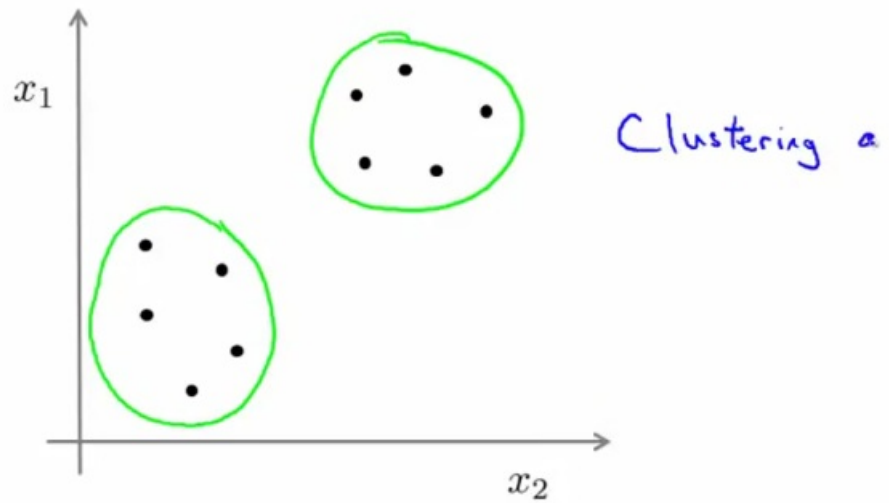


1. 没有任何标签的数据

Unsupervised learning



Training set: $\{\underline{x^{(1)}}, \underline{x^{(2)}}, x^{(3)}, \dots, \underline{x^{(m)}}\}$ ←

In unsupervised learning, you are given an unlabeled dataset and are asked to find "structure" in the data.

聚类算法：K均值算法（k-means algorithm）簇分配

分两类：

- （1）选两个聚类中心。分别计算离聚类中心距离来分类。
- （2）分别计算两个聚类的均值，然后重新选择两个聚类中心为这两个均值，重新分配聚类，（簇分配）。

K-means algorithm

Input:

- K (number of clusters) ←
- Training set $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ ←

$x^{(i)} \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

K-means algorithm

μ_1 μ_2
x x

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {

Cluster assignment step

for $i = 1$ to m

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

$\min_k \|x^{(i)} - \mu_k\|^2$
 \uparrow
 $c^{(i)}$

Move centroid

for $k = 1$ to K

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

$x^{(1)}, x^{(5)}, x^{(6)}, x^{(10)}$

$\rightarrow c^{(1)}=2, c^{(5)}=2, c^{(6)}=2, c^{(10)}=2$

$$\mu_2 = \frac{1}{4} [x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}] \in \mathbb{R}^n$$

