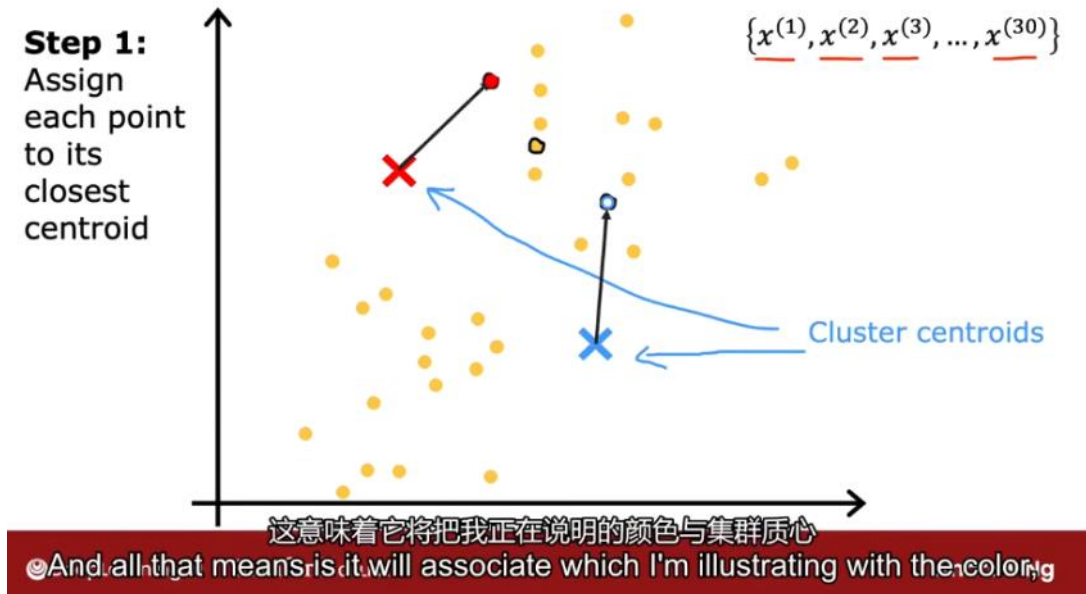


# K-means

2023年5月31日 9:02

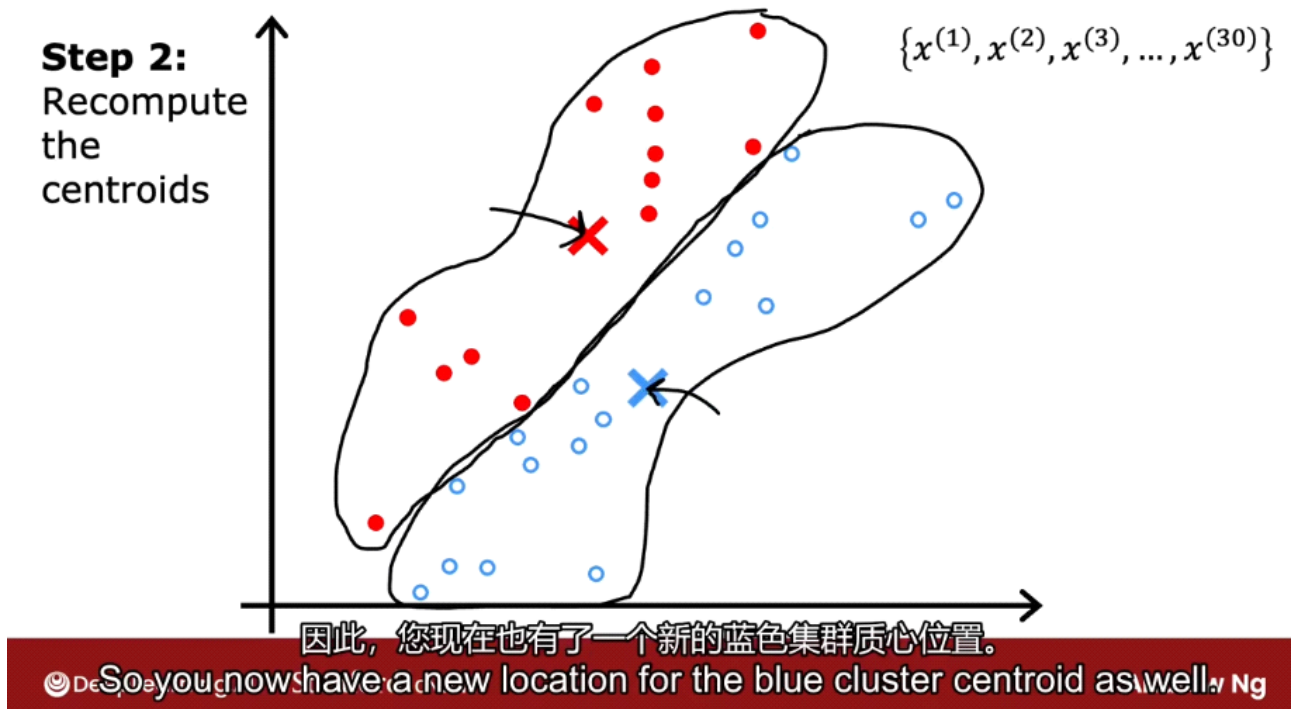
## 1.K-means Algorithm

第一步：随机设置聚类中心点，把每个数据点分配给离它最近的中心点



屏幕剪辑的捕获时间: 2023/5/31 9:15

第二步：计算分组后的两组数据各自的“平均”位置，然后将第一步中设置的聚类中心点移动至计算所得的“平均”位置，重复以上过程



屏幕剪辑的捕获时间: 2023/5/31 9:21

算法实现：（图中n为数据的特征数量，聚类中心的维数与数据的特征维数相同）

## K-means algorithm

$\mu_1, \mu_2$

$x^{(1)}, x^{(2)}, \dots, x^{(n)}$

$n=2$

$K=2$

Randomly initialize  $K$  cluster centroids  $\mu_1, \mu_2, \dots, \mu_K$

Repeat {

# Assign points to cluster centroids

for  $i = 1$  to  $m$

$c^{(i)} :=$  index (from 1 to  $K$ ) of cluster centroid closest to  $x^{(i)}$

# Move cluster centroids

for  $k = 1$  to  $K$

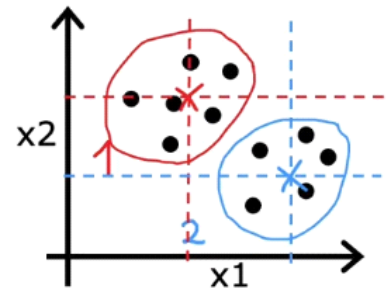
