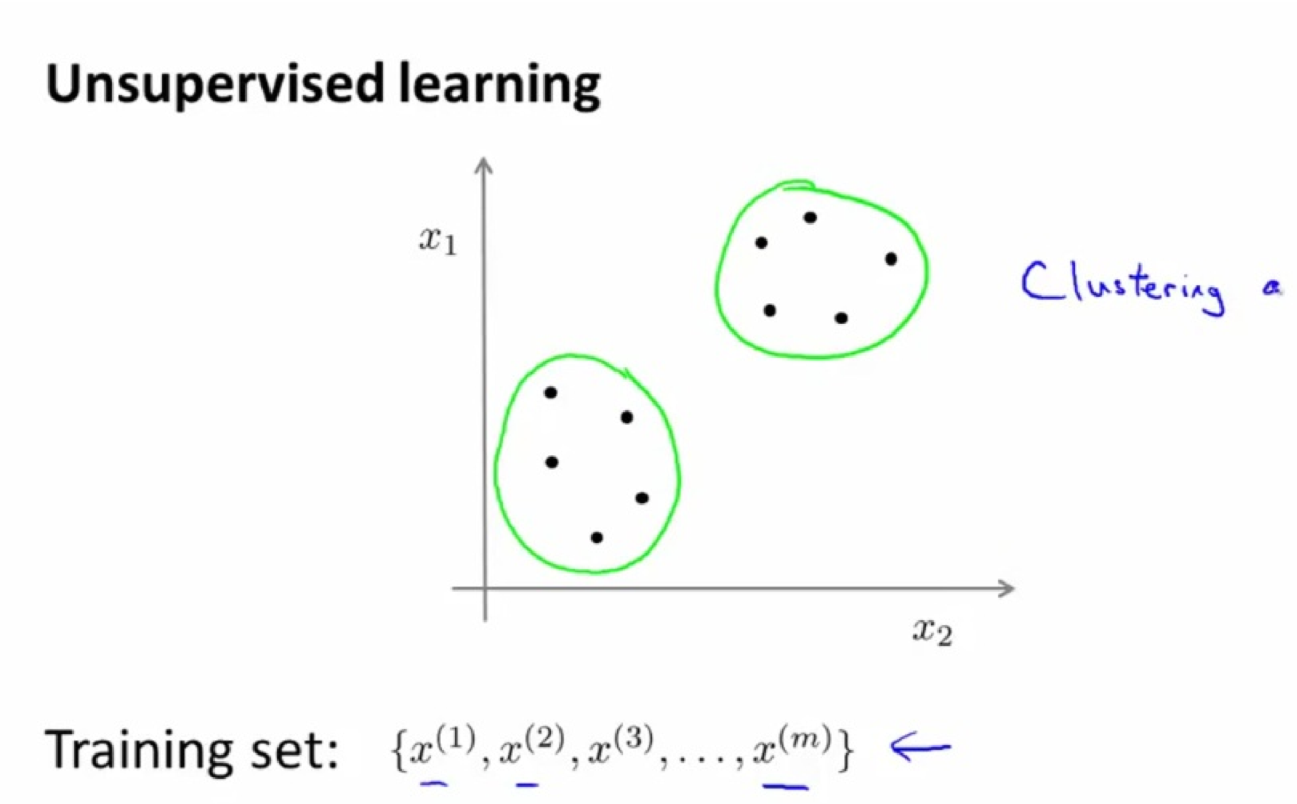
1. 没有任何标签的数据



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。

