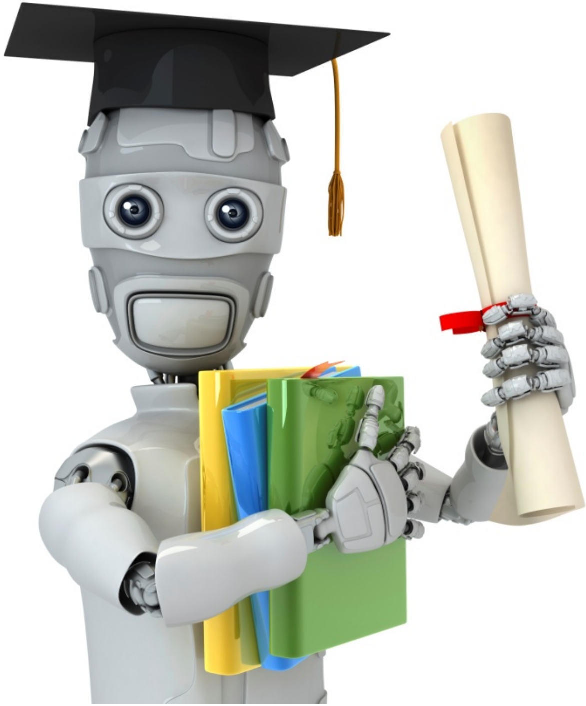
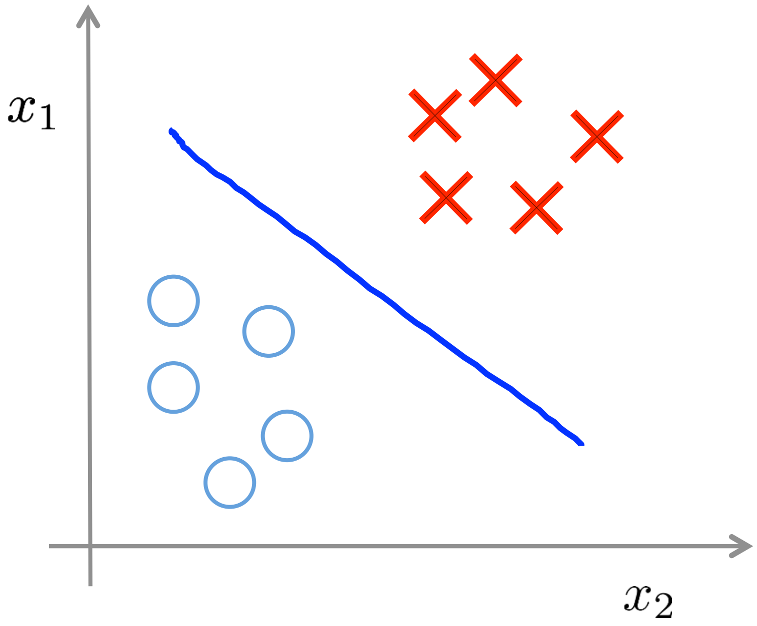


