5.2 K-均值算法基本思想

（1）随机选择一个*K*值，用以确定簇的总数。

（2）在数据集中任意选择*K*个实例，将它们作为初始的簇中心。

（3）计算这*K*个簇中心与其他剩余实例的简单欧氏距离（Euclidean Distance），用这个距离作为实例之间相似性的度量，将与某个簇相似度高的实例划分到该簇中，成为其成员之一。

$$Dis\tan ce(A\text{-}B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

其中：*A*、*B*为两个对象，x_1、y_1为对象*A*的属性，x_2、y_2为对象*B* 的属性。

（4）使用每个簇中的实例来计算该簇新的簇中心。

（5）如果计算得到新的簇中心等于上次迭代的簇中心，终止算法过程。否则，用新的簇中心作为簇中心并重复步骤（3）~（5）。

5.2 K-均值算法基本思想

终止条件可以是以下任何一个:

1)没有（或最小数目）对象被重新分配给不同的聚类。

2)没有（或最小数目）聚类中心再发生变化。

3)误差平方和局部最小。