Machine Learning and Data Analytics
ME 5013- Fall 2019

Lectures 16

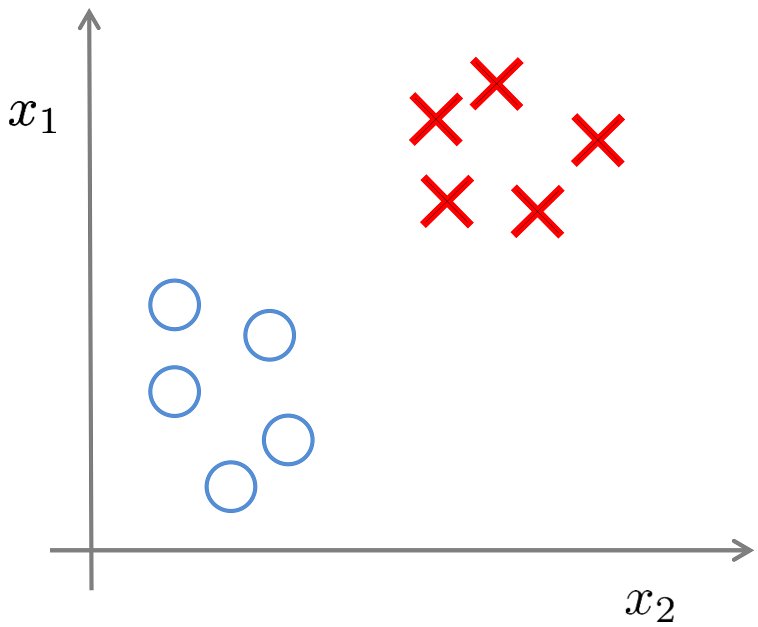
- K-means clustering

UTSA®
The University of Texas at San Antonio™

Adel Alaeddini, PhD
Associate Professor of Mechanical Engineering
Advanced Data Engineering Lab
adel.alaeddini@utsa.edu

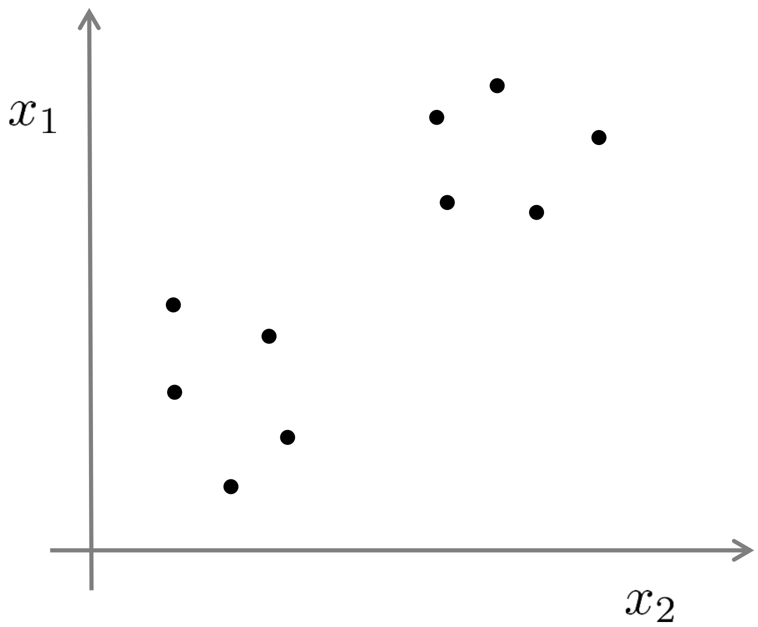
Disclosure: the slides are adopted from Andew Ng online Machine Learning course

Supervised learning



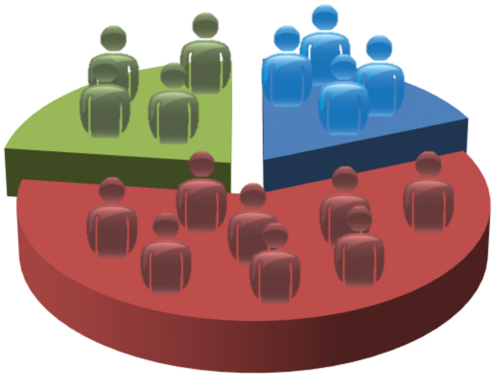
Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, \ (x^{(n)}, y^{(n)})\ \}$

