k-means 聚类

一、聚类与分类的区别

分类：类别是已知的，通过对已知分类的数据进行训练和学习，找到这些不同类的特征，再对未分类的数据进行分类。属于监督学习。

聚类：事先不知道数据会分为几类，通过聚类分析将数据聚合成几个群体。聚类不需要对数据进行训练和学习。属于无监督学习。

二、k-means 聚类

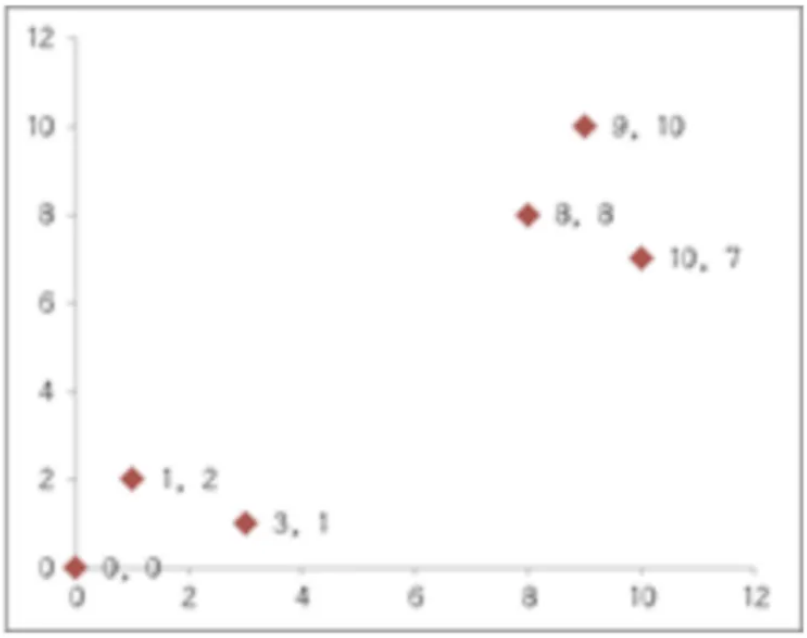
聚类算法有很多种，K-Means 是聚类算法中的最常用的一种，算法最大的特点是简单，好理解，运算速度快，但是只能应用于连续型的数据，并且一定要在聚类前需要手工指定要分成几类。

K-Means 聚类算法的大致意思就是“物以类聚，人以群分”：

- 首先输入 k 的值，即我们指定希望通过聚类得到 k 个分组；
- 从数据集中随机选取 k 个数据点作为初始质心；
- 对集合中每一个数据，计算与每一个质心的距离，离哪个质心距离近，就跟定哪个质心。
- 这时每一个质心都聚集了一些数据，这时候召开选举大会，每一群选出新的质心。
- 如果新的质心和老的质心之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），可以认为我们进行的聚类已经达到期望的结果，算法终止。
- 如果新质心和老质心距离变化很大，需要迭代3~5步骤。

举例：有6个点，从图上看应该可以分成两堆，前三个点一堆，后三个点另一堆。

	X	Y
P1	0	0
P2	1	2
P3	3	1
P4	8	8
P5	9	10
P6	10	7



1.设定 k 值为2

2.选择初始质心（就选 P1 和 P2）

3.计算元素与质心的距离：

	P1	P2
P3	3.16	2.24
P4	11.3	9.22
P5	13.5	11.3
P6	12.2	10.3

从上图可以看出，所有的元素都离 P2 更近，所以次站队的结果是：

A 组：P1

B 组：P2、P3、P4、P5、P6

4.召开选举大会：

A 组没什么可选的，质心就是自己

B 组有5个人，需要重新选质心，这里要注意选质心的方法是每个人 X 坐标的平均值和 Y 坐标的平均值组成的新的点，为新质心，也就是说这个质心是“虚拟的”。因此，B 组选出新质心的坐标为：Pn（（1+3+8+9+10）/5，（2+1+8+10+7）/5）=（6.2，5.6）。

综合两组，新的质心为 P1（0，0），Pn（6.2，5.6），而P2–P6重新成为两个质心的元素。

5.再次计算元素到质心的距离：

	P1	Pn
P2	2.24	6.3246
P3	3.16	5.6036
P4	11.3	3
P5	13.5	5.2154
P6	12.2	4.0497

这时可以看到P2、P3离P1更近，P4、P5、P6离P哥更近，所以第二次站队的结果是：

A 组：P1、P2、P3

B 组：P4、P5、P6（虚拟质心这时候消失）

6.第二届选举大会：

同样的方法选出新的虚拟质心：Pn1（1.33，1），Pn2（9，8.33），P1–P6都成为其元素。

7.第三次计算元素到质心的距离：

	Pn1	Pn2
P1	1.4	12
P2	0.6	10
P3	1.4	9.5
P4	47	1.1
P5	70	1.7
P6	56	1.7

这时可以看到 P1、P2、P3 离 Pn1 更近，P4、P5、P6离 Pn2 更近，所以第二次站队的结果是：

A 组：P1、P2、P3

B 组：P4、P5、P6

我们可以发现，这次站队的结果和上次没有任何变化了，说明已经收敛，聚类结束。