍

Iris 鸢尾花数据集是一个经典数据集，在统计学习和机器学习领域都经常被用作示例。数据集内包含 3 类共 150 条记录，每类各 50 个数据，每条记录都有 4 项特征：花萼长度、花萼宽度、花瓣长度、花瓣宽度，可以通过这4个特征预测鸢尾花卉属于（iris-setosa，iris-versicolour，iris-virginica）中的哪一品种。下图给出数据集前20条数据。



CSDN @abuzhy

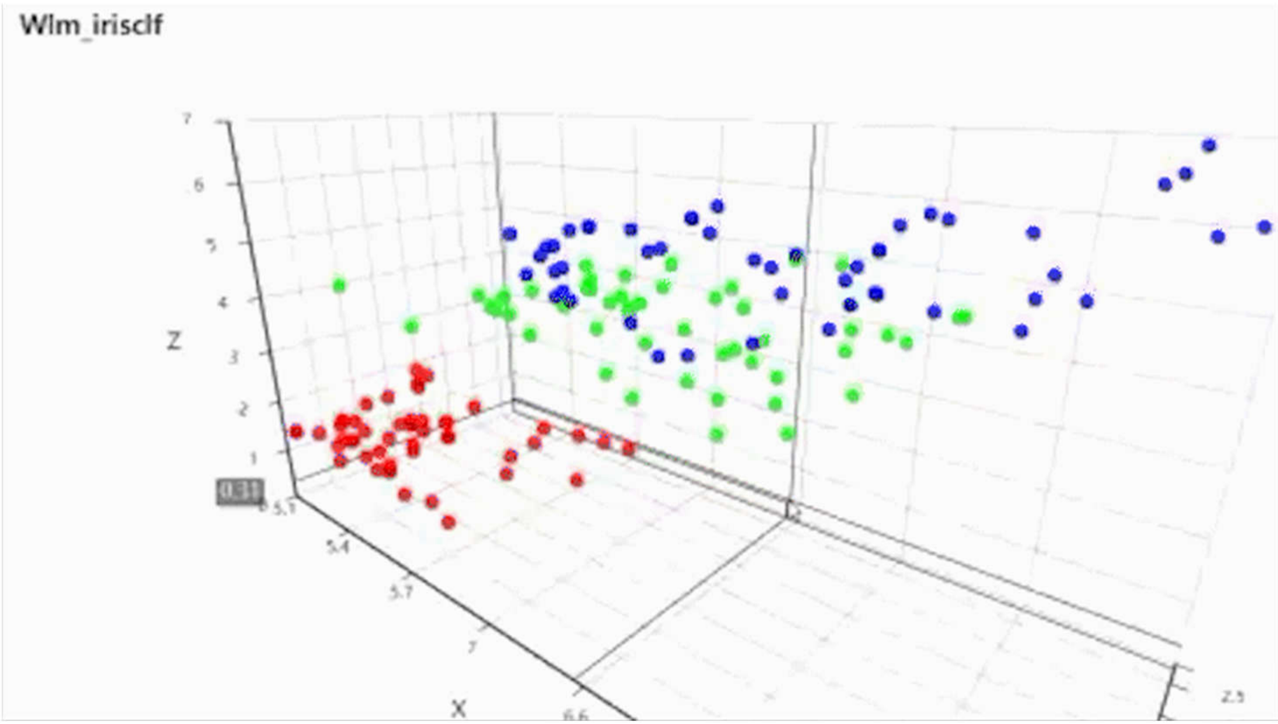
02

PART TWO

数据集可视化

数据集可视化

本项目使用的是pyecharts对sklearn中的鸢尾花数据进行可视化展示。用数据集前三个参数作为xyz坐标，可以大致看出这些数据点在空间中的分布。在下图的gif中，使用鼠标点击每个特征向量，即可显示出所属类别（红绿蓝代表三种鸢尾花）和前三个特征值（xyz坐标）。