Unsupervised learning

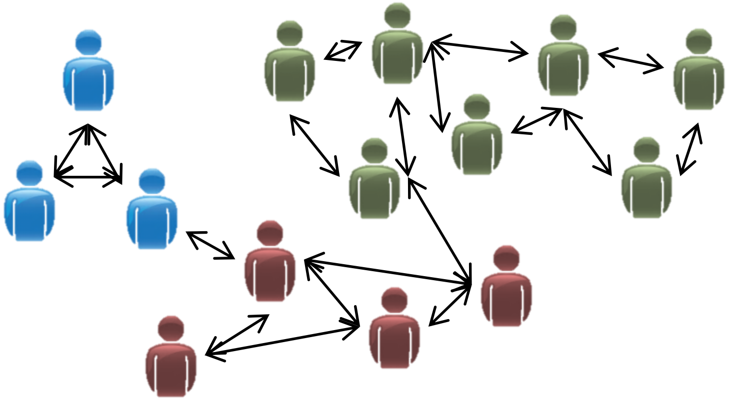


Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(n)}\}$

Applications of clustering



Market segmentation



Social network analysis

Organize computing clusters

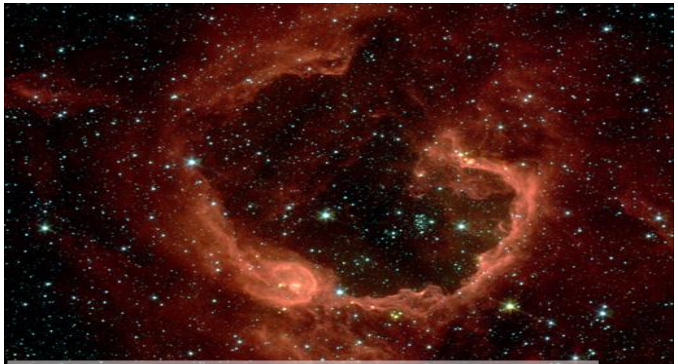
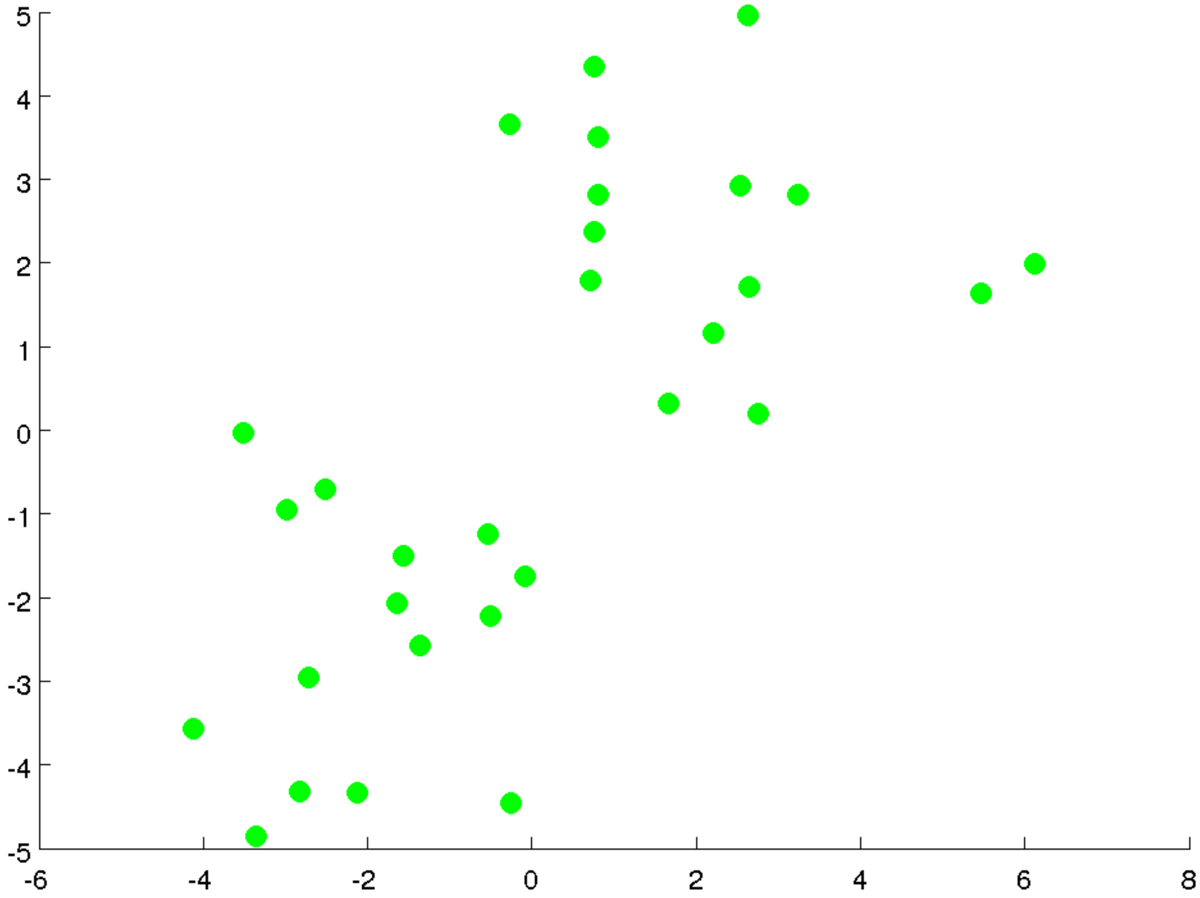


Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

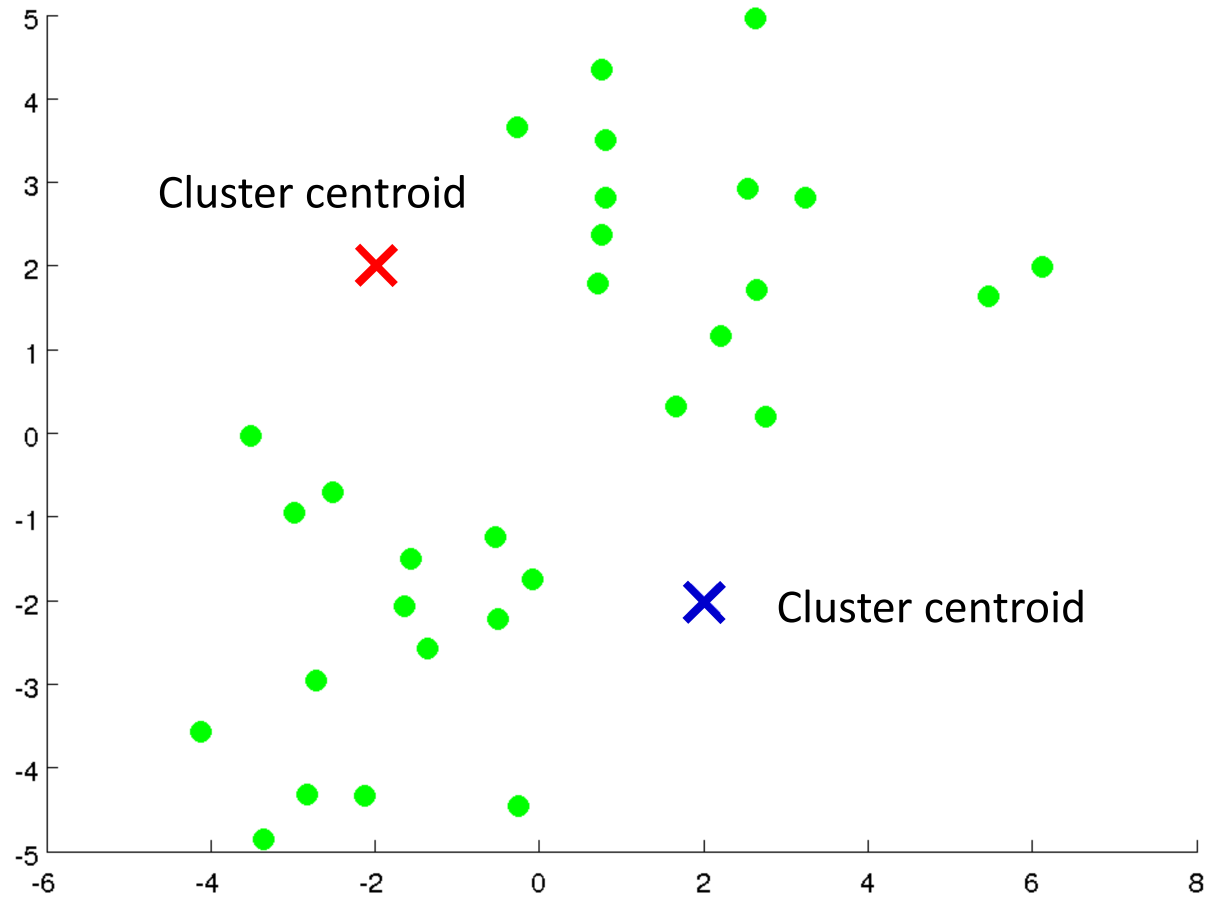
Astronomical data analysis

K-means algorithm