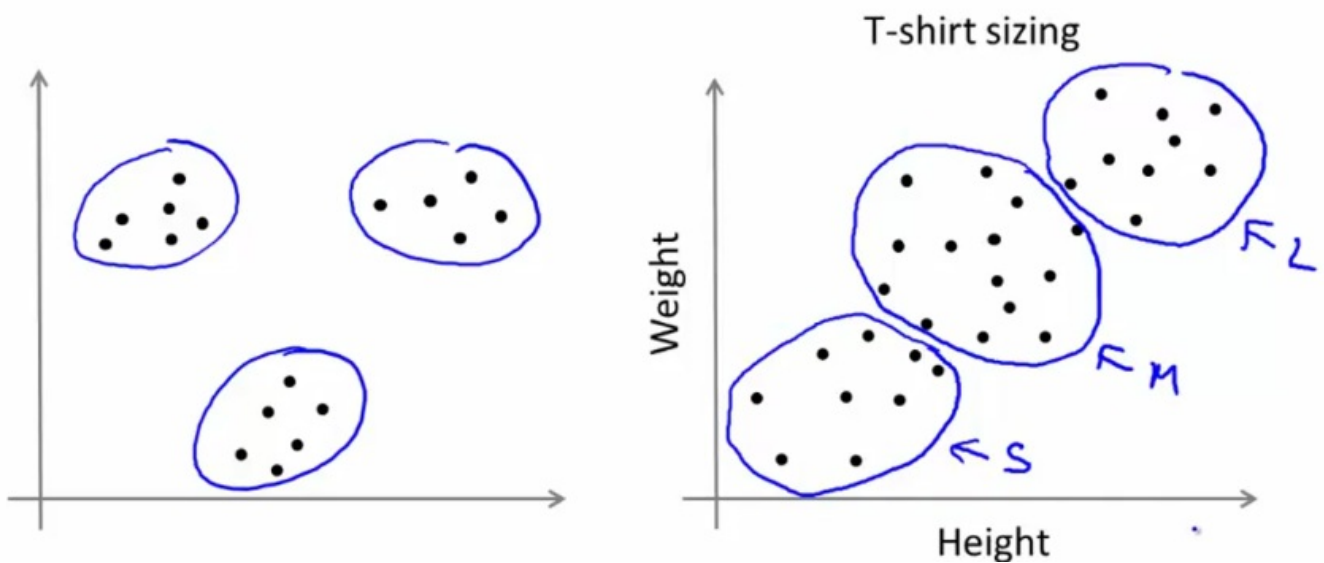
将 μ_2 移动到

如果有一个聚类中心没有被分配到点，那么通常直接移除那个聚类中心。这样就得到 $K-1$ 个簇。

不可分聚类上执行K均值算法：

K-means for non-separated clusters

S, M, L



K均值算法优化目标：各个样本点和他所属的聚类中心距离平方之和最小。

K-means optimization objective

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

→ μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$)

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

$x^{(i)} \rightarrow 5$

$c^{(i)} = 5$

$\mu_{c^{(i)}} = \mu_5$

K

$k \in \{1, 2, \dots, K\}$

Optimization objective:

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

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

有时候也叫做
Distortion



K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {
Cluster assignment step
Minimize $J(\dots)$ w.r.t $c^{(1)}, c^{(2)}, \dots, c^{(m)}$ ←
(holding μ_1, \dots, μ_K fixed)

for $i = 1$ to m

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

move centroid

for $k = 1$ to K

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

}

minimize $J(\dots)$ w.r.t μ_1, \dots, μ_K

随机初始化选择聚类中心：

Random initialization

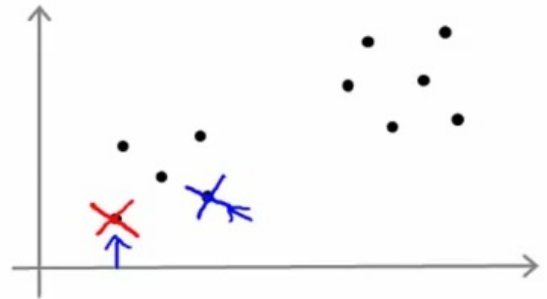
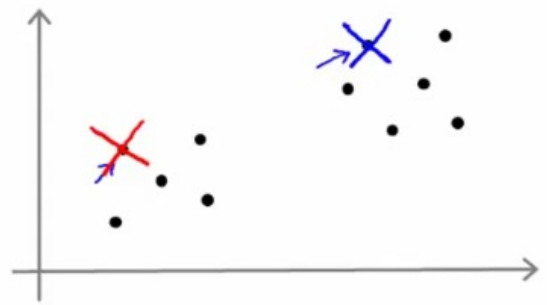
Should have $K < m$

$K=2$

Randomly pick K training examples.

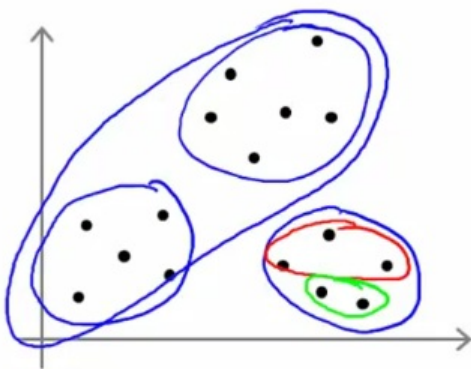
Set μ_1, \dots, μ_K equal to these K examples.