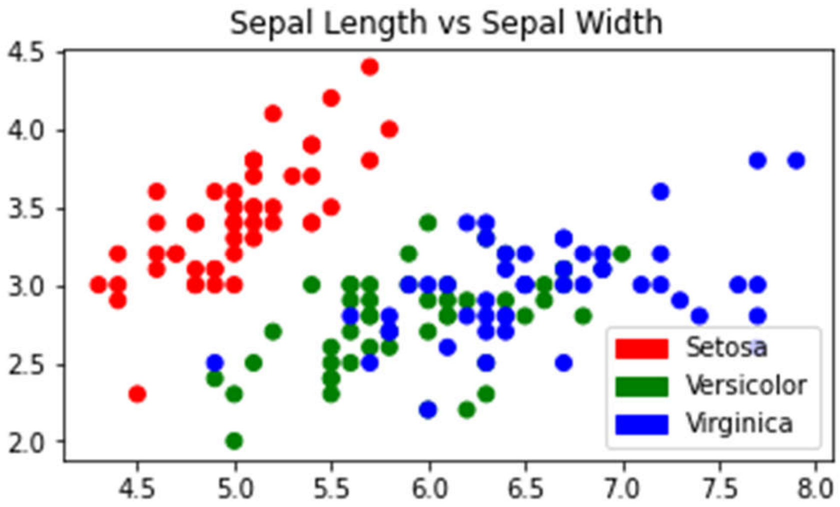
03

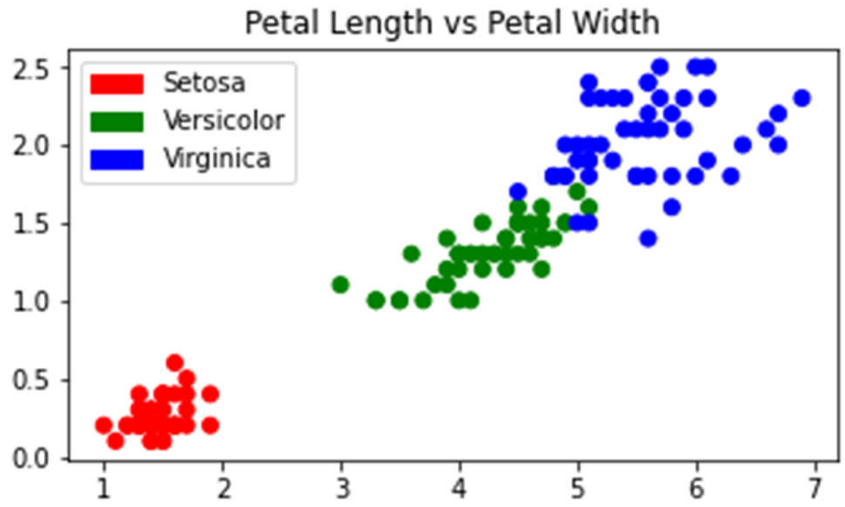
PART THREE

基于SK_learn的实现

系统实现

- 数据集前二维特征分布图，后二维特征分布图
  左图以花萼长，花萼宽为横纵坐标，绘制150个样本的特征分布
  右图以花瓣长，花瓣宽为横纵坐标，绘制150个样本的特征分布