Clustering

Unsupervised learning introduction

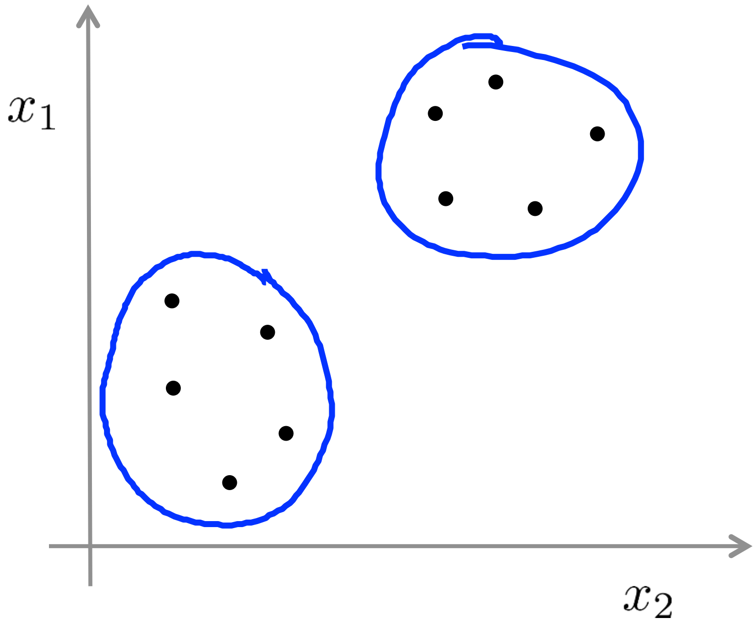
Machine Learning

Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

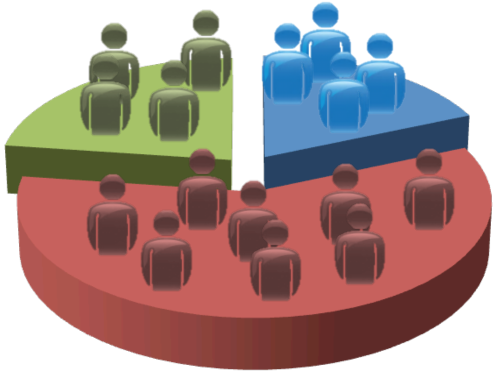
Unsupervised learning



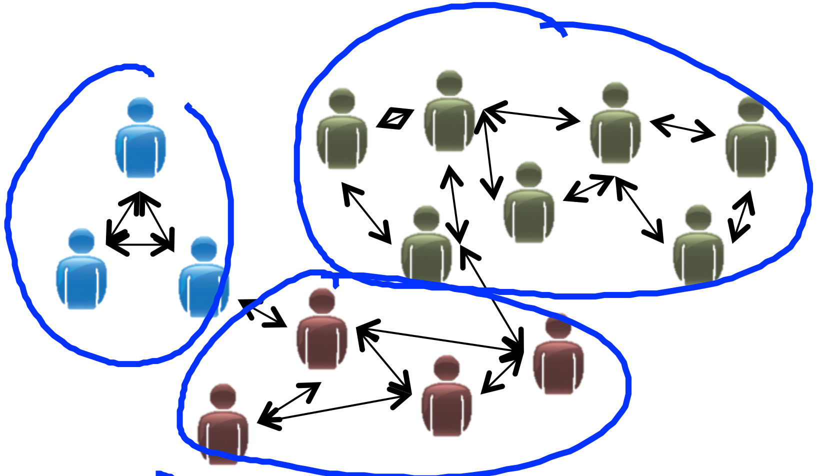
Clustering algorithm

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

Applications of clustering

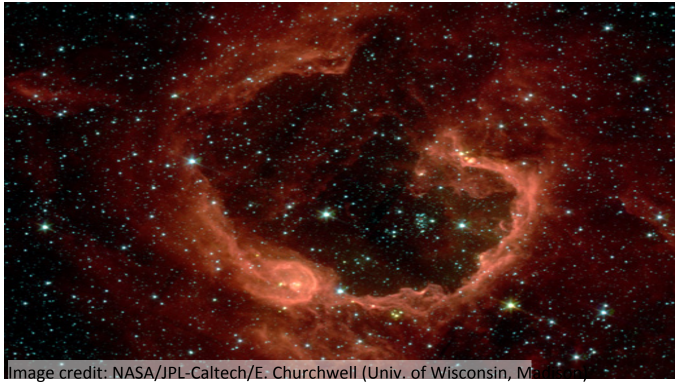


Market segmentation



