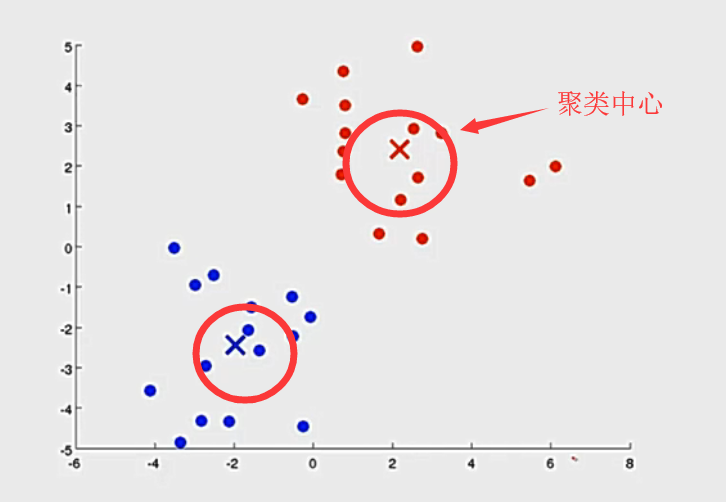
1、K-means（K均值）

（1）聚类中心：

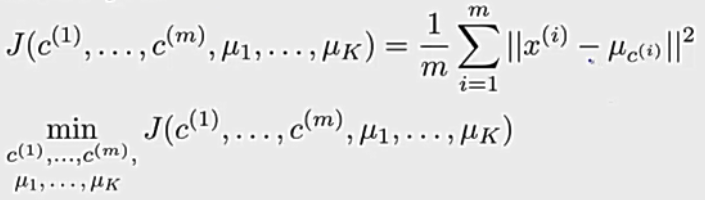


（2）原理：

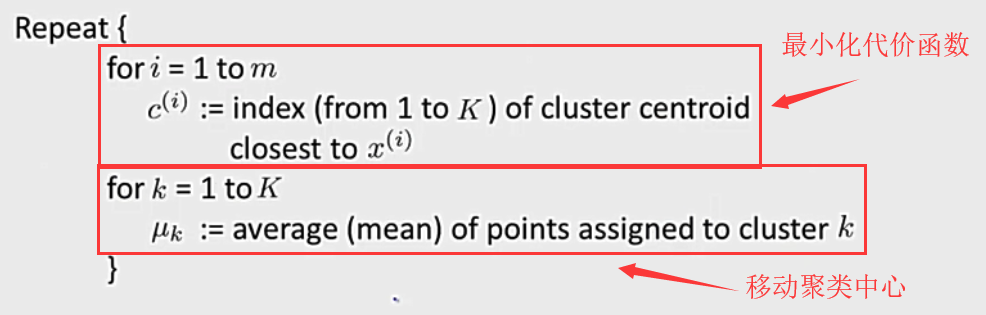
1、参数

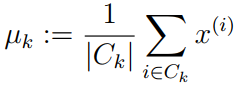


2、代价函数



3、执行步骤**（需要随机初始化聚类中心）**





找到最近的中心：

for i=1:length(idx)

distanse = pdist2(centroids,X(i,:));

% compute the distance(K,1) pdist2 is a good function

[C,idx(i)]=min(distanse); % find the minimum

end

移动聚类中心：

for i=1:K

centroids(i,:) = mean( X( find(idx==i) , :) ); %

end

**K-means算法工作流程：**

1. 确定k个初始点作为质心，将每一个数据点分配到一个簇中
2. 计算每一个簇的质心：该簇所有点的平均值（即：重新计算质心）
3. 重复上述过程，直到数据点的簇分类结果不变为止

