一、概述

第四章主要学习无监督学习，学习一些基本模型以及几个实例。

二、内容

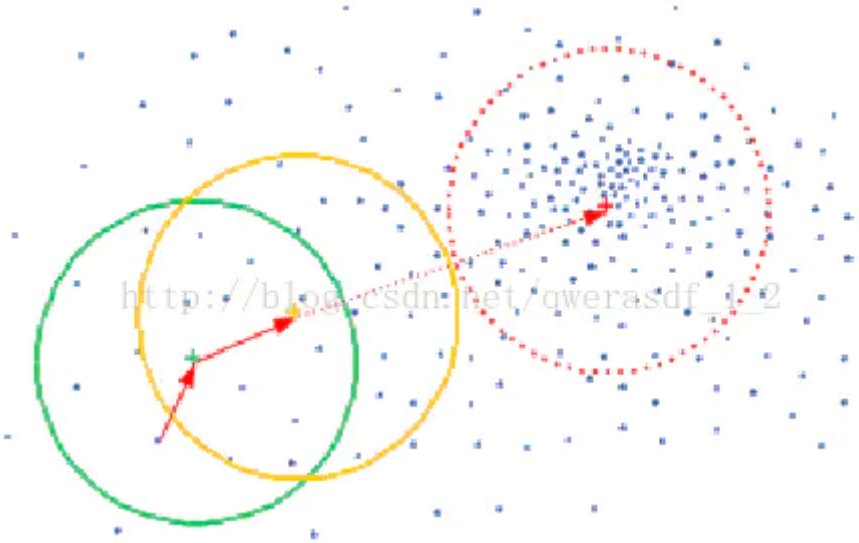
1、k-means算法

K-means一般指K均值聚类算法，是一种迭代求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。此算法简单，好理解。

2、均值漂移聚类模型

均值漂移算法是一种基于密度梯度上升的非参数方法，它经常被应用在图像识别中的目标跟踪，数据聚类，分类等场景。

其核心思想是：首先随便选择一个中心点，然后计算该中心点一定范围之内所有点到中心点的距离向量的平均值，计算该平均值得到一个偏移均值，然后将中心点移动到偏移均值位置，通过这种不断重复的移动，可以使中心点逐步逼近到最佳位置。这种思想类似于梯度下降方法，通过不断的往梯度下降的方向移动，可以到达梯度上的局部最优解或全局最优解。



3.凝聚层次聚类

层次聚类，顾名思义，就是一层一层的进行聚类，通常，我们可以把整个数据集看成一颗树，每一个数据点就是该树的叶子，而一个节点就是该节点下的叶子总数。故而，对这个数据集进行聚类分析，就相当于找到各种叶子对应的节点，这种寻找方式就是层次聚类算法。

一般的，层次聚类算法有两种：自下而上的算法和自上而下的算法。在自下而上的算法中，刚开始每个数据点（即每个叶子）都被看成一个单独的集群，然后将这些集群不断的合

并，直到所有的集群都合并成一个巨型集群，这种自下而上的合并算法也叫做凝聚层次聚类算法。

而相反的，在自上而下的算法中，刚开始所有的叶子被当做一个巨型集群，然后对这个集群进行不断的分解，直到所有的集群都变成一个个单独的数据点，即巨型集群被分解成单独的叶子节点，这种自上而下的的分解算法也叫做分裂层次聚类算法。

凝聚法的具体计算过程可以描述为：

1，将数据集中的所有的数据点都当做一个独立的集群

2，计算两两之间的距离，找到距离最小的两个集群，并合并这两个集群为一个集群，认为距离越小，两者之间的相似度越大，越有可能是一个集群。

3，重复上面的步骤2，直到聚类的数目达到设定的条件，表示聚类过程完成。

上面的计算过程看似简单，但有一个关键的难点在于：数据点或集群之间的距离计算，这种集群间距离的计算方法有很多种，下面几种比较常见的计算方法：

1，SingleLinkage: 又叫做nearest-neighbor，其本质就是取两个集群中距离最近的两个样本的距离作为这两个集群的距离。这种方式有一个缺点，会造成一种Chaning的效果，即明明两个集群相距甚远，但由于其中个别点的距离比较近而把他们计算成距离比较近。

2，CompleteLinkage: 这种计算方式是上面的SingleLinkage算法的反面，即取两个集群中距离最远的两个点的距离作为这两个集群的距离，所以它的缺点也和上面的算法类似。

3，AverageLinkage: 由于上面两种算法都存在一定的问题，都会被离群点带到沟里去，所以本算法就考虑整体的平均值，即把两个集群中的点两两的距离都计算出来后求平均值，作为两个集群的距离。有的时候，并不是计算平均值，而是取中值，其背后的逻辑也是类似的，但取中值更加能够避免个别离群数据点对结果的干扰。

一旦我们通过上面的公式计算出来两个集群的相似度，我们就可以对这两个集群进行合并。

4.评估模型

这里是对数据可以分几类，分成几类的效果最好。对此使用silhouette_score指标值确定最优分类数目。如案例4.5中分成五类。