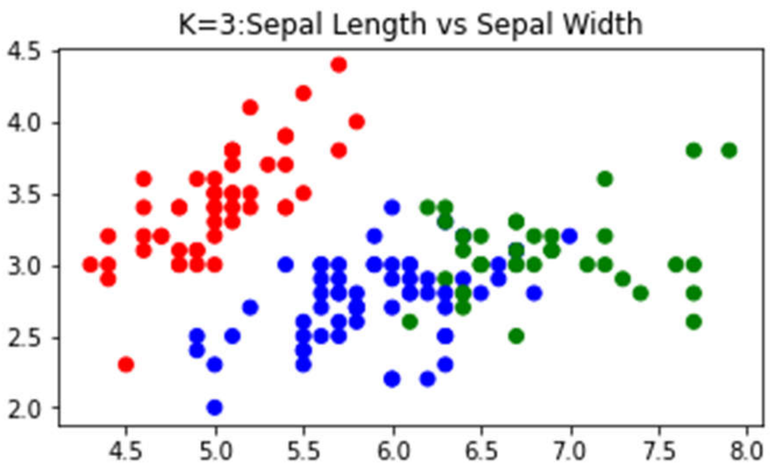


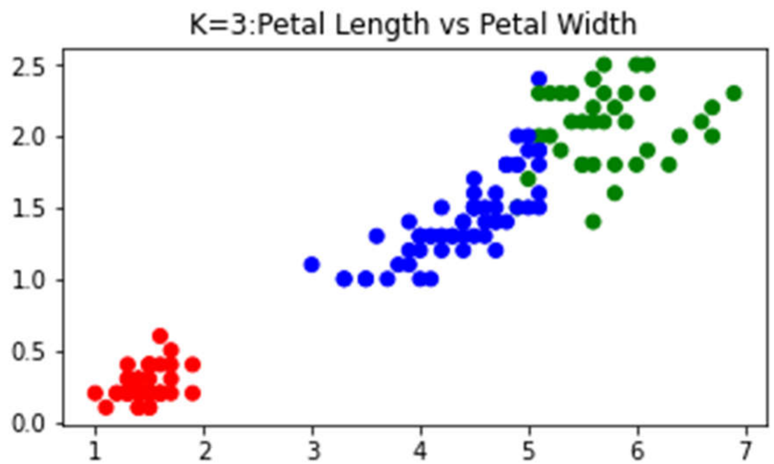


系统实现

- 令K=3进行分类
- 0 1 2代表聚类标号，不代表原数据集样本label。每次运行代码都会变化。但是标号分布不变
- 给出3个最终聚类中心
- 聚类后前两维数后两维散点分布

```
cluster_labels:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1
 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1
 1 2]
cluster_centers:
[[5.006      3.428      1.462      0.246     ]
 [6.85       3.07368421 5.74210526 2.07105263]
 [5.9016129  2.7483871  4.39354839 1.43387097]]
KMeans(n_clusters=3)
```



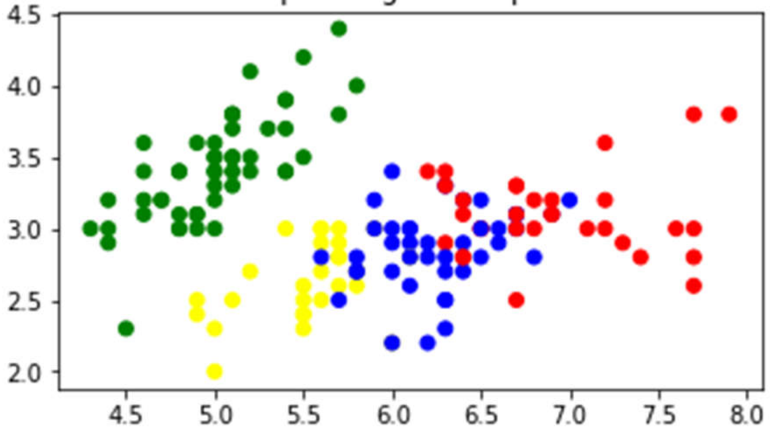


系统实现

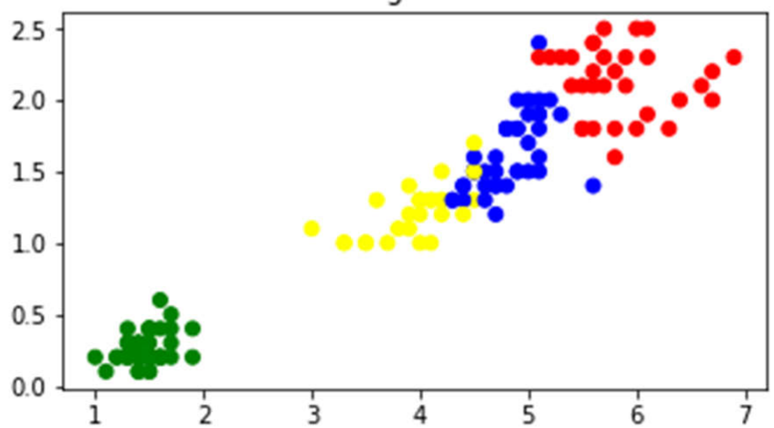
- 令K=4进行分类
- 0 1 2 3代表聚类标号，不代表原数据集样本label。每次运行代码都会变化。但是标号分布不变
- 给出4个最终聚类中心
- 聚类后前两维数后两维散点分布

```
cluster_labels:
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 2 3 2 3 2 3 3 3 3 2 3 2 3 3 2 3 2 3 2 2
 2 2 2 2 2 3 3 3 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 2 3 3 0 2 0 0 0 0 3 0 0 0 2
 2 0 2 2 0 0 0 0 2 0 2 0 2 0 0 2 2 0 0 0 0 0 2 2 0 0 0 2 0 0 0 2 0 0 0 2 2
 0 2]
cluster_centers:
[[6.9125     3.1        5.846875   2.13125   ]
 [5.006      3.428      1.462      0.246     ]
 [6.2525     2.855      4.815      1.625     ]
 [5.53214286 2.63571429 3.96071429 1.22857143]]
Text(0.5, 1.0, 'K=4:Petal Length vs Petal Width')
```