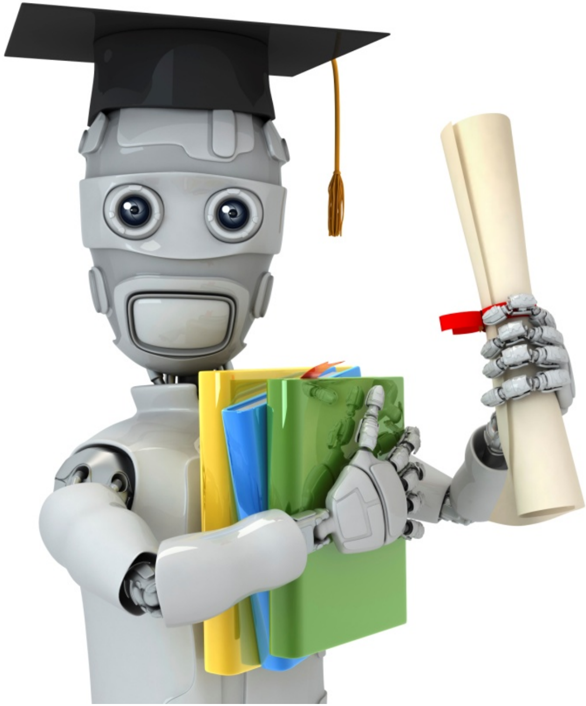
Social network analysis

Organize computing clusters



Astronomical data analysis

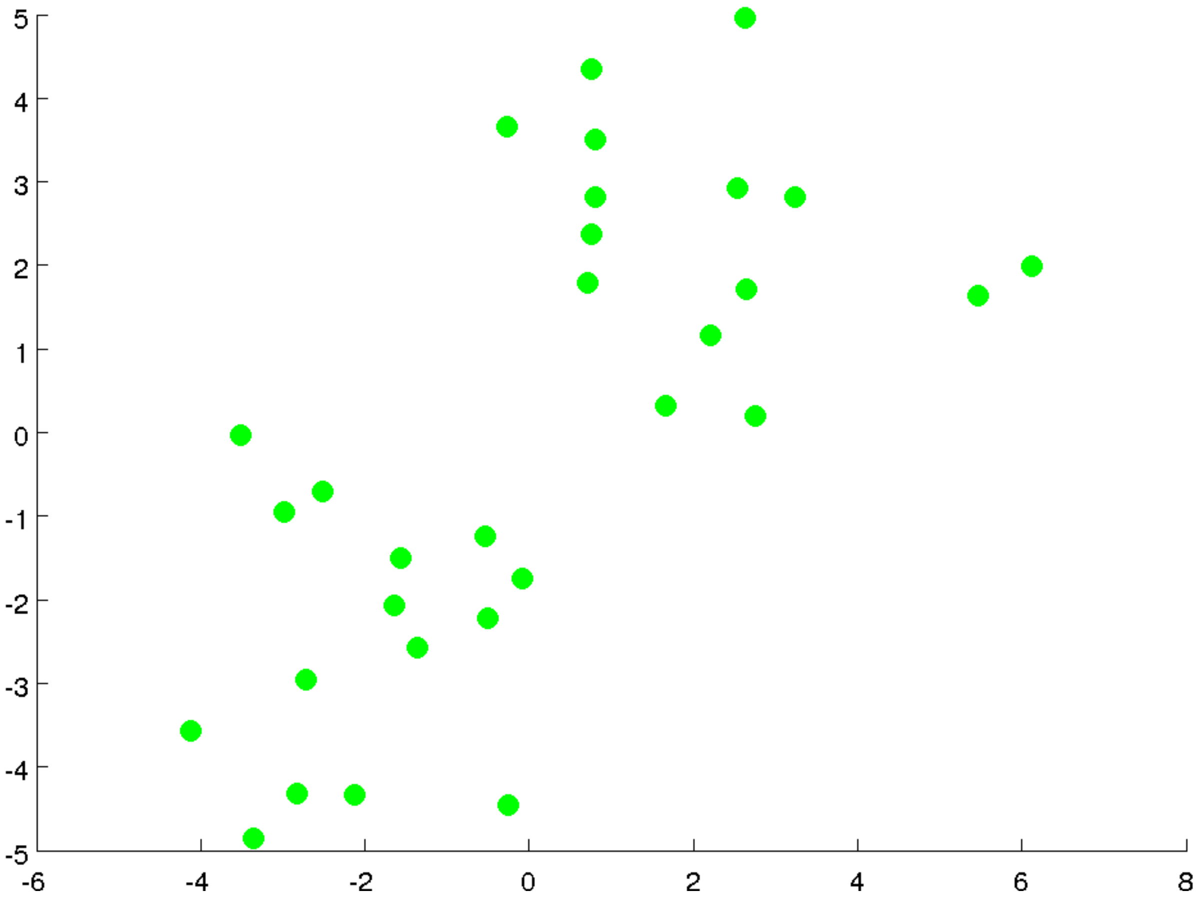
Andrew Ng



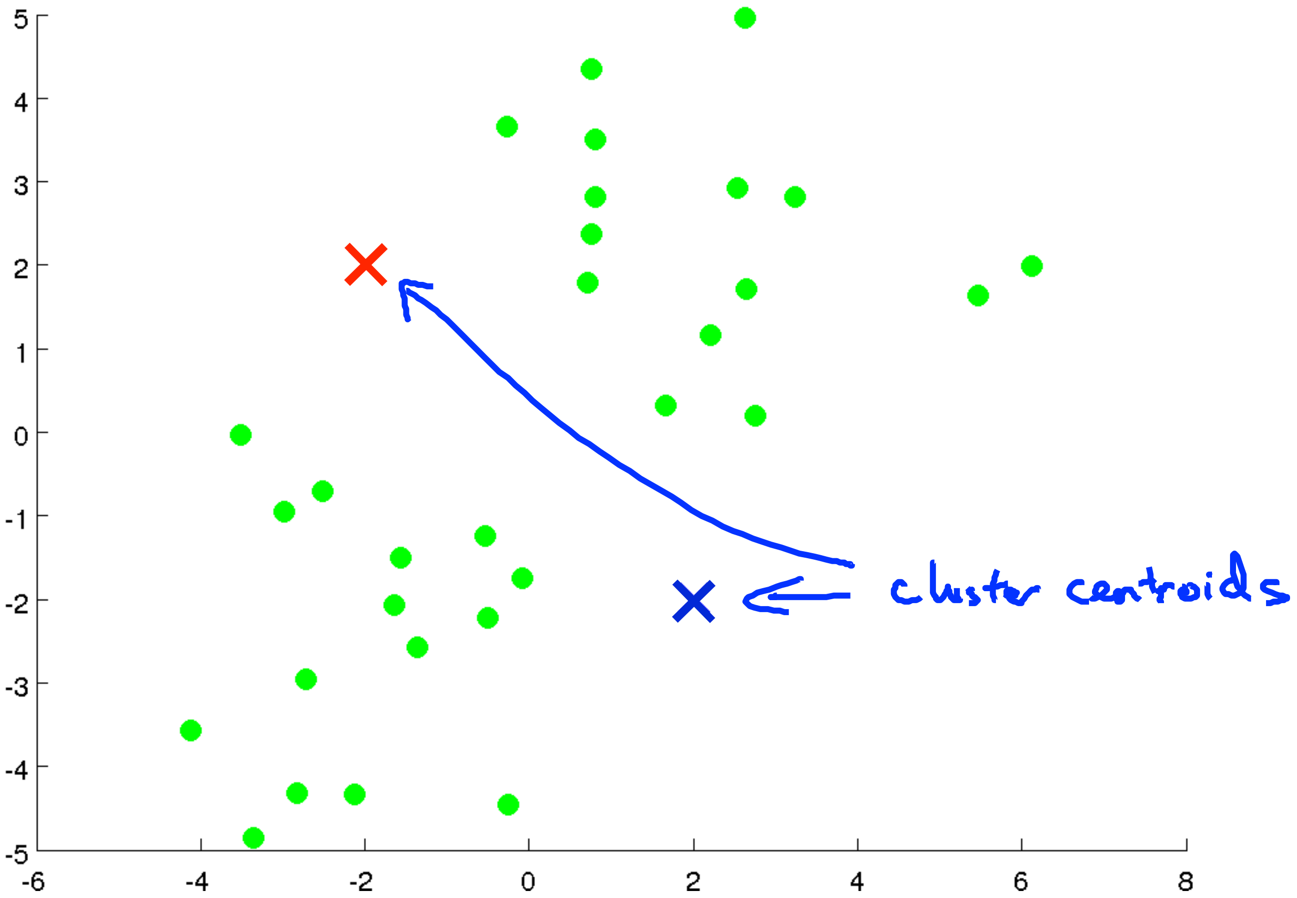
Machine Learning

Clustering

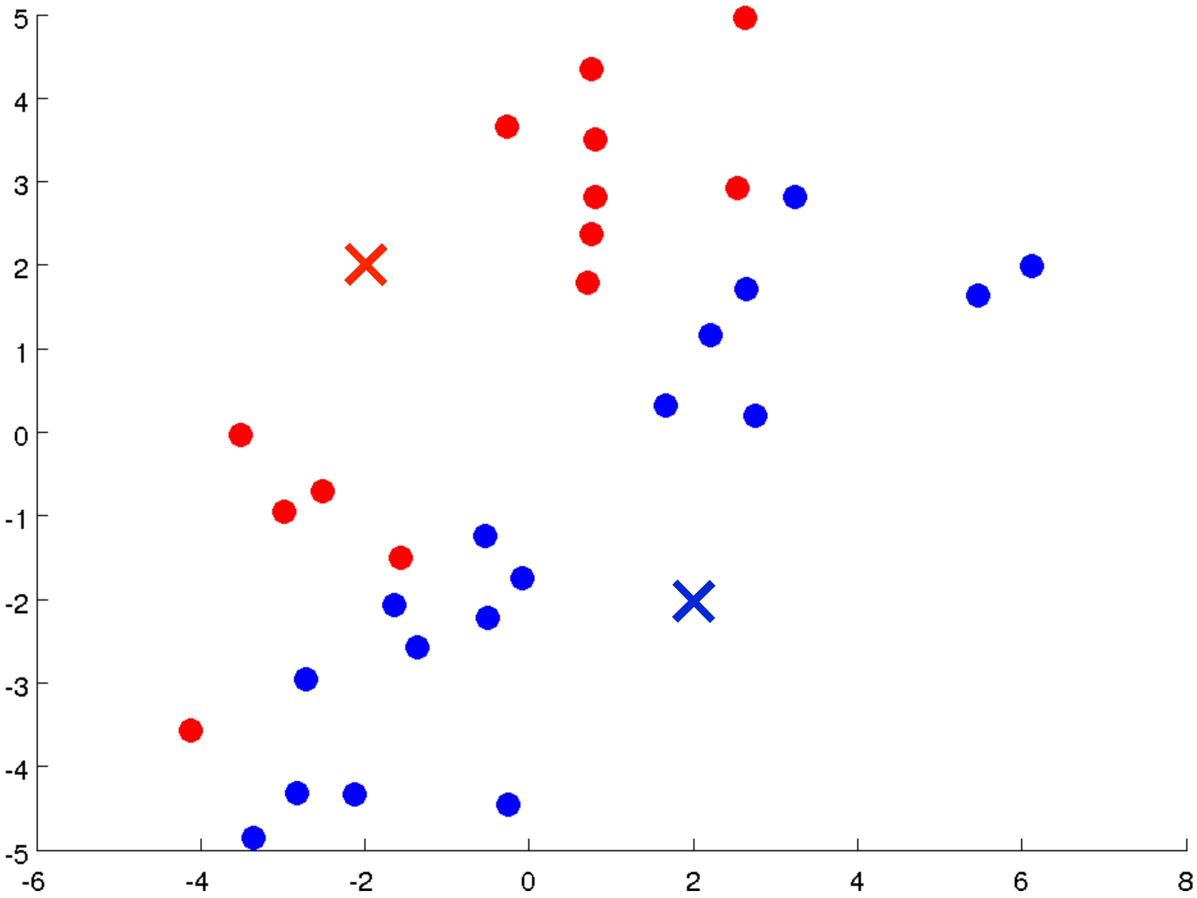
K-means algorithm



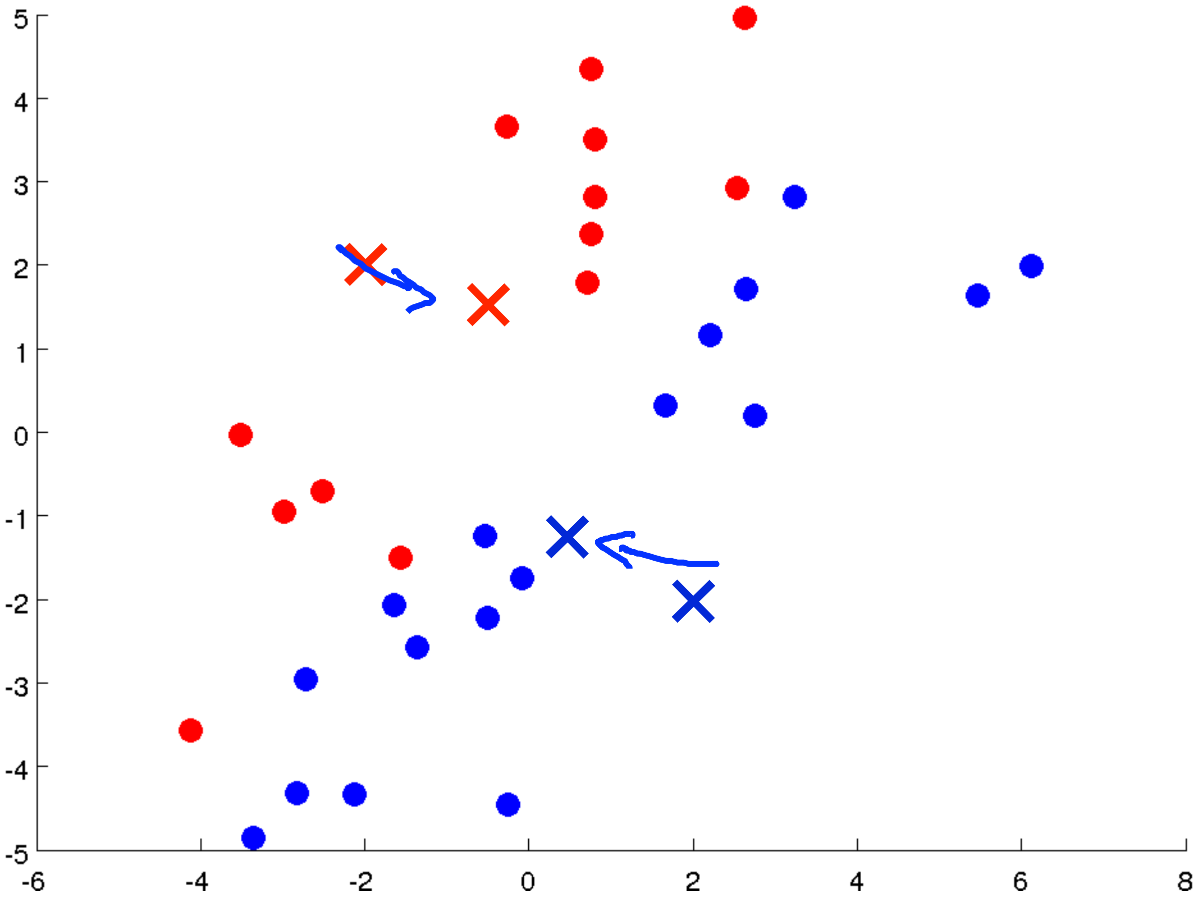
Andrew Ng



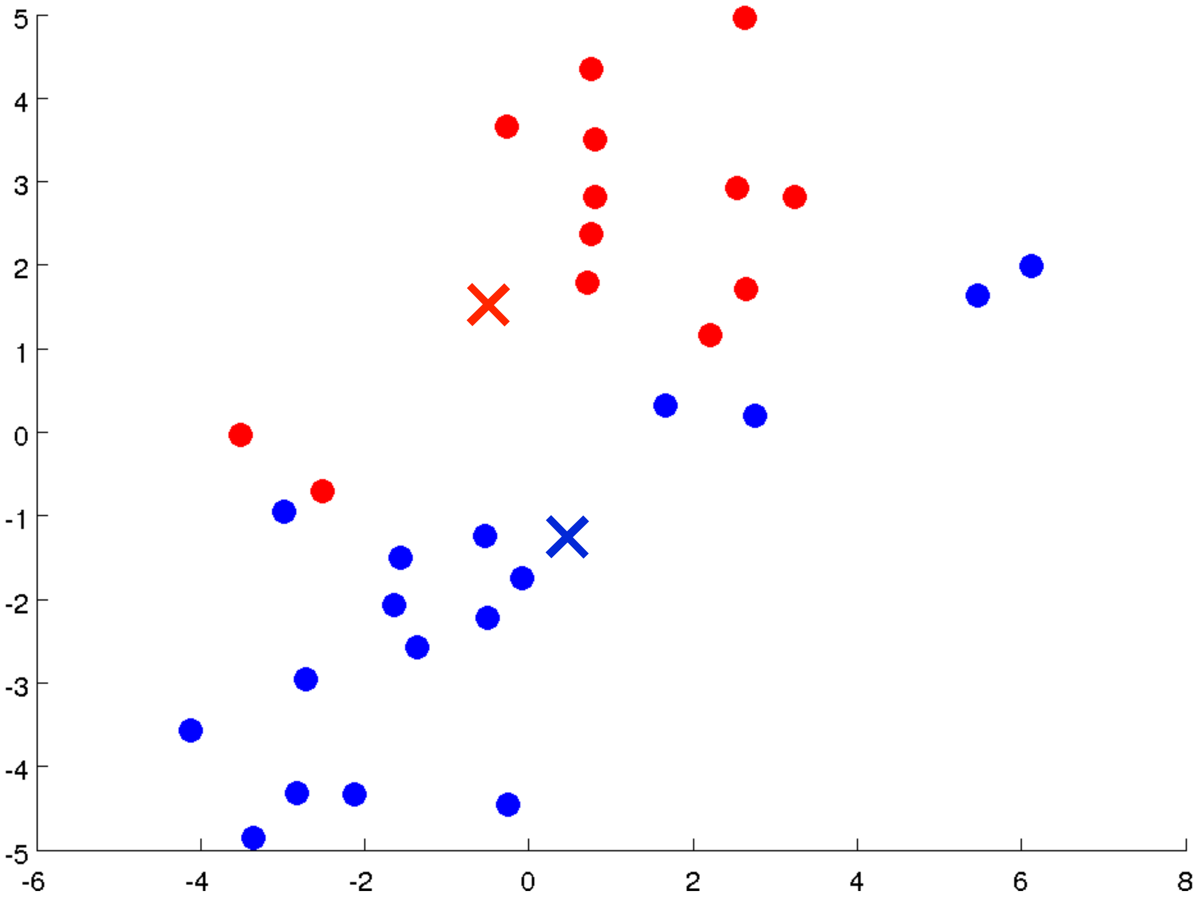
cluster centroids