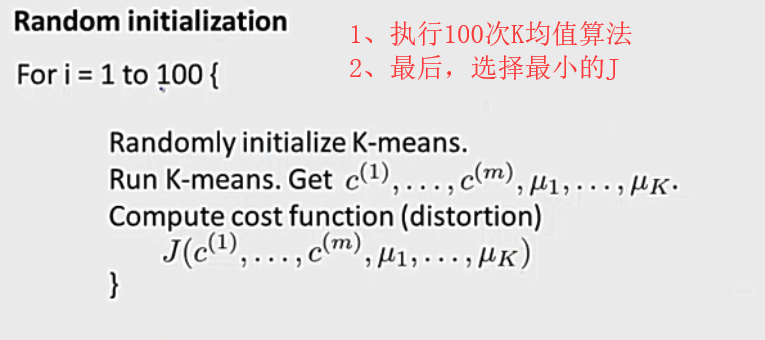
4、随机初始化

方式一：聚类中心个数K<样本数量m

方式二：随机选择聚类中心个数K

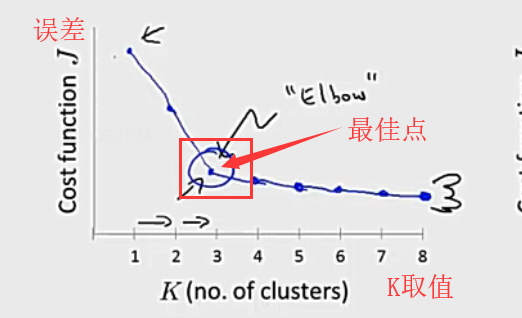
方式三：任意选取样本，作为聚类中心。

步骤：



5、选取最佳的聚类中心数量K：（肘部法则）

但该方法不一定有效果。

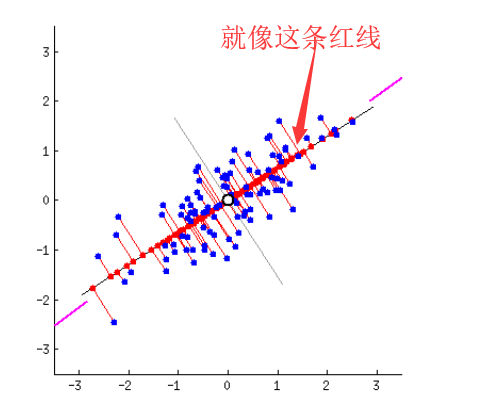


2、降维

（1）PCA：主成分分析法**（无监督学习）**

1、**思想：**找到一个平面（直线）等，使得数据投影在上面的距离最小

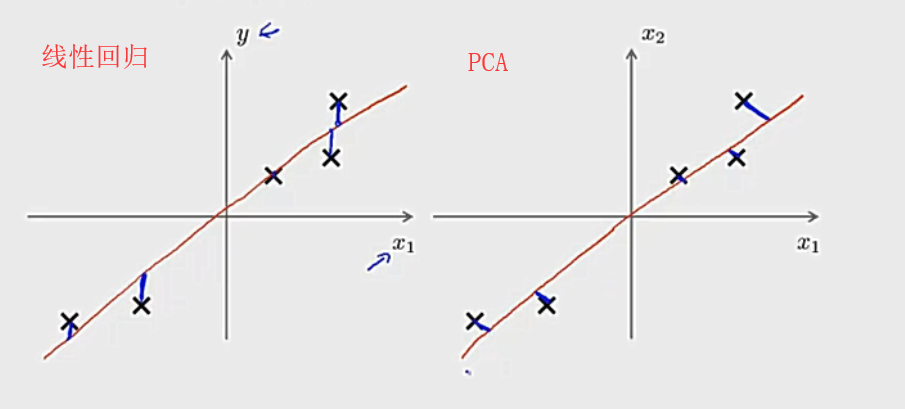
**通俗的讲：**通过沿着数据最大方差方向旋转坐标轴，来识主要特征。



该直线，也就是数据方差最大的位置。（**数据的最大方差->数据最重要的信息**）