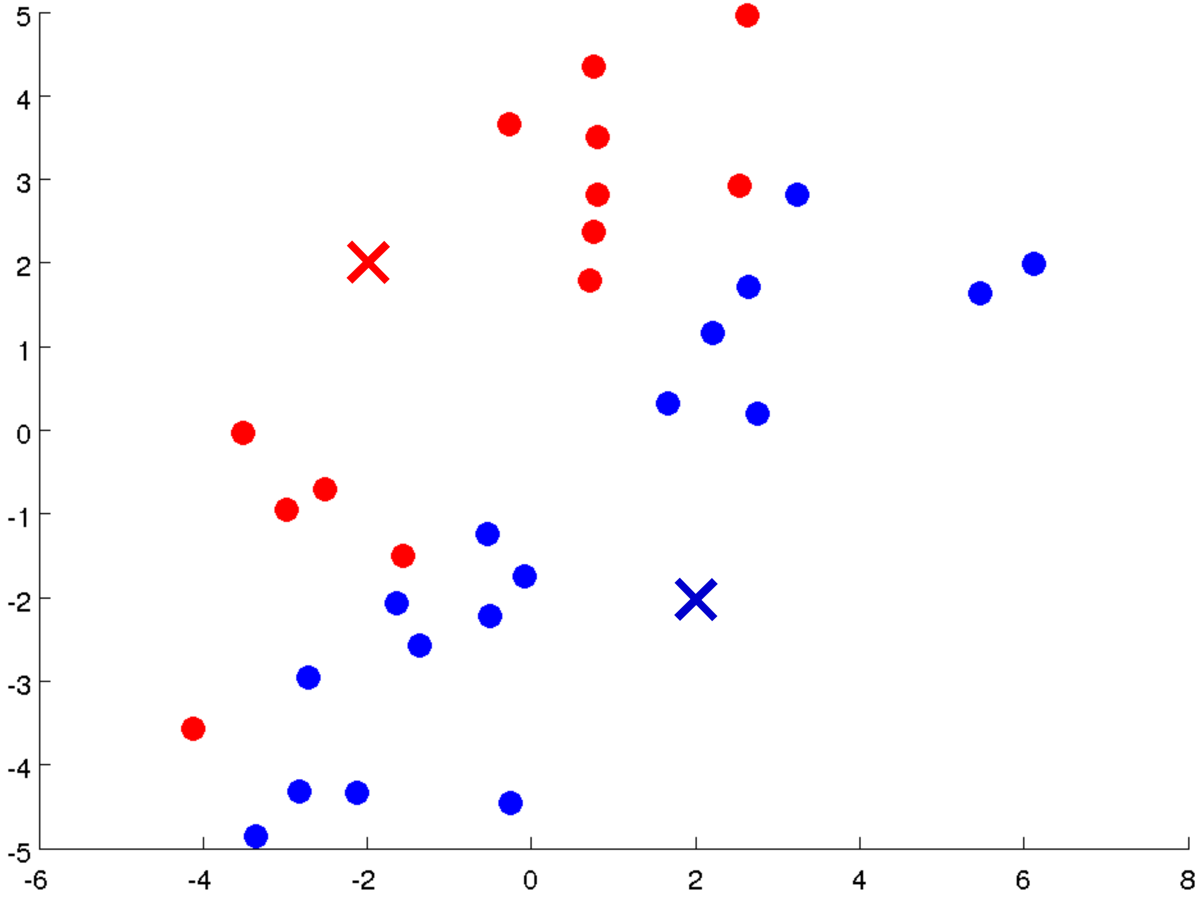


K-means algorithm

- **Random selection of cluster centroids**
- **Loop**
    - **Cluster assignment**
    - **Move centroids**