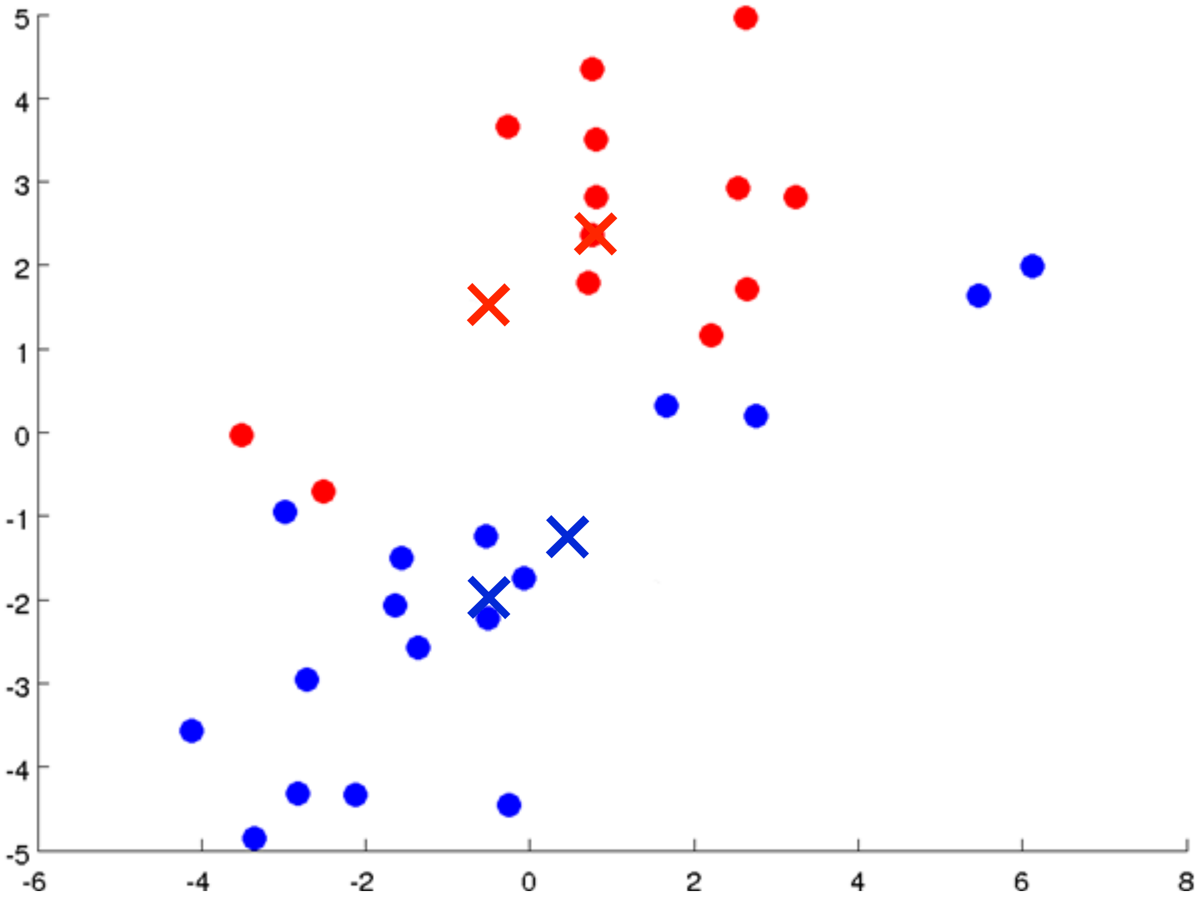
Andrew Ng



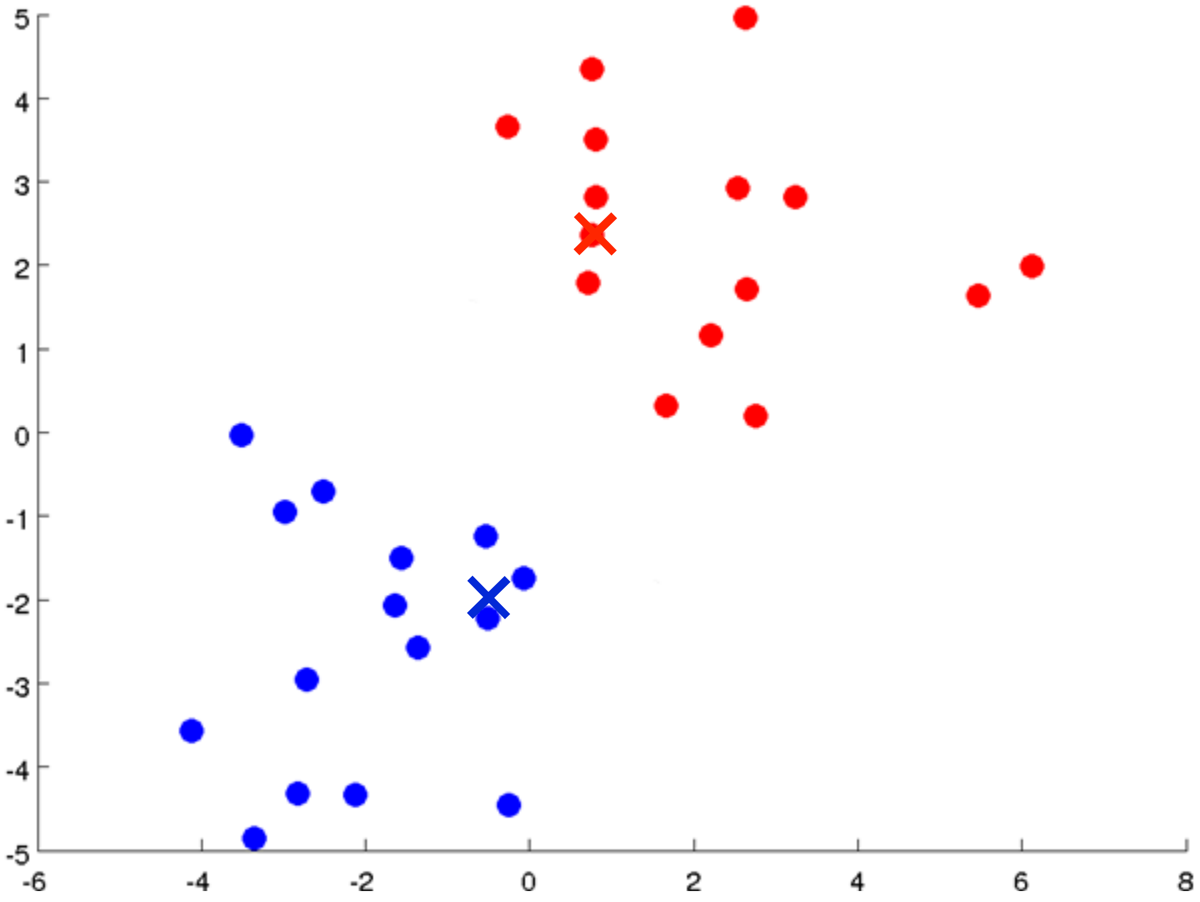
Andrew Ng



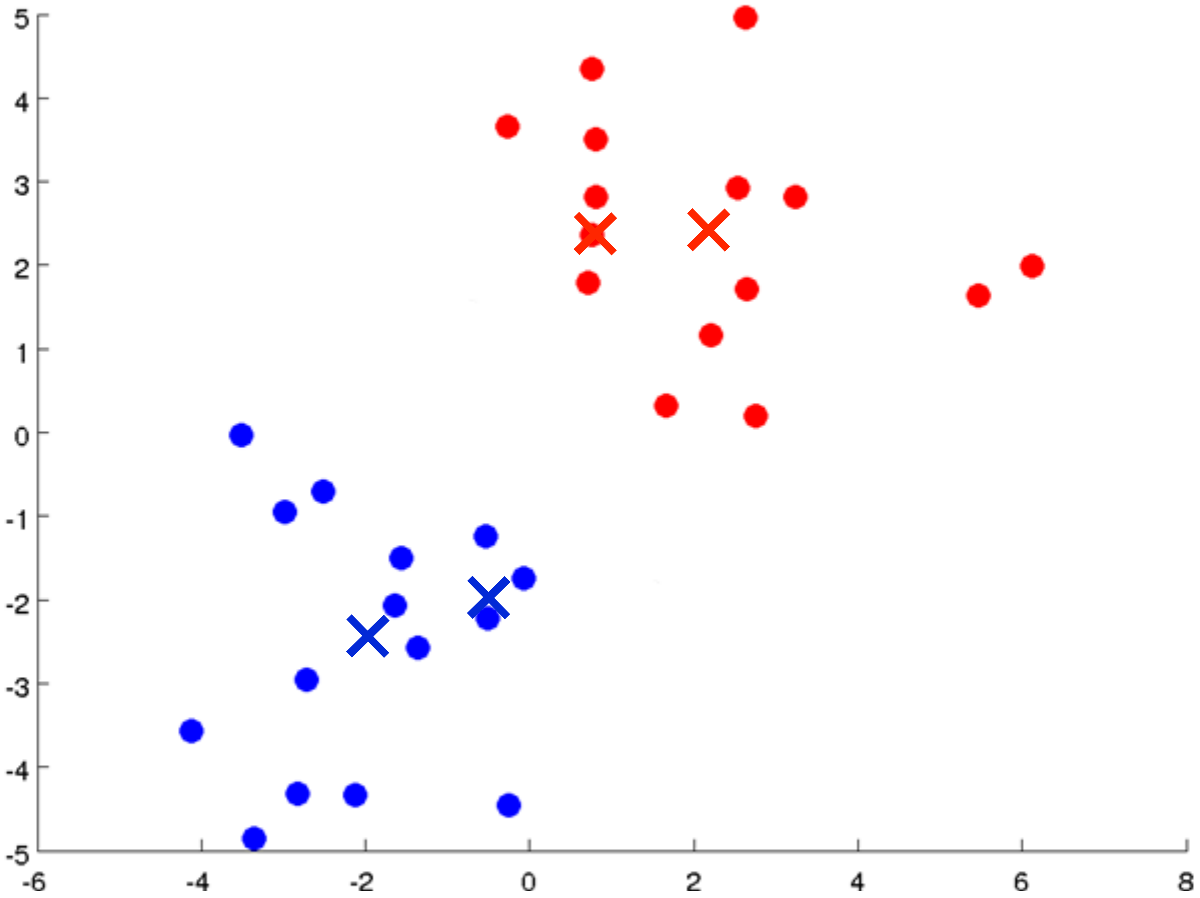
Andrew Ng



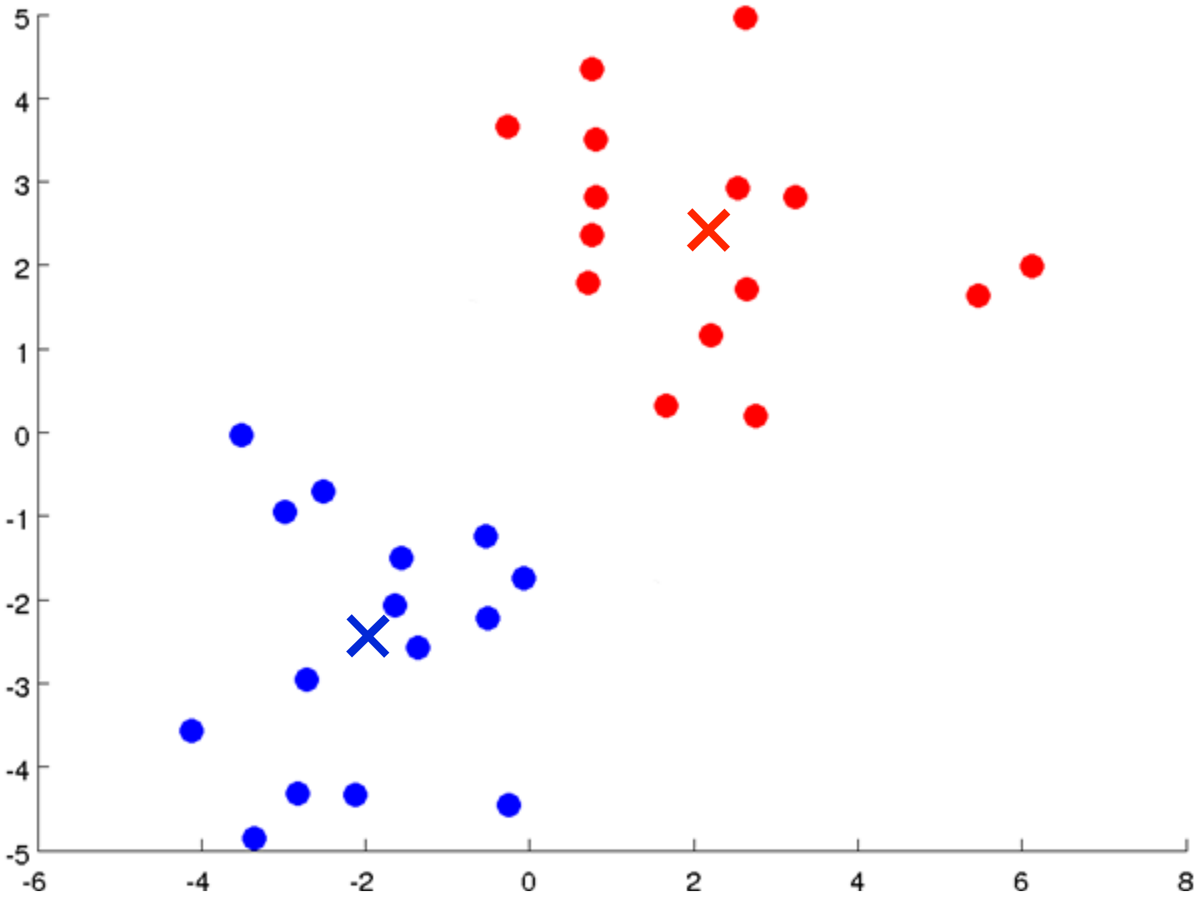
Andrew Ng



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



Andrew Ng



Andrew Ng

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)

K-means algorithm

μ_1 × μ_2 ×

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 $x^{(i)}$