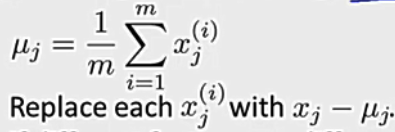
（投影距离：指的是正交投影，区别于线性回归）

2、PCA于线性回归的区别：



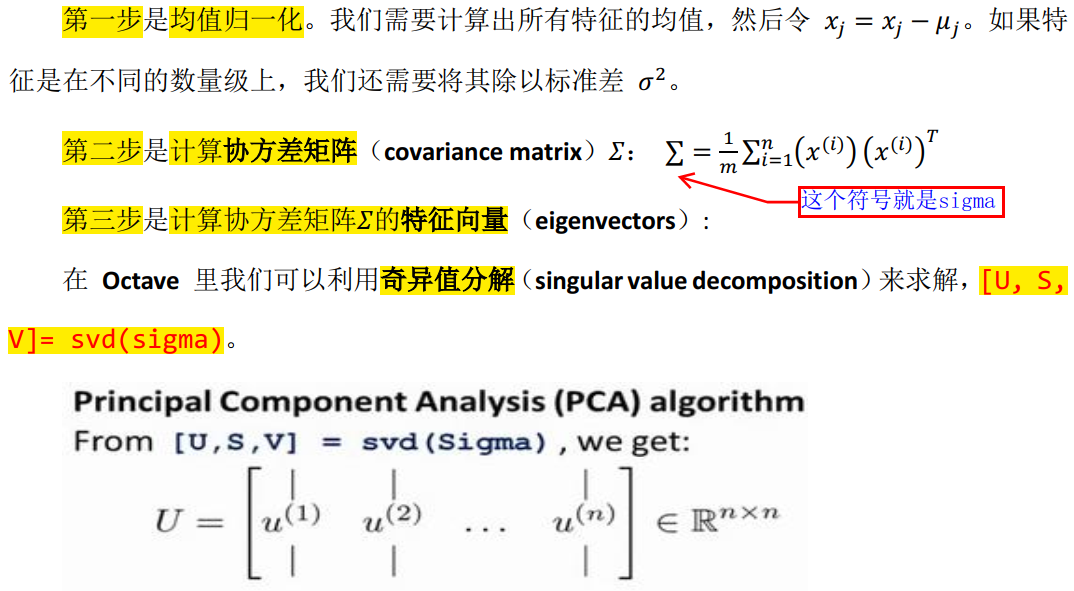
3、需要执行的步骤：

1）均值归一化：使得所有数据的均值为0

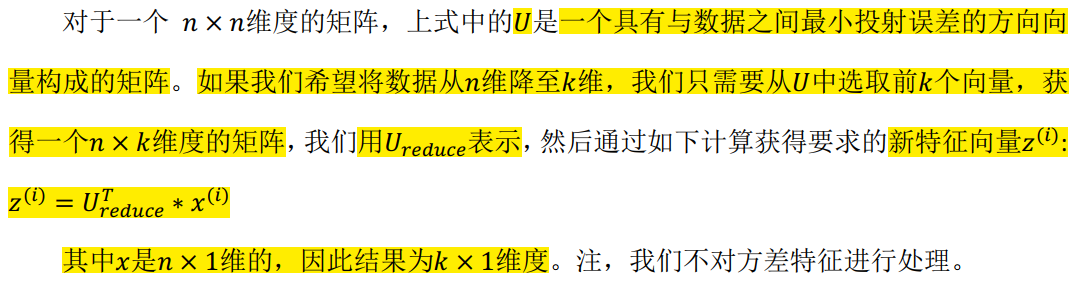


2）特征缩放：

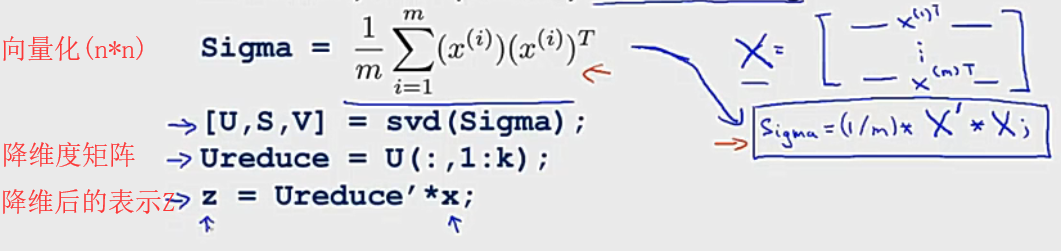
4、降维步骤：



**特征值的大小**表示对应**特征向量（u）的重要程度**，这里的U矩阵向量：按照按照特征值大小排序好了

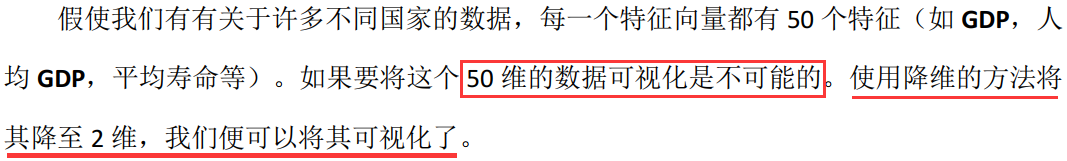


5、PCA公式：