K=4:Sepal Length vs Sepal Width



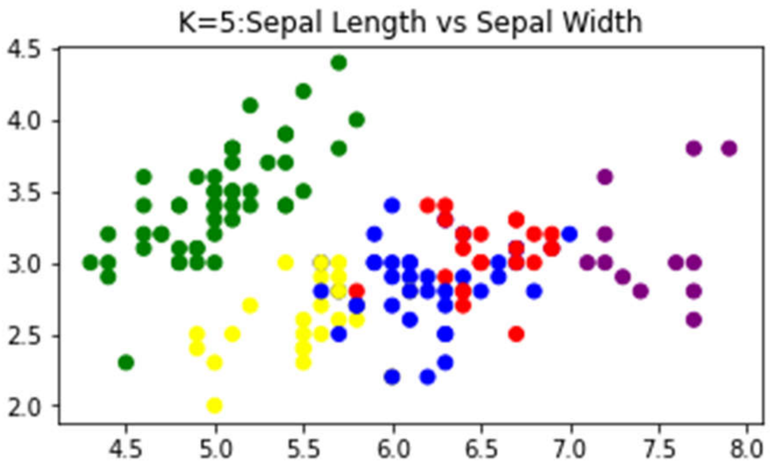
K=4:Petal Length vs Petal Width

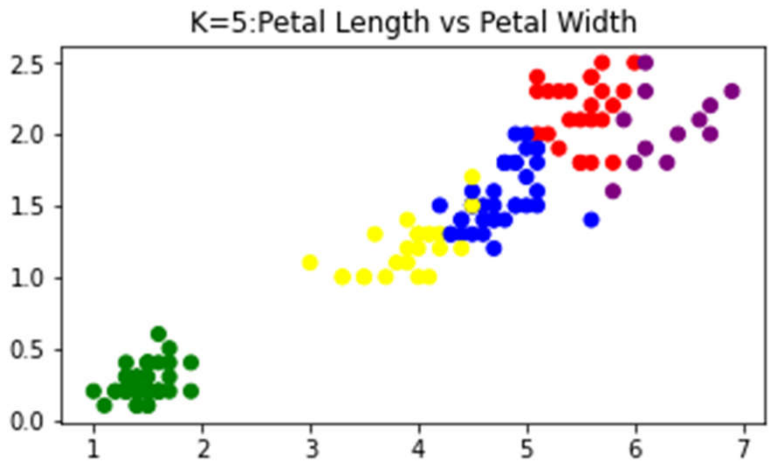


系统实现

- 令K=5进行分类
- 0 1 2 3 4代表聚类标号，不代表原数据集样本label。每次运行代码都会变化。但是标号分布不变
- 给出5个最终聚类中心
- 聚类后前两维数后两维散点分布

```
cluster_labels:
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 2 2 2 3 2 3 3 2 3 2 3 2 2 3 2 3 2 3 2 2
 2 2 2 2 2 3 3 3 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 2 3 3 0 2 4 0 0 4 3 4 0 4 0
 0 0 2 0 0 0 4 4 2 0 2 4 2 0 4 2 2 0 4 4 4 0 2 2 4 0 0 2 0 0 0 2 0 0 0 2 0
 0 2]
cluster_centers:
[[6.52916667 3.05833333 5.50833333 2.1625    ]
 [5.006      3.428      1.462      0.246     ]
 [6.20769231 2.85384615 4.74615385 1.56410256]
 [5.508      2.6        3.908      1.204     ]
 [7.475      3.125      6.3        2.05      ]]
Text(0.5, 1.0, 'K=5:Petal Length vs Petal Width')
```





04

PART FOUR

自编程实现

自编程实现

例题：

$$X_1=[0,0]^T \quad X_2=[1,0]^T \quad X_3=[0,1]^T \quad X_4=[1,1]^T$$

Step1： ① 取K=2，并选：$Z_1(1)=X_1=[0,0]^T \quad Z_2(1)=X_2=[1,0]^T$

Step2： ② 计算距离，聚类：

$$X_1:\left.\begin{array}{l} D_1=\|X_1-Z_1(1)\|=0 \\ D_2=\|X_1-Z_2(1)\|=\sqrt{(0-1)^2+(0-0)^2}=\sqrt{1}\end{array}\right\}\Rightarrow D_1<D_2\Rightarrow X_1\in S_1(1)$$