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

$k \rightarrow c^{(i)}$

Move centroid

for k = 1 to K

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

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

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

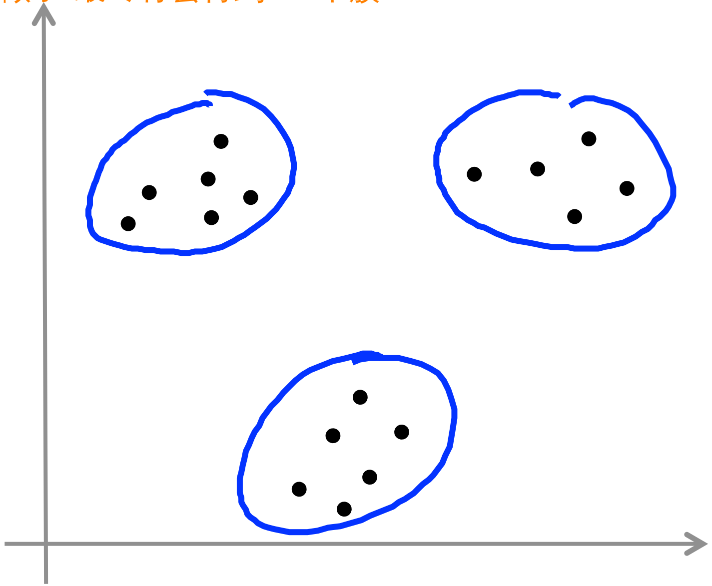
}

Andrew Ng

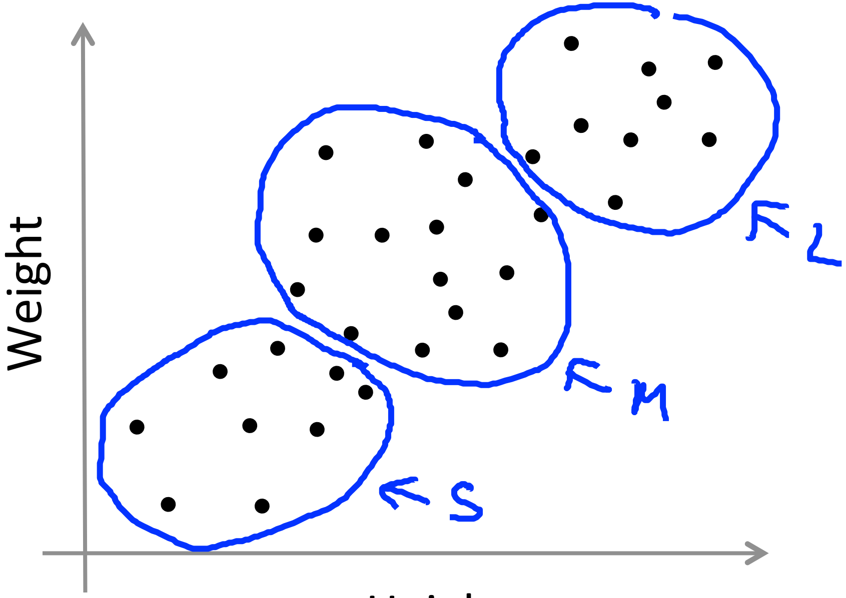
K-means for non-separated clusters

S, M, L