**（PS：n为特征向量的维数）**

6、可视化



7、PCA使用：  
 1）可以提高算法的速度：由于特征值数量少了，计算速度快；

压缩数据；减小对内存、硬盘空间的使用。

2）不可以用于防止过拟合：即使特征数量可以减少，但是PCA不涉及数据标签y，有可能会损失有价值的数据。

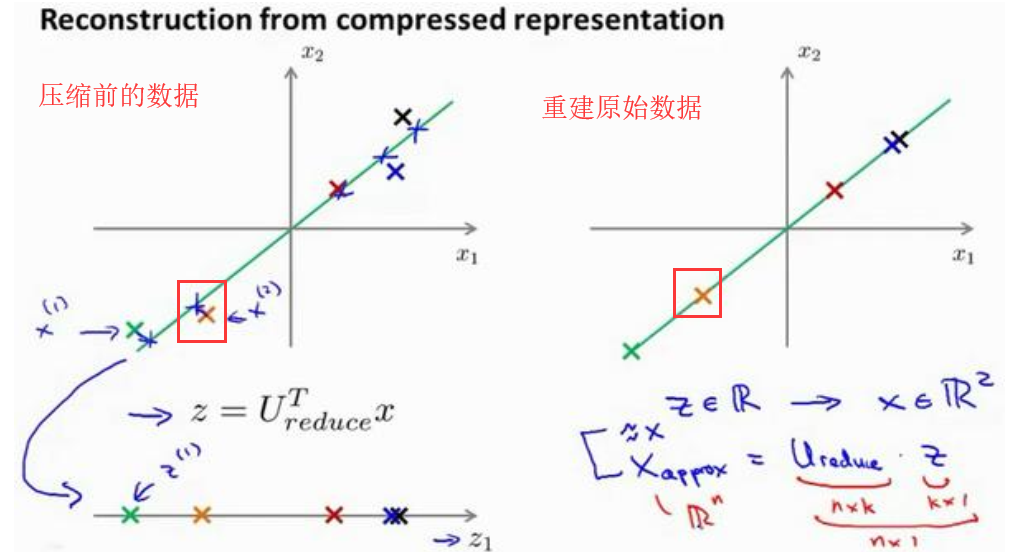
3）可用于数据加密（降维之后，数据原有含义改变）

4）去除数据中的噪声，是机器学习任务更加精确。

（2）压缩重现



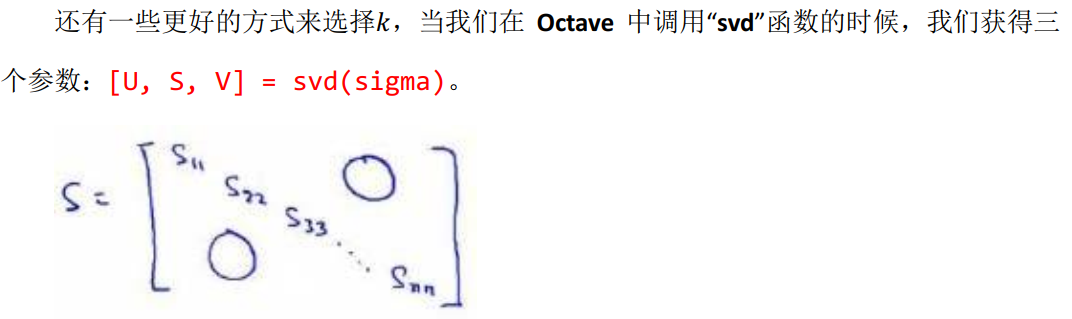
具体如下图所示：



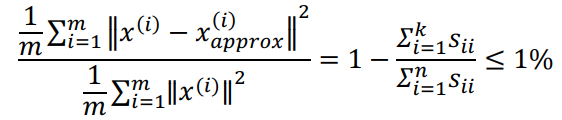
（3）主成分变量的数量（K）选择

1、方差保留的百分比

使用svd计算即可：平均均方误差/训练集均方误差