$\mu_k :=$  average (mean) of points assigned to cluster  $k$

}

$$\mu_1 = \frac{1}{4} [x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}]$$

, 集群 1 的新集群质心。

the new cluster centroid for cluster 1.



DeepLearning.AI

Stanford

Andrew Ng

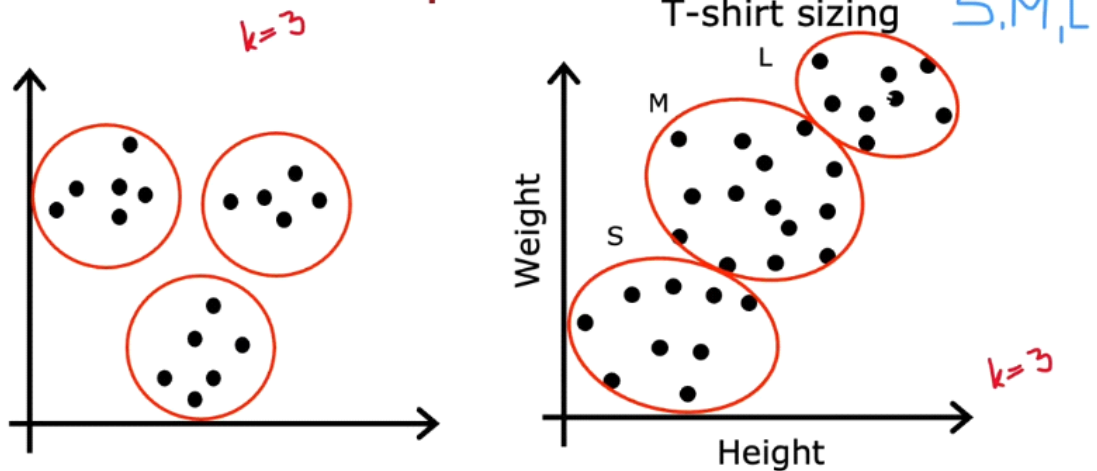
屏幕剪辑的捕获时间: 2023/5/31 10:50

如果有哪一次循环后, 有某个聚类没有分得数据, 那么可以直接消除这个聚类。

K-means算法对于分类并不明显的数据集也很有用, 比如下图中根据不同身高体重的人所

买衣服的不同, 判断小、中、大三种型号的衣服都是哪些顾客在买。

## K-means for clusters that are not well separated



将这些人与潜在的集

fit these individuals well with

DeepLearning.AI

Stanford

Andrew Ng

屏幕剪辑的捕获时间: 2023/5/31 10:56

## 2. 优化目标(Cost function in K-means)

K-means的cost函数

## K-means optimization objective

$c^{(i)}$  = index of cluster (1, 2, ..., K) to which example  $x^{(i)}$  is currently assigned