我要问的问题是 既然我们要让μk移动到分配给它的那些点的均值处 那么如果 存在一个 没有点分配给它的聚类中心 那怎么办?
通常在这种情况下 我们就直接移除 那个聚类中心
如果这么做了 最终将会得到K-1个簇



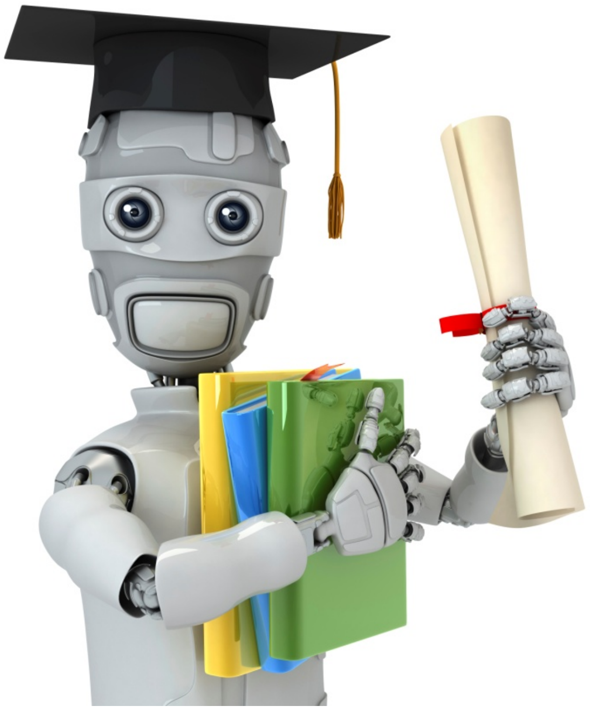
T-shirt sizing



如果就是要K个簇 不多不少 但是有个 没有点分配给它的聚类中心
你所要做的是 重新随机找一个聚类中心 但是直接移除那个中心
是更为常见的方法 当你遇到了一个 没有分配点的 聚类中心
不过在实际过程中 这个问题不会经常出现

假设你是T恤制造商 你找到了一些人 想把T恤卖给他们 然后
你搜集了一些 这些人的 身高和体重的数据 我猜 身高体重更
重要一些 然后你可能 收集到了这样的样本 一些关于 人们
身高和体重的样本 就像这个图所表示的 然后你想确定一

下T恤的大小 假设我们要设计 三种不同大小的t恤 小号 中号和大号 那么小号应该是多大的? 中号呢? 大号呢?