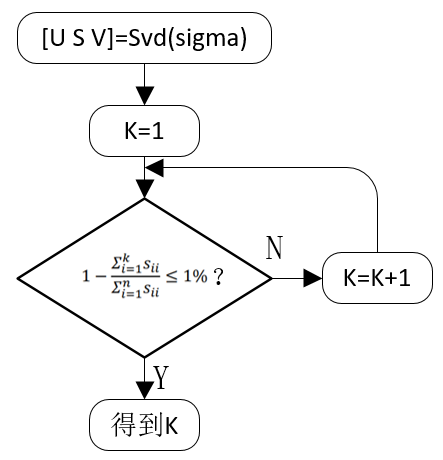
Eg：方差保留99%



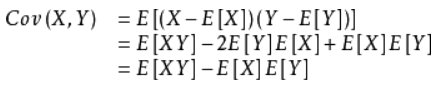
2、K的选择

选择上述满足方差保留比的K值即可。

寻找K的流程图如下：



补充知识：

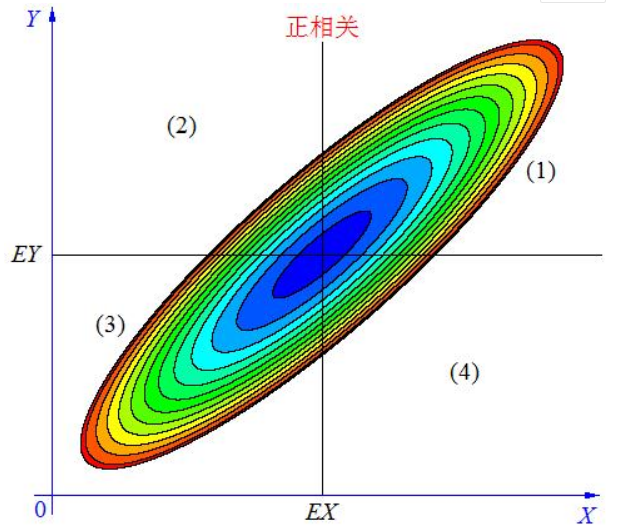
1、协方差：**协方差矩阵是一个对称的矩阵，且主对角线是各个维度上的方差。**

（1）含义：描述两个数据之间的关系

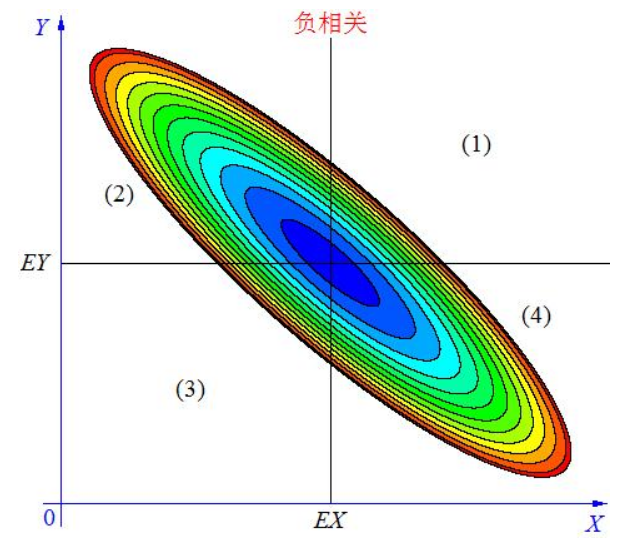
度量各个维度偏离其均值的程度。协方差的值如果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，结果为负值就说明负相关的，如果为0，也是就是统计上说的“相互独立”。