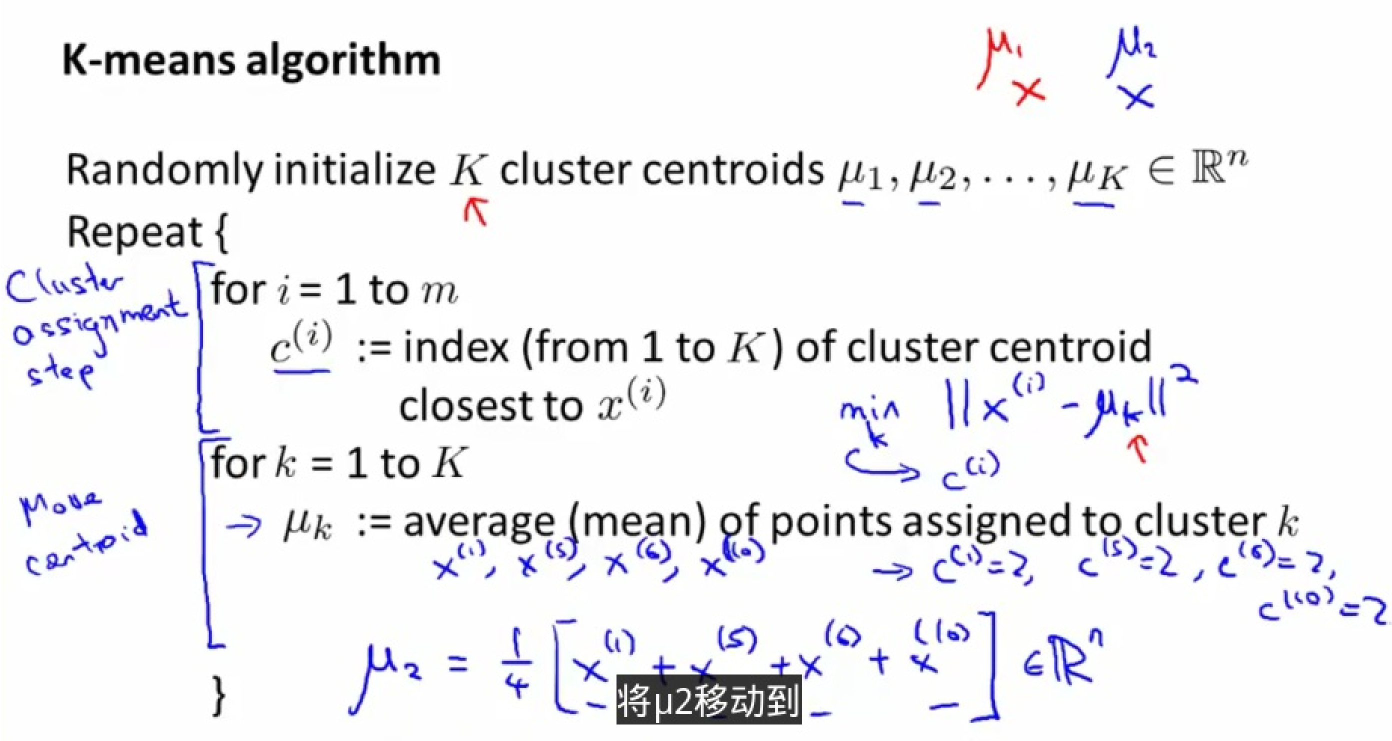
（2）分别计算两个聚类的均值，然后重新选择两个聚类中心为这两个均值，重新分配聚类，（簇分配）。

K-means algorithm

Input:

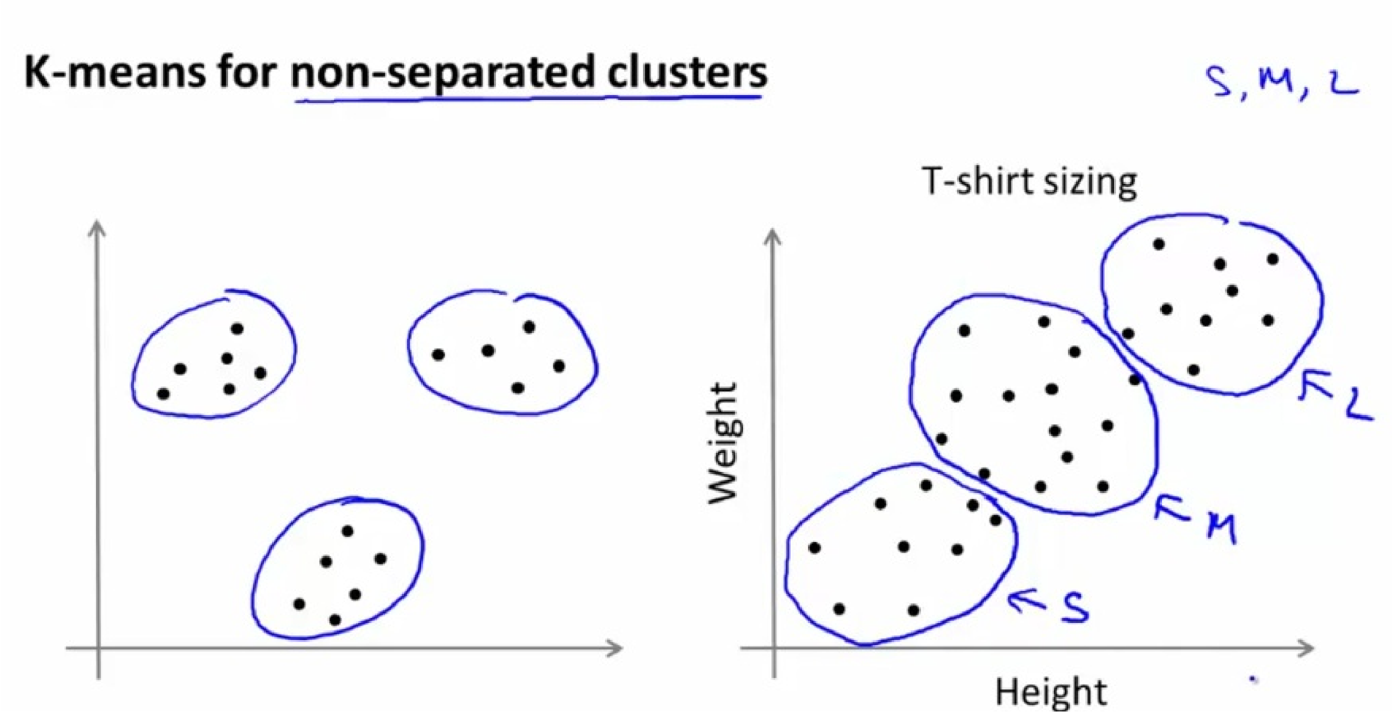
- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

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



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,...,K) to which example $x^{(i)}$ is currently assigned

→ μ_k = cluster centroid $\underline{k}$ ($\mu_k \in \mathbb{R}^n$) K $k \in \{1,2,\ldots,K\}$

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned $x^{(i)} \rightarrow \underline{5}$ $\underline{c^{(i)} = 5}$ $\underline{\mu_{c^{(i)}}} = \mu_5$

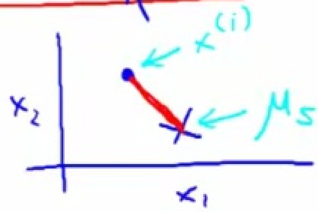
Optimization objective:

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

$$\min_{\substack{\rightarrow c^{(1)}, \ldots, c^{(m)}, \\ \rightarrow \mu_1, \ldots, \mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

Distortion

有时候也叫做



K-means algorithm

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

Cluster assignment step

Minimize $J(\ldots)$ wrt $\boxed{c^{(1)}, c^{(2)}, \ldots, c^{(m)}}$ ←

(holding $\mu_1, \ldots, \mu_K$ fixed)