Machine Learning

Clustering

Optimization objective

K-means optimization objective

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

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

K $\quad k \in \{1,2,\ldots,K\}$

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

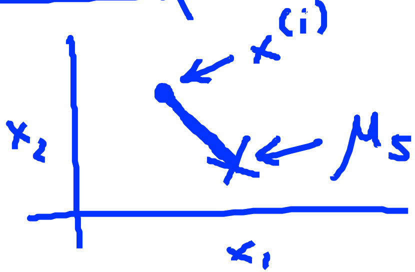
$x^{(i)} \to 5 \qquad c^{(i)} = 5 \qquad \mu_{c^{(i)}} = \mu_5$

Optimization objective:

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

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

Distortion



K均值算法中 有时候也叫做distortion cost function

Andrew Ng

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 $x^{(i)}$

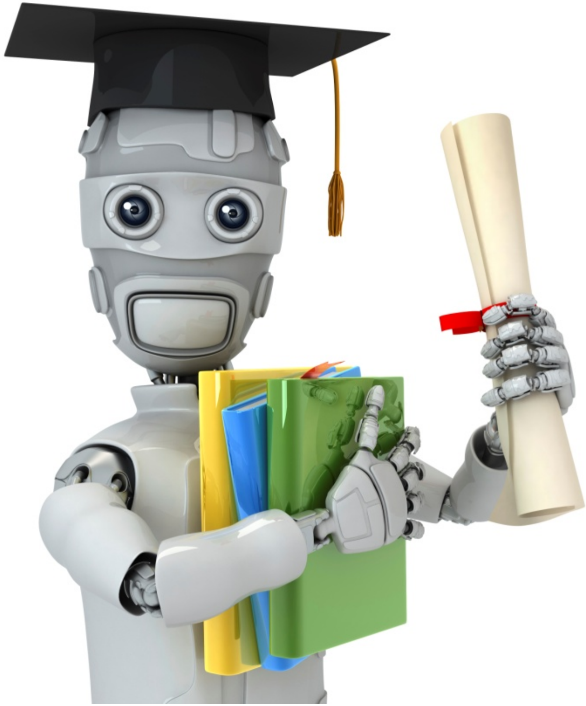
move centroid

for k = 1 to K

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

}

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



Machine Learning

Clustering

Random initialization

K-means algorithm

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

Repeat {

for i = 1 to m

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

for k = 1 to K

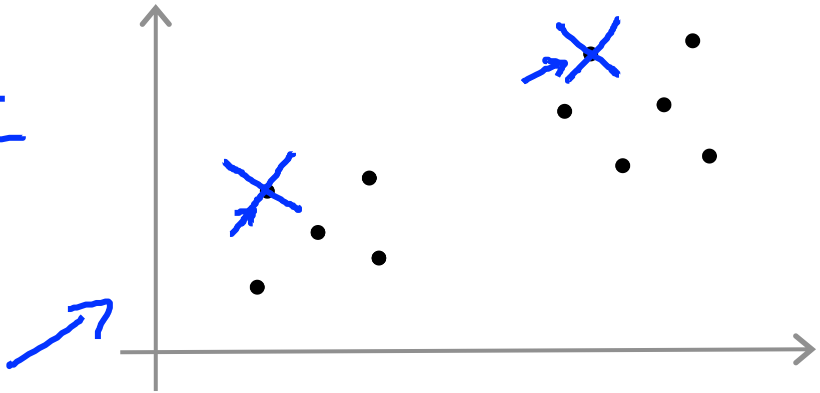
μ_k := average (mean) of points assigned to cluster k

}

Random initialization

$K=2$

Should have $K < m$



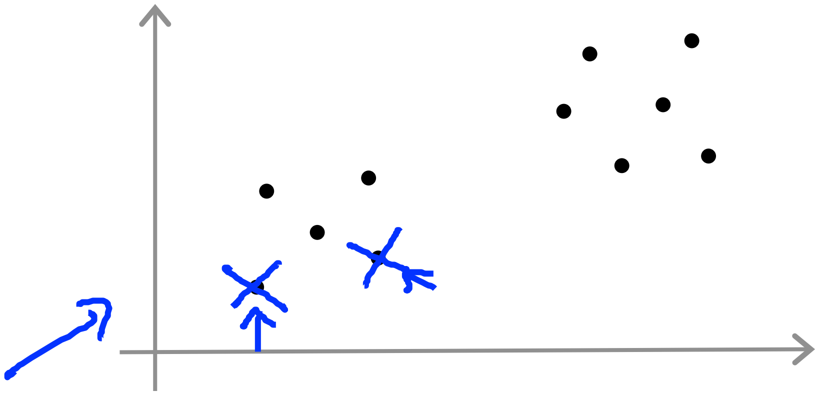
Randomly pick K training examples.

Set $\mu_1, \dots, \mu_K$ equal to these K examples.

$\mu_1 = x^{(i)}$

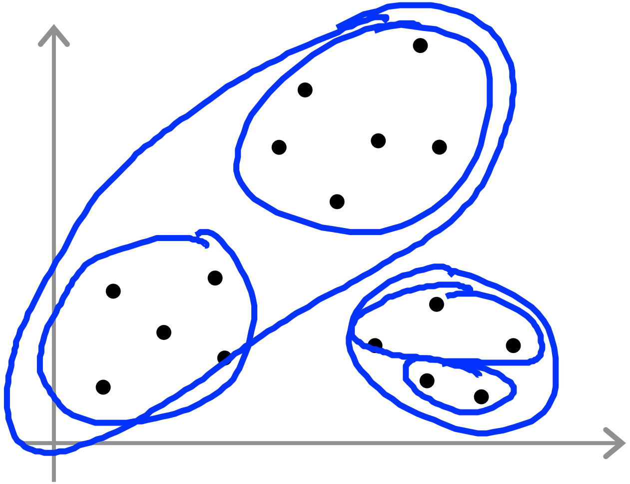
$\mu_2 = x^{(j)}$

$\vdots$

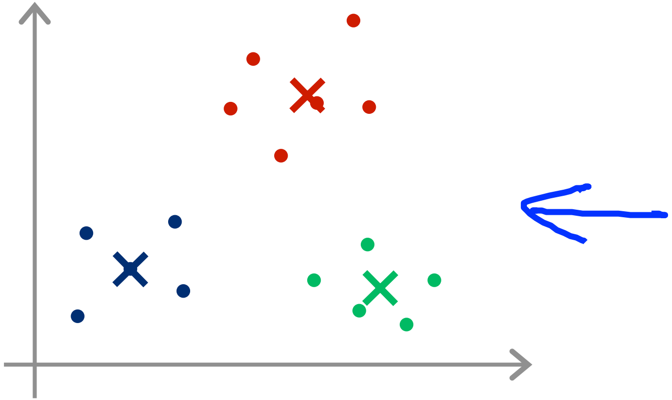


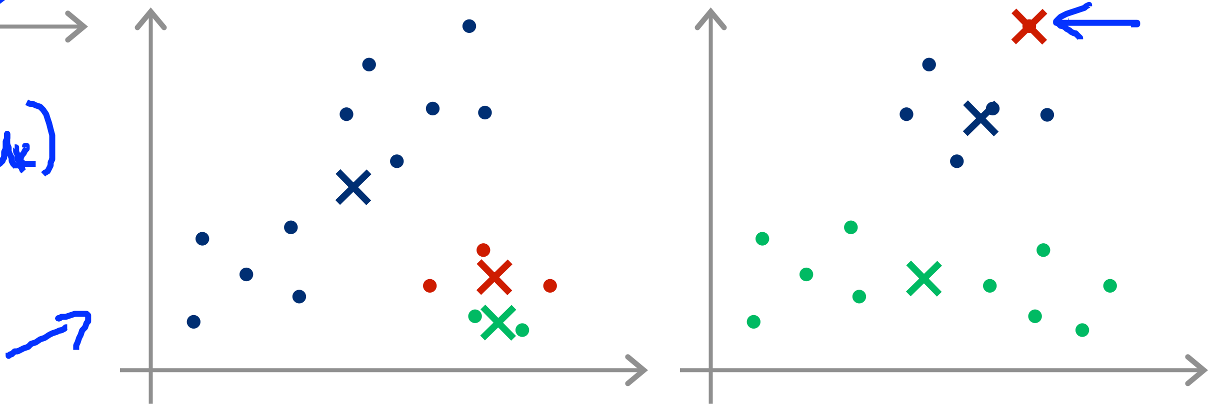
K均值方法 最终可能会得到 不同的结果 取决于 聚类簇的初始化方法 因此也就取决于随机的初始化
尤其是如果K均值方法落在局部最优的时候

Local optima



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





我们能做的是 初始化K均值很多次 并运行K均值方法很多次

Random initialization

For i = 1 to 100 {

Randomly initialize K-means.