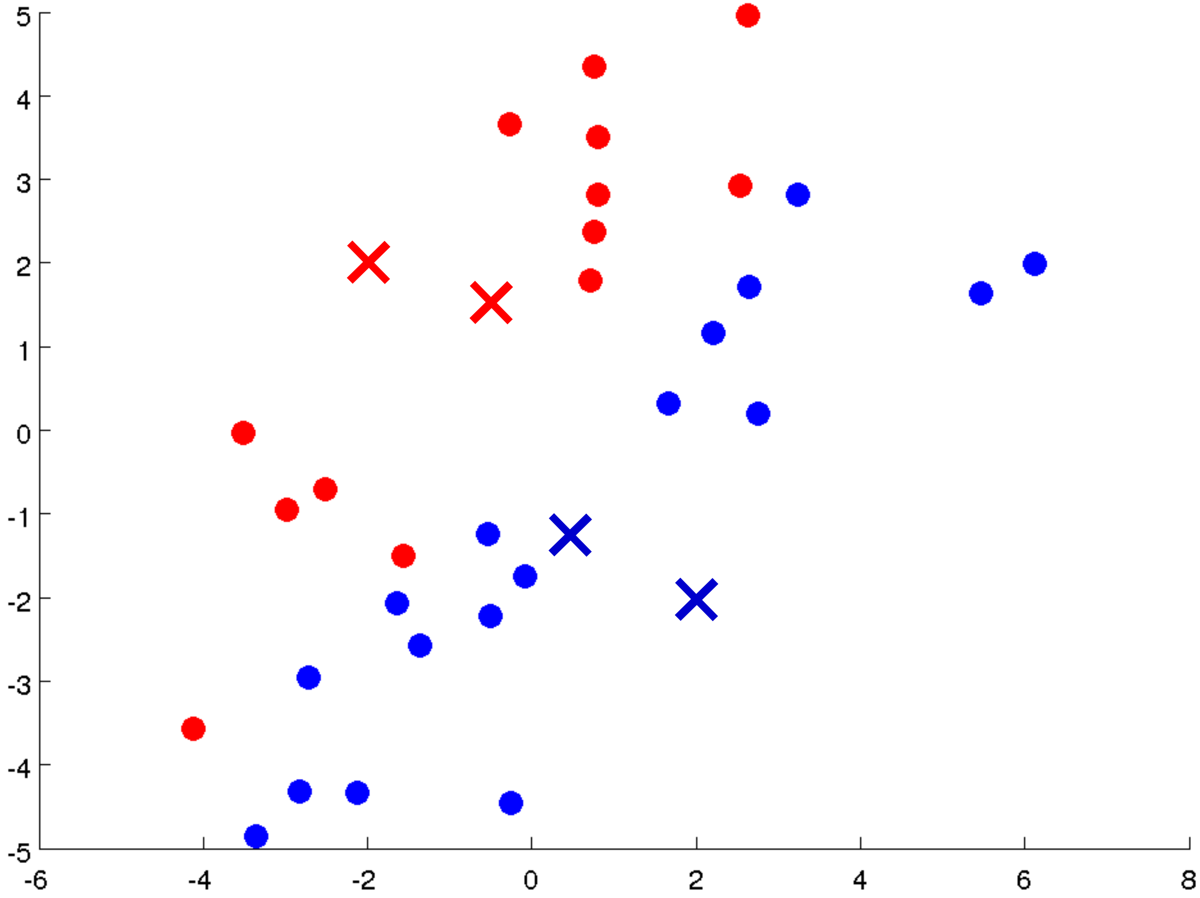


- **Cluster assignment**