Repeat {

for i = 1 to m

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

move centroid

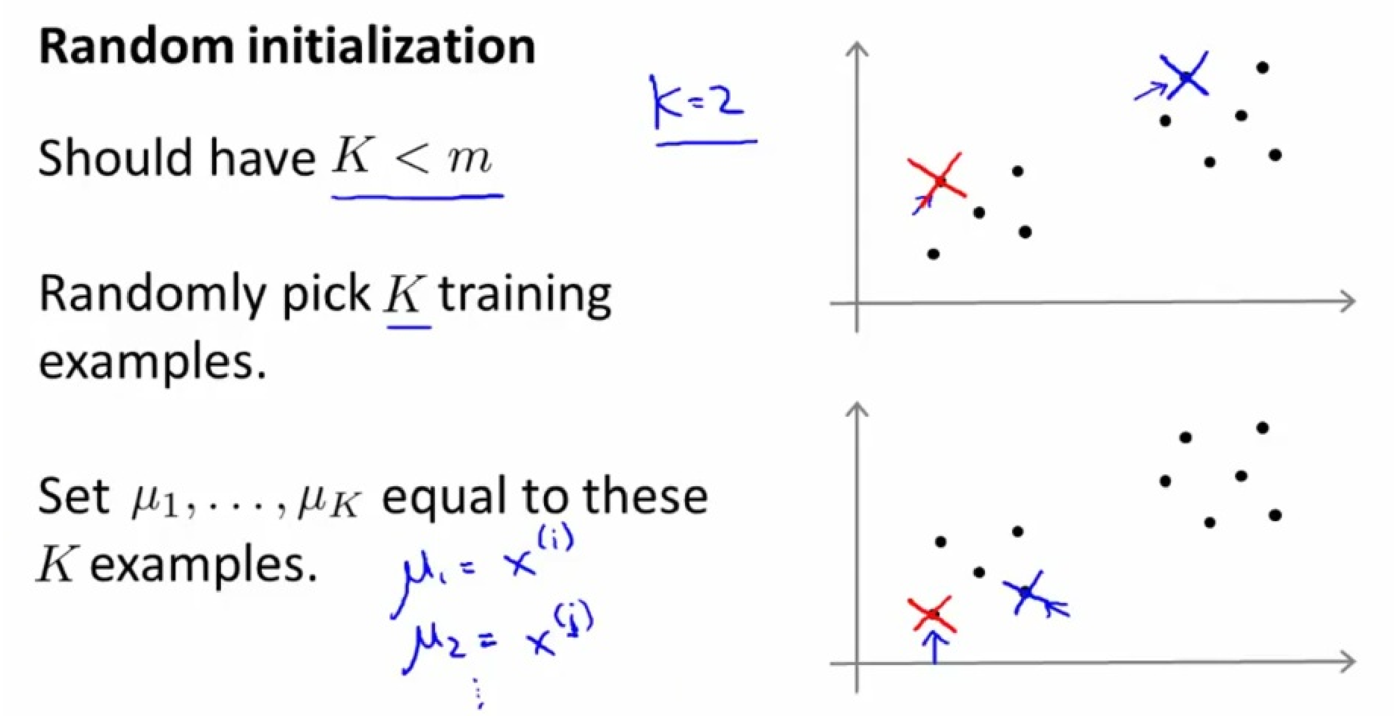
for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