$$X_2:\left.\begin{array}{l} D_1=\|X_2-Z_1(1)\|=\sqrt{1} \\ D_2=\|X_2-Z_2(1)\|=0\end{array}\right\}\Rightarrow D_2<D_1\Rightarrow X_2\in S_2(1)$$

$$X_3:\left.\begin{array}{l} D_1=\|X_3-Z_1(1)\|=\sqrt{(0-0)^2+(1-0)^2}=\sqrt{1} \\ D_2=\|X_3-Z_2(1)\|=\sqrt{(0-1)^2+(1-0)^2}=\sqrt{2}\end{array}\right\}\Rightarrow D_1<D_2\Rightarrow X_3\in S_1(1)$$

$$X_4:\left.\begin{array}{l} D_1=\|X_4-Z_1(1)\|=\sqrt{(1-0)^2+(1-0)^2}=\sqrt{2} \\ D_2=\|X_4-Z_2(1)\|=\sqrt{(1-1)^2+(1-0)^2}=\sqrt{1}\end{array}\right\}\Rightarrow D_2<D_1\Rightarrow X_4\in S_2(1)$$

$$S_1(1)=\{X_1,X_3\} \qquad S_2(1)=\{X_2,X_4\}$$

自编程实现

例题：

Step3： 初始clusterAssment

-1	Inf
-1	Inf
-1	Inf
-1	inf

本次clusterAssment

1	D1=0
2	D2=0
1	D1=1
2	D2=1

样本聚类号都发生了改变，继续迭代

Step4： 计算新的聚类中心

$$\boldsymbol{Z}_1(2)=\frac{1}{N_1}\sum_{X\in S_1(1)}\boldsymbol{X}=\frac{1}{2}(\boldsymbol{X}_1+\boldsymbol{X}_3)=\frac{1}{2}\left(\begin{bmatrix}0\\0\end{bmatrix}+\begin{bmatrix}0\\1\end{bmatrix}\right)=\begin{bmatrix}0\\0.5\end{bmatrix}$$

$Z_2(2)$类似

自编程实现

原PPT算法计算量大，迭代基线条件难以达到（需两轮聚类中心相等），在此基础加以改进：

算法简介（设样本数m，聚类数k）

Step1：初始化聚类中心

方法：随机从m个样本中取出不重复的k个样本，作为初始样本中心c1,c2,…,ck

Step2：聚类

方法：计算第j个样本Yj对k个样本中心的距离dj1,dj2,…,djk，取出最小值dji=min{dj1,dj2,…,djk}。那么Yj的聚类号就是i。将所有样本全部聚类。结果存储到一格m*2的表clusterAssment中。此表第一列代表第一个样本的聚类号，第二列代表第一个样本到聚类中心的距离。

Step3：判断基线条件

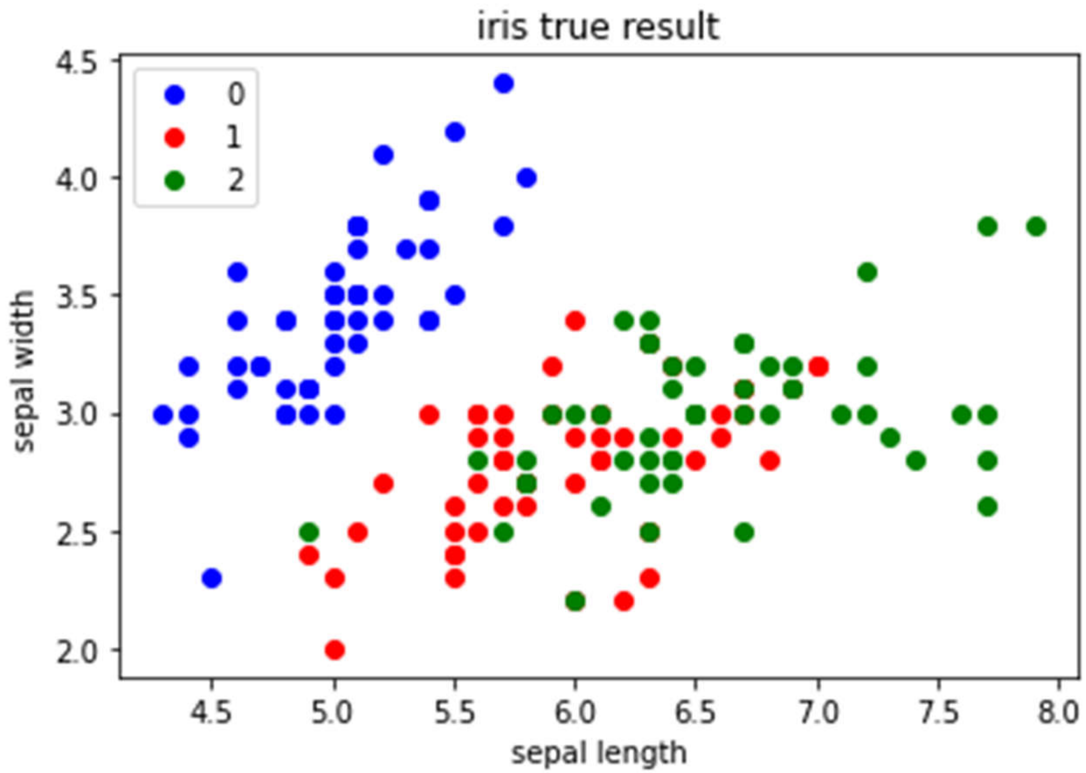
方法：逐个样本判断聚类号是不是同上一轮迭代的聚类号发生了改变，若全部样本都没改变，则迭代结束。 若存在样本聚类号发生变化，则继续迭代。