$\mu_k$  = cluster centroid  $k$

$\mu_{c^{(i)}}$  = cluster centroid of cluster to which example  $x^{(i)}$  has been assigned

### Cost function

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

$\min_{c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K} J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$

Distortion

J 正在计算的内容。

the distortion cost function, that's just what this formula J is computing. Ng

K-means 执行中的两个步骤，就是两种不同的减小 J 的方式

第一步是固定  $\mu$ ，调整  $c$ ；第二步是固定  $c$ ，调整  $\mu$

### Cost function for K-means

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

Repeat {

# Assign points to cluster centroids

for  $i = 1$  to  $m$

$c^{(i)} :=$  index of cluster  
centroid closest to  $x^{(i)}$

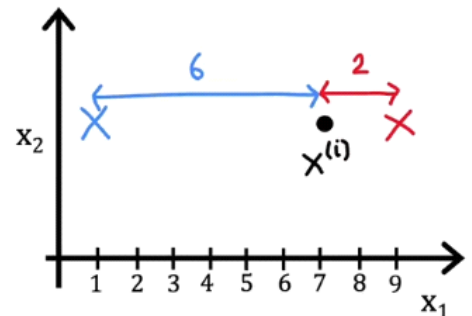
# Move cluster centroids

for  $k = 1$  to  $K$

$\mu_k :=$  average of points in cluster  $k$

}

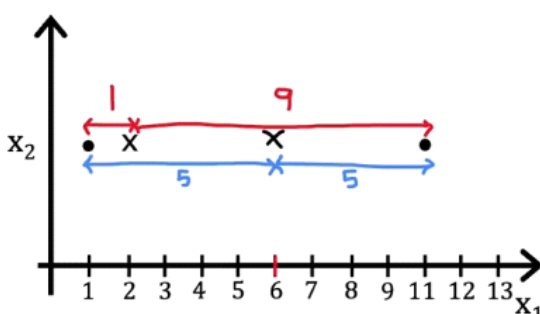
Cl 的值以尝试最小化 J。



DeepLearning.AI Stanford the values for Cl to try to minimize J.