理解协方差矩阵的关键就在于牢记它计算的是不同维度之间的协方差，而不是不同样本之间。

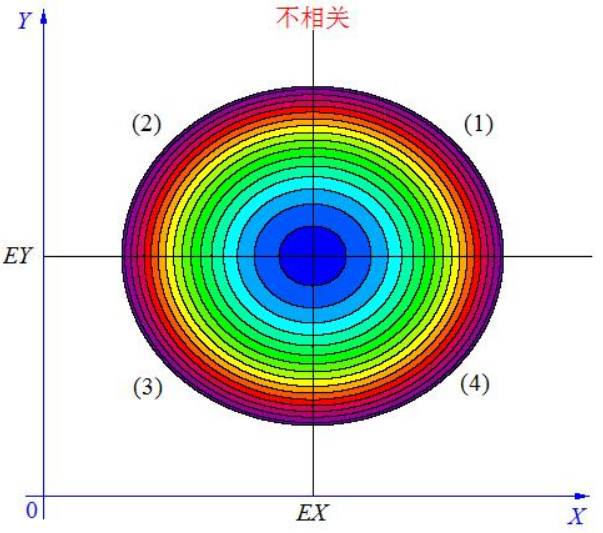
附：1：正相关



2：负相关



3：不相关



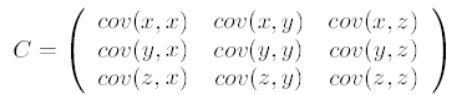
2、相关系数：



1、也可以反映两个变量变化时是同向还是反向，如果同向变化就为正，反向变化就为负。

2、由于它是标准化后的协方差，因此更重要的特性来了：它消除了两个变量变化幅度的影响，而只是单纯反应两个变量每单位变化时的相似程度。

3、协方差阵：



4、相似度表达式：

1）相似度 = 1/（1+距离）

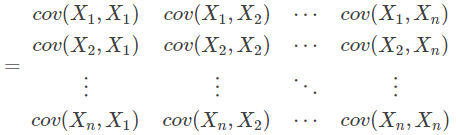
2）距离：欧式距离

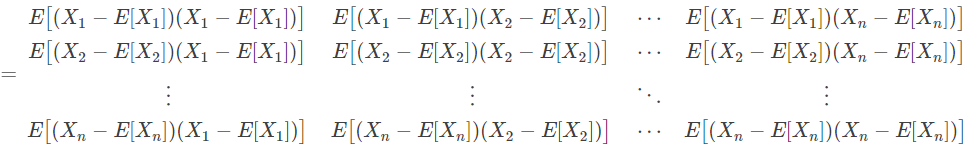
皮尔逊相关系数 = 0.5+0.5\*corrcoef()

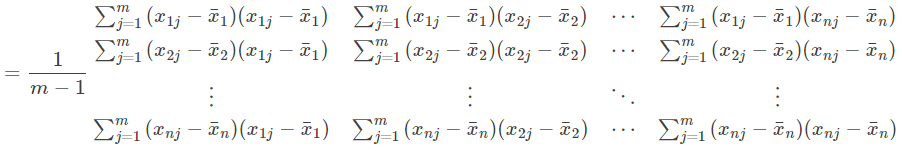
余弦相似度：，归一化之后的表达式

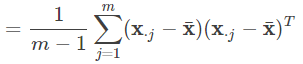
附：协方差矩阵推导：









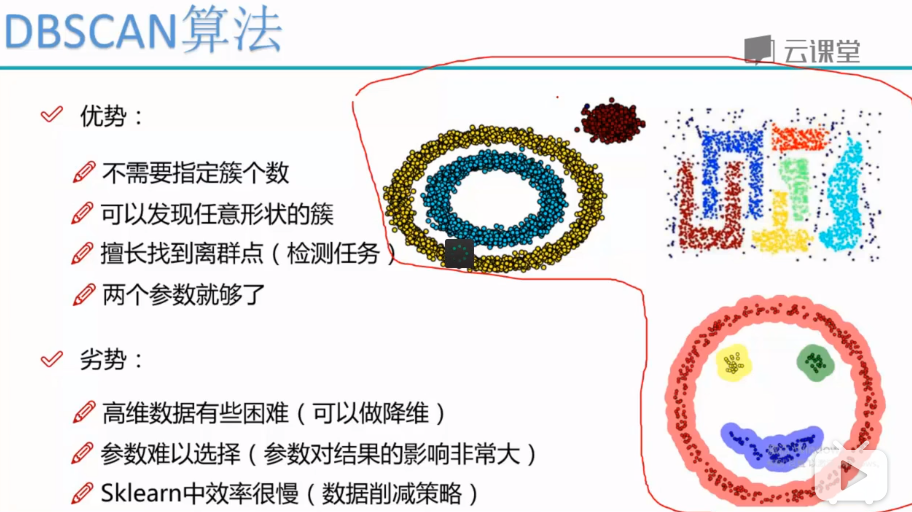


但之前使用了均值归一化，故此处的均值=0，因此有下面的式子：

4、聚类算法：

（1）K-means：度量聚类效果的指标：使用误差平方和（SSE：sum of squared error）

（2）二分K-means：克服K-means收敛于局部最小值问题

（3）DBSCAN：密度聚类算法

5、聚类效果评估：

轮廓系数评估：