Step4：更新聚类中心

方法：新的聚类中心是此类所有样本的均值

系统实现

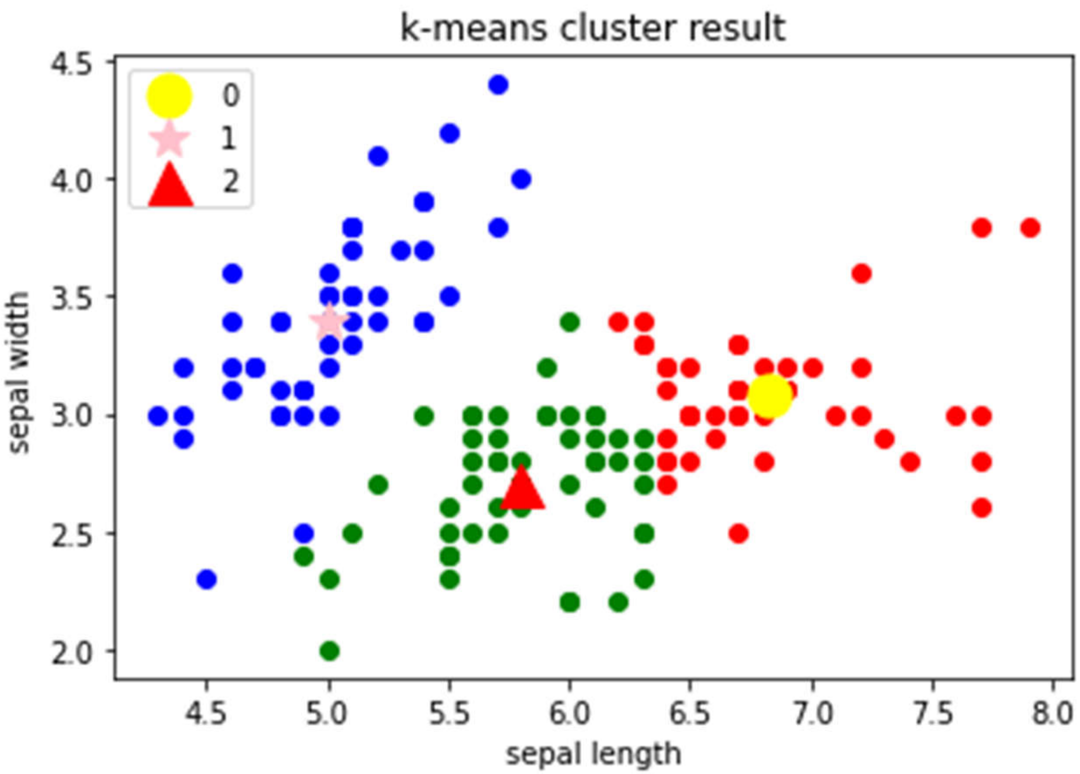
- 自编程实现：
- 原数据集分布



系统实现

- 自编程实现：
- 令K=3进行分类
- 初始中心，迭代次数，最终距离趋于最小稳定。因为聚类不再变化
- 聚类前两维数特征散点及最终聚类中心

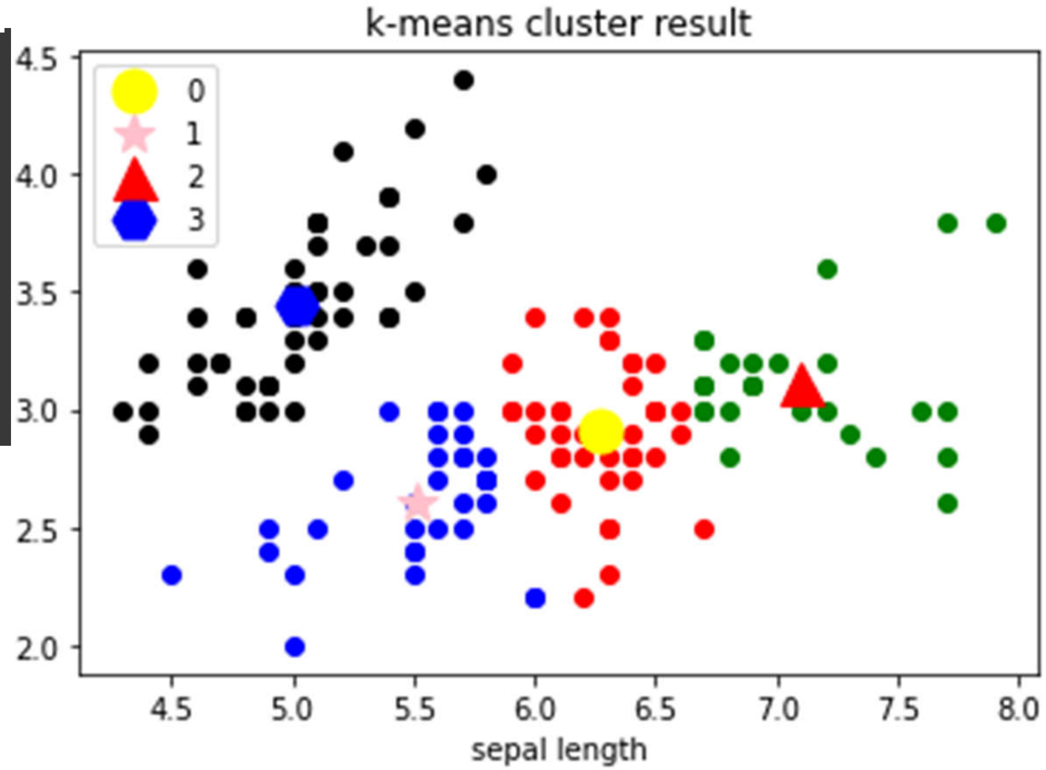
```
最初的中心= [[7.6 3. ]
 [5.4 3. ]
 [5.9 3. ]]
第1次迭代所有样本到聚类中心距离的平方为:  60.570000
第2次迭代所有样本到聚类中心距离的平方为:  42.281045
第3次迭代所有样本到聚类中心距离的平方为:  38.863737
第4次迭代所有样本到聚类中心距离的平方为:  38.291968
第5次迭代所有样本到聚类中心距离的平方为:  38.135644
第6次迭代所有样本到聚类中心距离的平方为:  38.055060
第7次迭代所有样本到聚类中心距离的平方为:  37.980634
第8次迭代所有样本到聚类中心距离的平方为:  37.859100
第9次迭代所有样本到聚类中心距离的平方为:  37.783402
第10次迭代所有样本到聚类中心距离的平方为:  37.694864
第11次迭代所有样本到聚类中心距离的平方为:  37.636365
第12次迭代所有样本到聚类中心距离的平方为:  37.535779
第13次迭代所有样本到聚类中心距离的平方为:  37.454640
第14次迭代所有样本到聚类中心距离的平方为:  37.355678
第15次迭代所有样本到聚类中心距离的平方为:  37.290519
第16次迭代所有样本到聚类中心距离的平方为:  37.229337
第17次迭代所有样本到聚类中心距离的平方为:  37.201302
第18次迭代所有样本到聚类中心距离的平方为:  37.155048
第19次迭代所有样本到聚类中心距离的平方为:  37.141172
```