$$\begin{aligned}\mu_1 &= x^{(i)} \\ \mu_2 &= x^{(j)} \\ &\vdots\end{aligned}$$

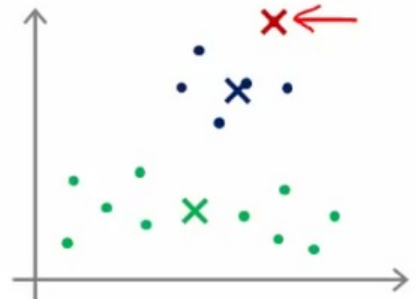
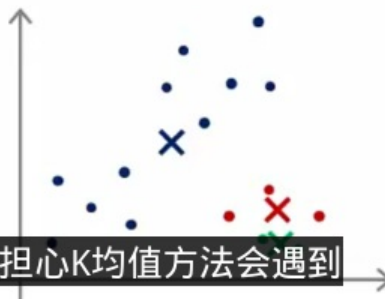
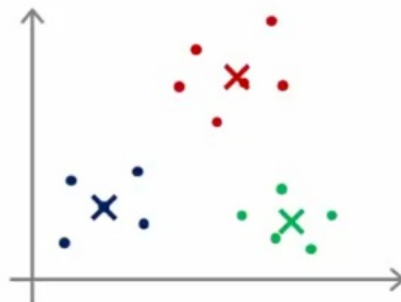


局部最优：

Local optima



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



如果你担心K均值方法会遇到

多次随机初始化，避免陷入局部最优

Random initialization

For $i = 1$ to 100 { 50 - 1000

→ Randomly initialize K-means.

Run K-means. Get $c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K$.

Compute cost function (distortion)

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

}

Pick clustering that gave lowest cost $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$

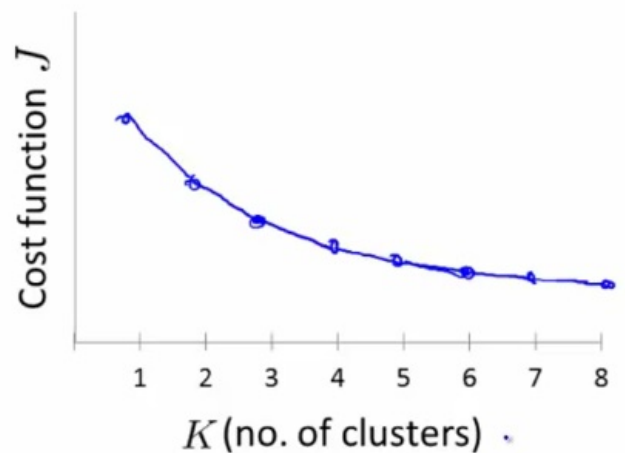
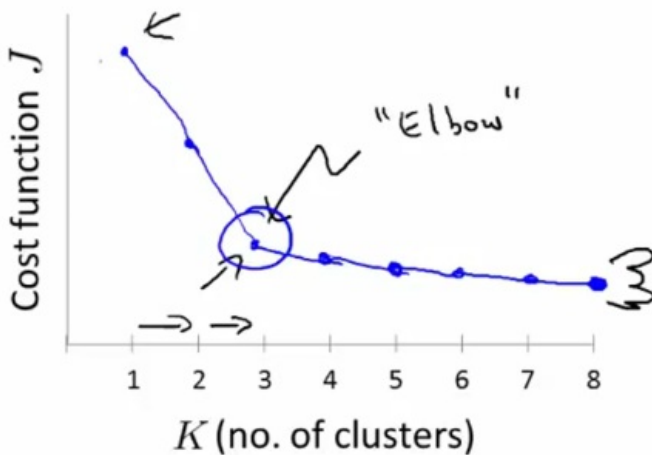
K = 2 - 10

保证你能找到更好的聚类数据

选择类型数K--肘部法则：

Choosing the value of K

Elbow method:



为了什么目的而选择聚类：

Choosing the value of K

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

$K=3$ S, M, L

$K=5$ XS, S, M, L, XL

E.g.