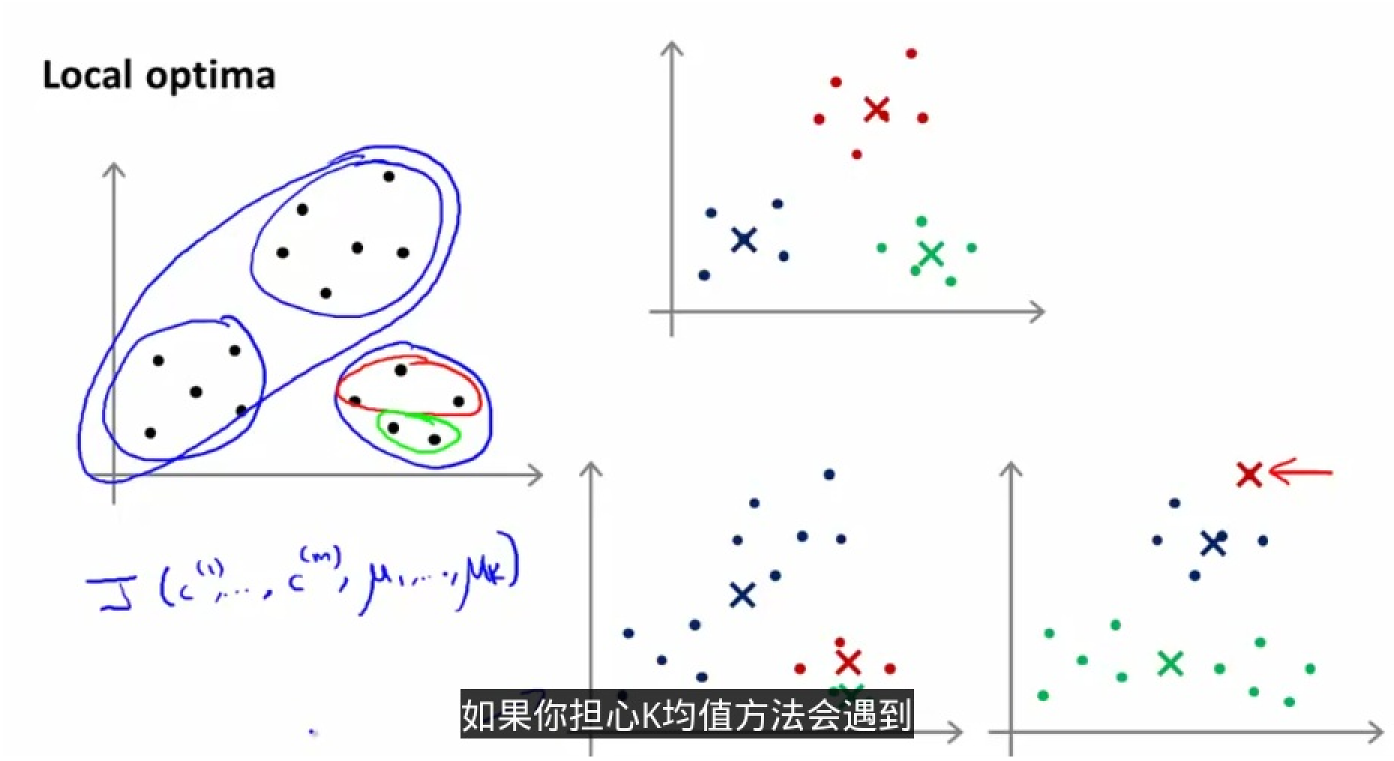
}

minimize $J(\ldots)$ wrt $\boxed{\mu_1, \ldots, \mu_K}$

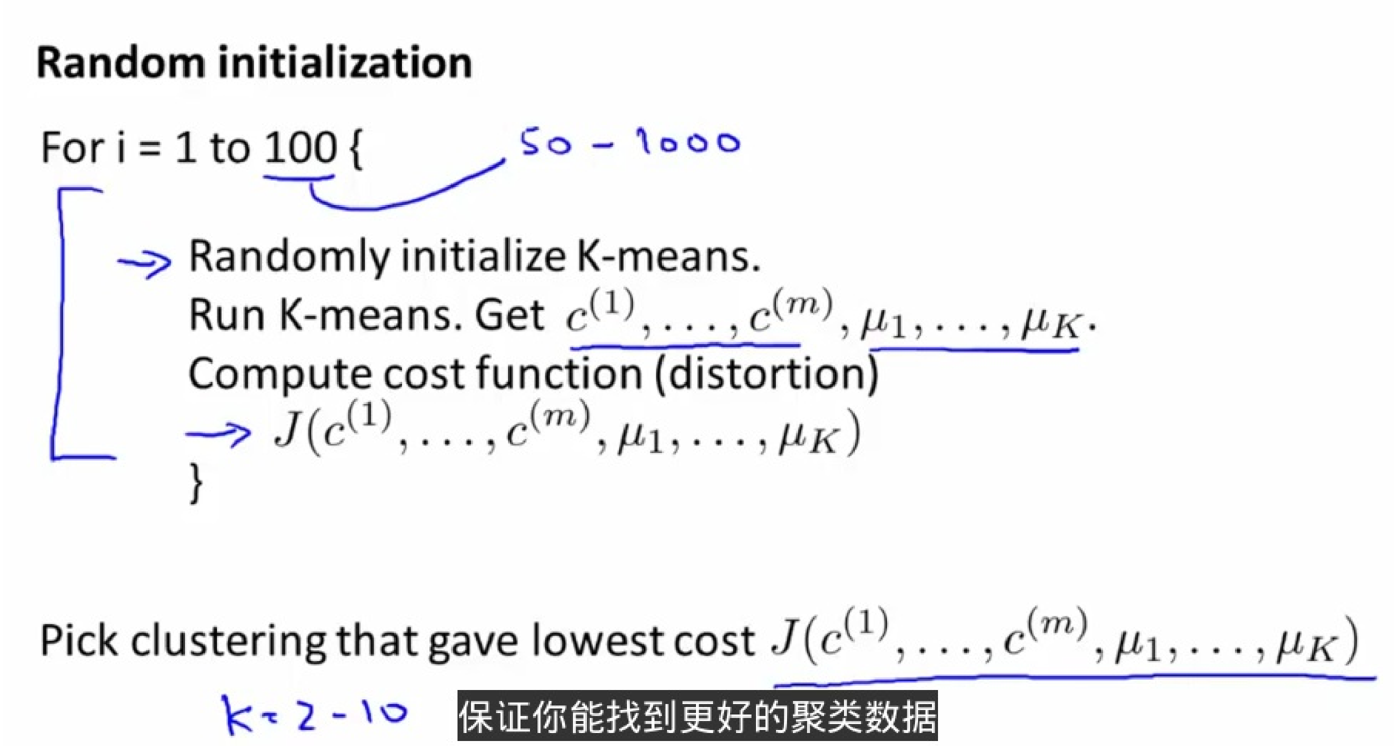