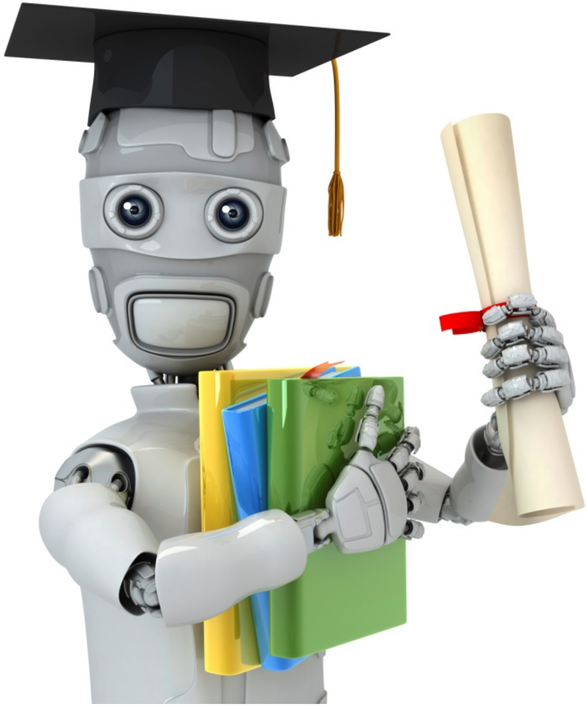
Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.

Compute cost function (distortion)

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

}

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

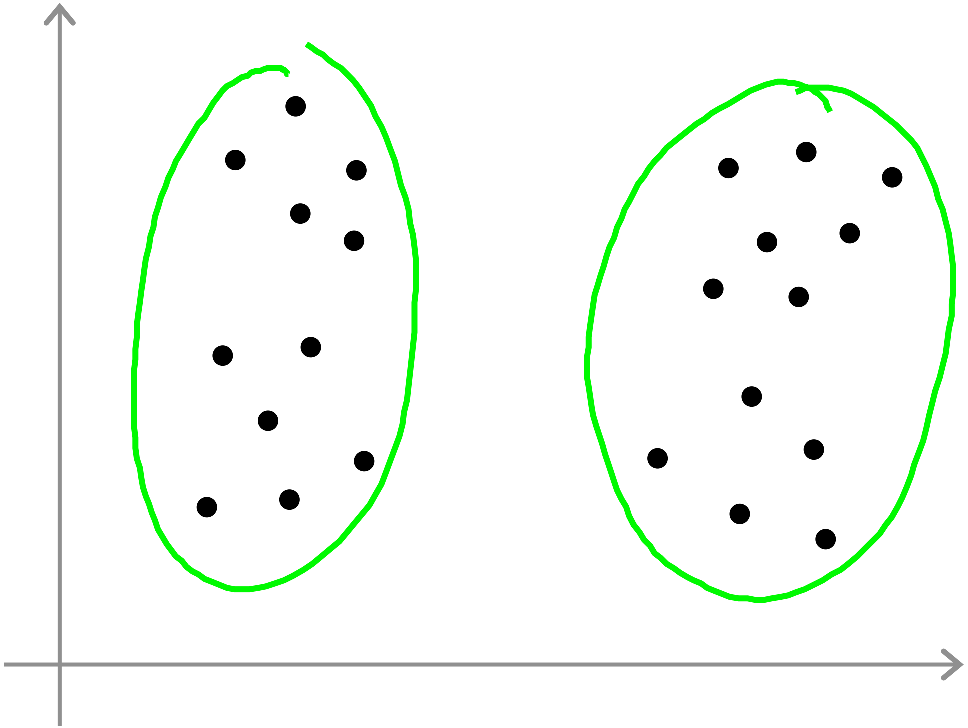


Machine Learning

Clustering

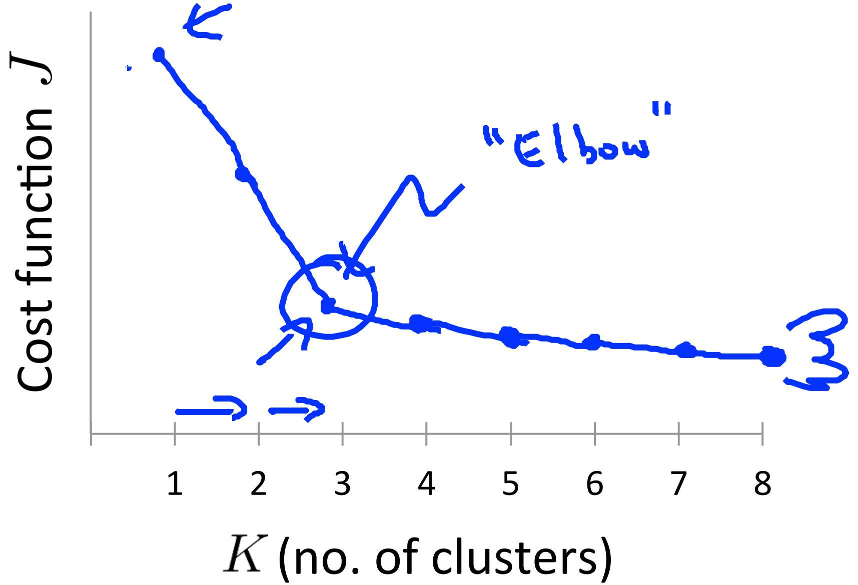
Choosing the number of clusters

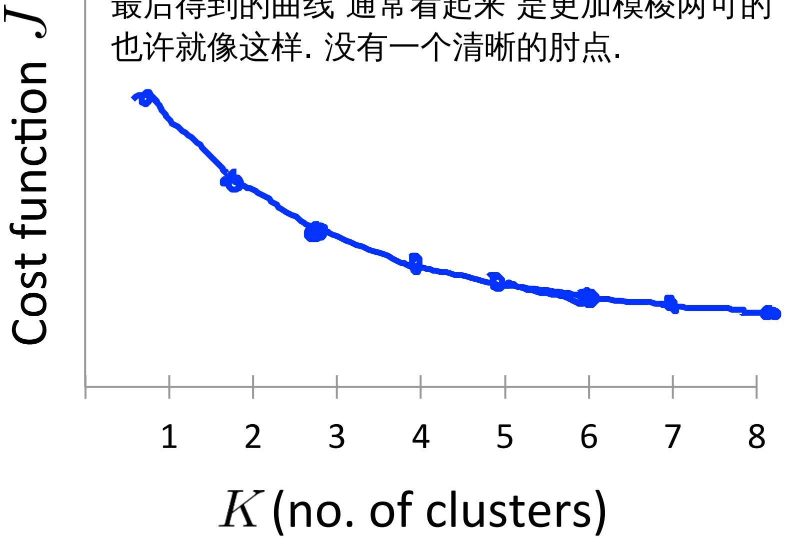
What is the right value of K?



Choosing the value of K

Elbow method:





而事实证明肘部法则 并不那么常用 其中一个原因是如果你把这种方法 用到一个聚类问题上 事实上你最后得到的曲线 通常看起来 是更加模棱两可的 也许就像这样. 没有一个清晰的肘点.

最后 有另外一种方法 来考虑如何选择K的值

Choosing the value of K

通常人们使用 K-均值聚类算法 是为了某些后面的用途 或者说某种下游的目的

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.

看不同的聚类数量能为 后续下游的目的提供多好的结果

K=3 S, M, L

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

E.g.

具体来说 你要做的是 从T恤生意的角度 来思考这个事情 然后问"如果我有5个分段 那么我的T恤有多适合我的顾客？