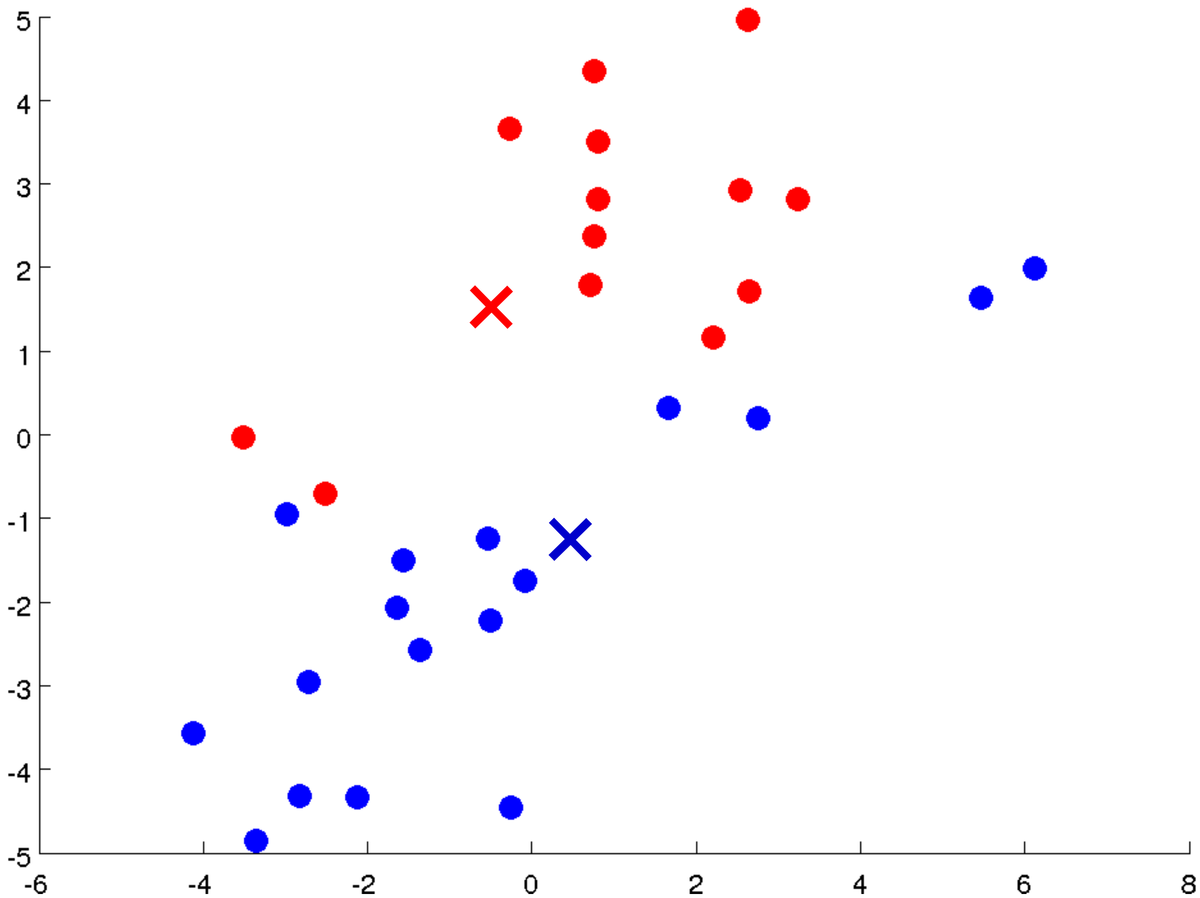
K-means algorithm

- **Move centroids**