随机初始化选择聚类中心：



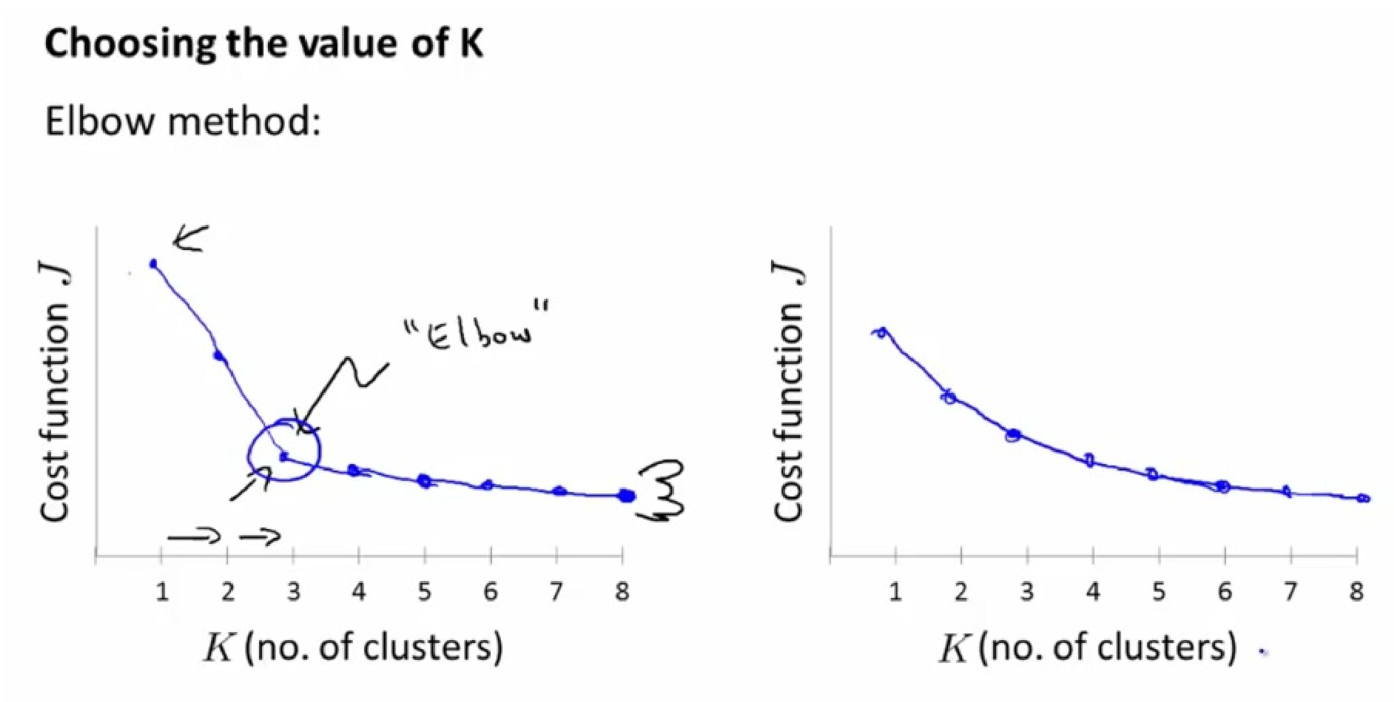
局部最优：



多次随机初始化，避免陷入局部最优



选择类型数K--肘部法则：



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.