系统实现

- 自编程实现：
- 令K=4进行分类
- 初始中心，迭代次数，最终距离趋于最小稳定。因为聚类不再变化
- 聚类前两维数特征散点及最终聚类中心

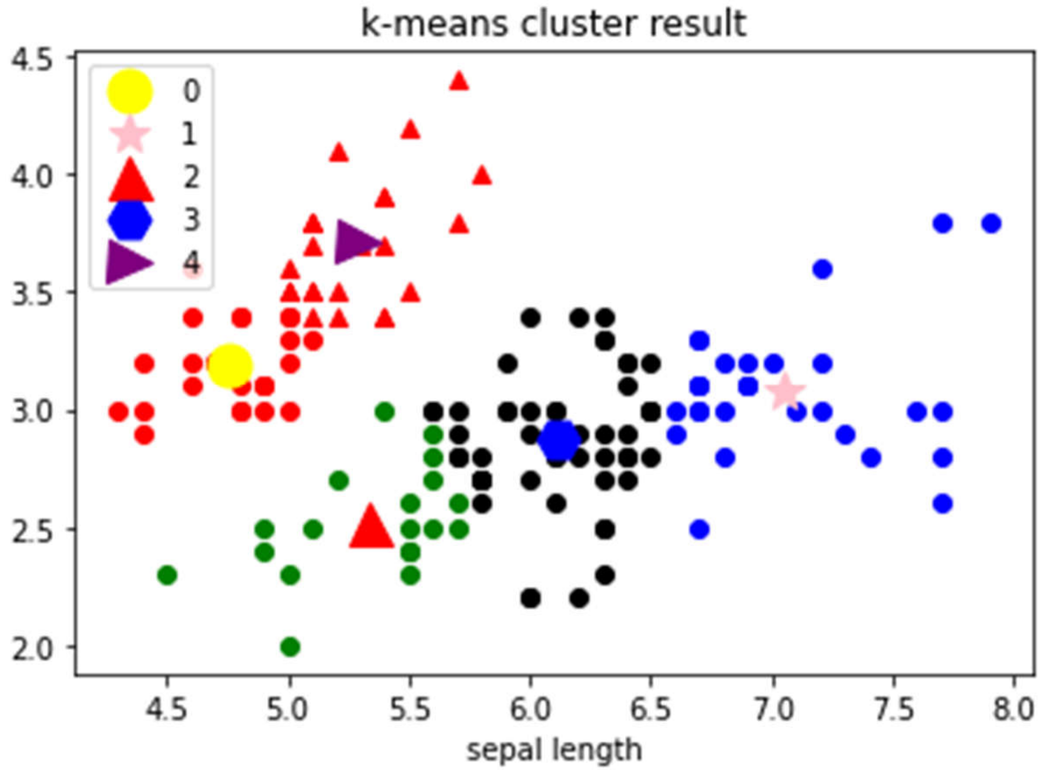
```
最初的中心= [[6.3 2.7]
 [5.9 3. ]
 [6.6 3. ]
 [4.6 3.1]]
第1次迭代所有样本到聚类中心距离的平方为:  46.760000
第2次迭代所有样本到聚类中心距离的平方为:  36.179001
第3次迭代所有样本到聚类中心距离的平方为:  35.052248
第4次迭代所有样本到聚类中心距离的平方为:  33.262579
第5次迭代所有样本到聚类中心距离的平方为:  30.274314
第6次迭代所有样本到聚类中心距离的平方为:  28.651774
第7次迭代所有样本到聚类中心距离的平方为:  28.509572
```



系统实现

- 自编程实现：
- 令K=5进行分类
- 初始中心，迭代次数，最终距离趋于最小稳定。因为聚类不再变化
- 聚类前两维数特征散点及最终聚类中心

```
最初的中心= [[4.8 3. ]
 [6.1 2.8]
 [5.5 2.4]
 [5.7 2.8]
 [4.9 3.1]]
第1次迭代所有样本到聚类中心距离的平方为:  62.690000
第2次迭代所有样本到聚类中心距离的平方为:  29.792161
第3次迭代所有样本到聚类中心距离的平方为:  26.047008
第4次迭代所有样本到聚类中心距离的平方为:  24.513531
第5次迭代所有样本到聚类中心距离的平方为:  23.974134
第6次迭代所有样本到聚类中心距离的平方为:  23.675362
第7次迭代所有样本到聚类中心距离的平方为:  23.404731
第8次迭代所有样本到聚类中心距离的平方为:  23.312118
第9次迭代所有样本到聚类中心距离的平方为:  23.106492
第10次迭代所有样本到聚类中心距离的平方为:  23.053022
第11次迭代所有样本到聚类中心距离的平方为:  22.947111
第12次迭代所有样本到聚类中心距离的平方为:  22.929565
```



感谢您的聆听

请导师点评指正