**Kmeans原理介绍**

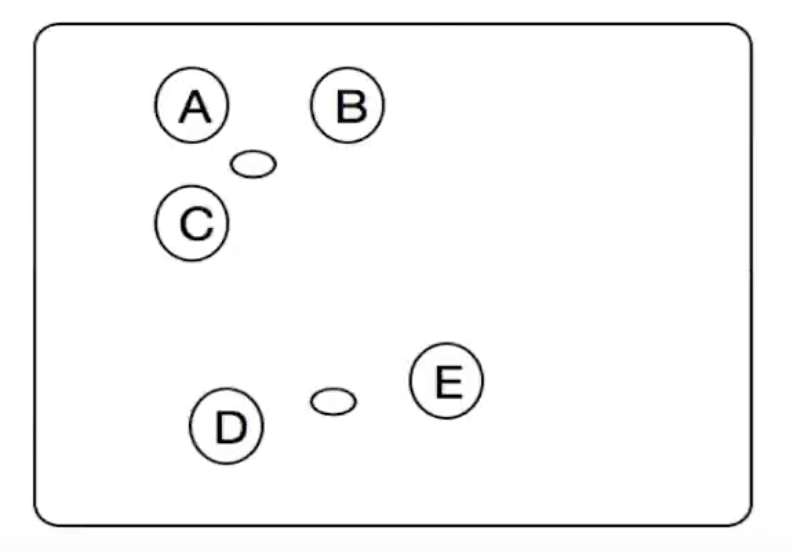
聚类kmeans算法是一个无监督学习过程。一般是用来对数据对象按照其特征属性进行分组。经常被应用在客户分群、欺诈检测、图像分析等领域。K-means是最有名并且最经常使用的聚类算法。

**算法介绍：**

Kmeans算法的基本思想是初始化随机给定K个簇中心，按照最邻近原则把待分类样本点分割到各个簇，然后按照平均法重新计算各个簇的质心，从而确定簇心，一直迭代，直到簇心的移动距离小于给定的值。

**迭代过程：**

1. 选择k个点作为初始聚类中心。（k需要我们程序自己设置）
2. 计算其余所有点到聚类中心的距离，并把每个点划分到离它最近的聚类中。最常用的衡量距离的函数为欧几里得距离，也叫做欧式距离。
3. 重新计算每个聚类中所有点的平均值，并将作为新的聚类中心点。
4. 重复2、3步的过程，直至聚类中心不再发生变化，或者算法达到预定的迭代次数（程序自己设置），又或者聚类中心的改变小于预定的阈值。



Kmeans有两个重要的问题

1. 选择K值

k是我们kmeans算法的关键值。

Spark ml中提供computecost。所有点到其最近的中心点的平方和来评估聚类效果。同样的迭代次数和算法跑的次数，这个值越小代表聚类的效果越好。

1. 初始聚类中心点的选择
   1. 随机模式：会造成的聚类的结果和数据的实际分布相差很大。
   2. k-means++: 初始的聚类中心之间的相互距离要尽可能远。
2. 从输入的数据点集合中随机选择一个点作为第一个初始点。
3. 计算数据集中所有点与最新选择的中心点的距离 D(x)
4. 选择下一个中心点，p(x)=D(x)2/ED(x)2 最大
5. 重复b、c步过程，直到k个初始点选择完成。