Andrew Ng

## Moving the centroid

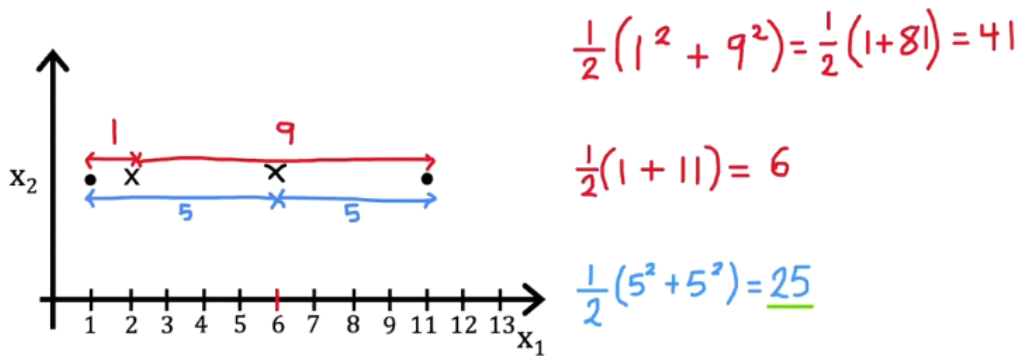


$$\frac{1}{2}(1^2 + 9^2) = \frac{1}{2}(1 + 81) = 41$$

$$\frac{1}{2}(1 + 11) = 6$$

$$\frac{1}{2}(5^2 + 5^2) = 25$$

# Moving the centroid



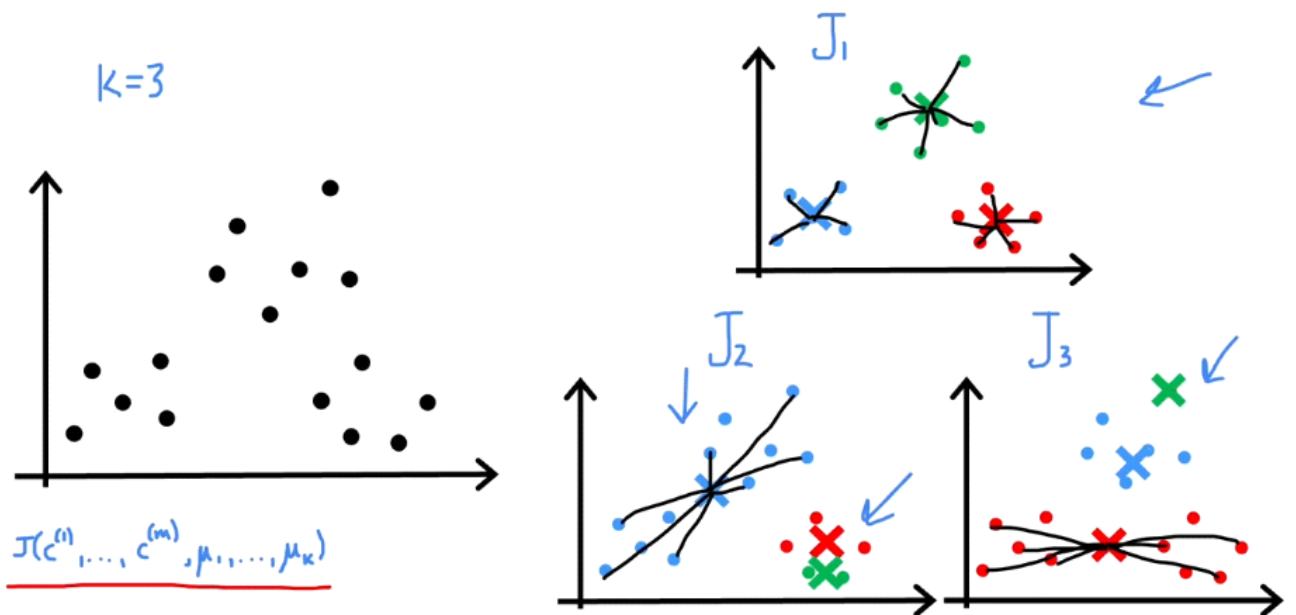
屏幕剪辑的捕获时间: 2023/5/31 11:26

## 3.K-means初始化

理由

K-means的初始点选的不一样, 分类结果可能有很大差别, 因为聚类毕竟不是有标准答案的任务, 怎么分都是有道理的

如果试图对m组数据进行聚类分析, 想要分出K个聚类, 那么取数据集中的K个数据点作为初始点, 用K-means进行聚类分析, 如此重复一定次数, 得到许多不同聚类方法, 最后根据最低成本原则, 选出最合适的聚类方法。



J 失真最小的选项。

the one with the smallest distortion of the smallest cost function  $J$  Ng

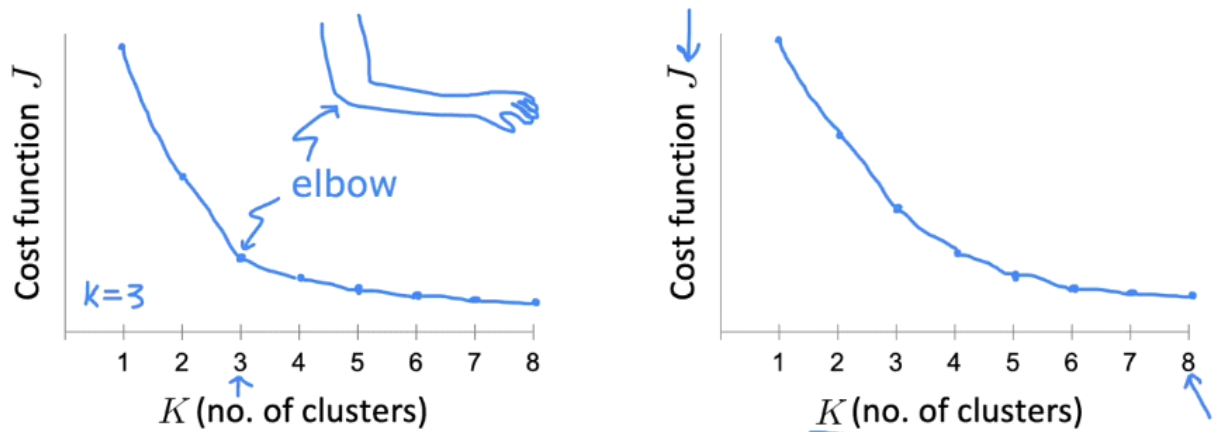
屏幕剪辑的捕获时间: 2023/6/2 16:13

选择聚类数量的方法

1.Elbow method (有些道理, 但用处有限)

## Choosing the value of $K$

Elbow method:



DeepLearning.AI

Stanford more clusters will pretty much

Andrew Ng

屏幕剪辑的捕获时间: 2023/6/2 16:19

不是所有案例都能出现这种“肘部”，而且只要 $K$ 变大， $J$ 一定会变小，没有明显突变时难以确定 $K$ 值。

2.根据实际需要选择聚类个数，比如调研衣服尺码与购买人群的关系。