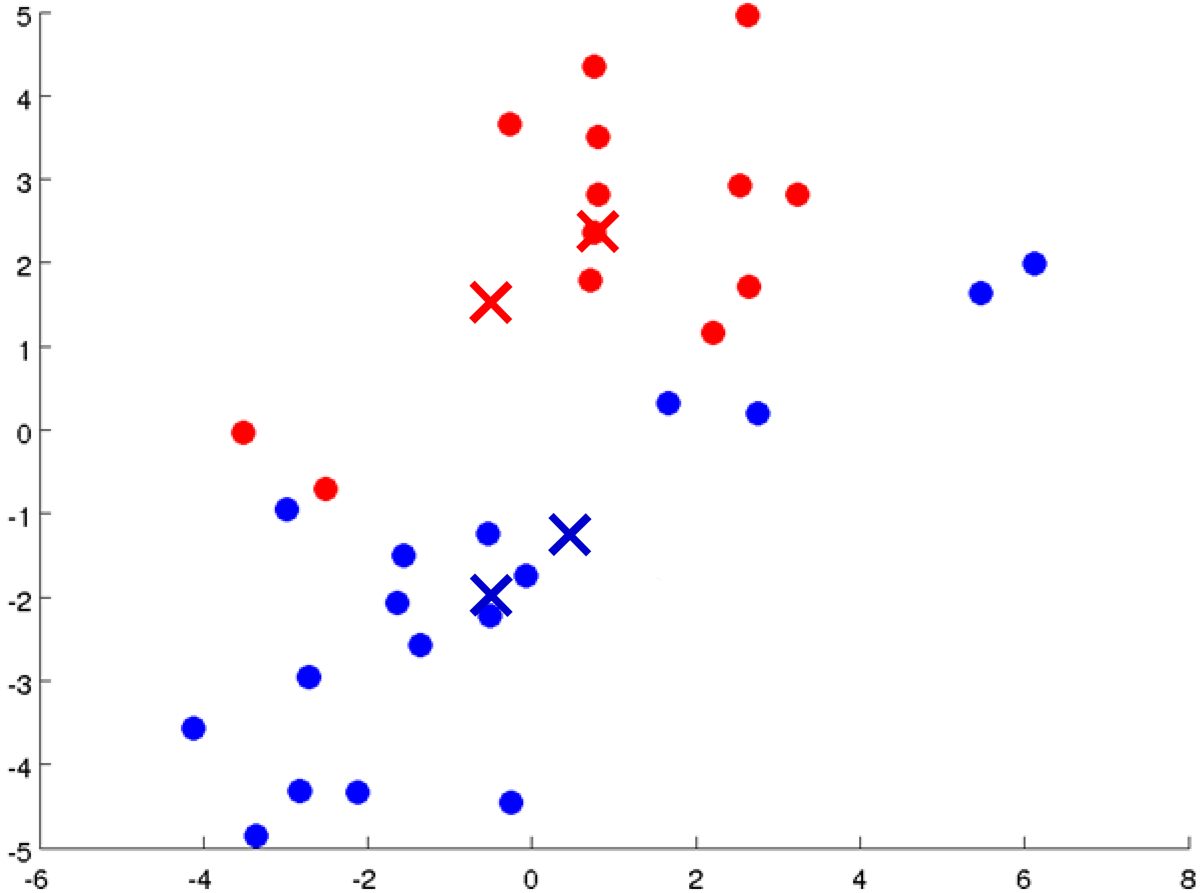


K-means algorithm

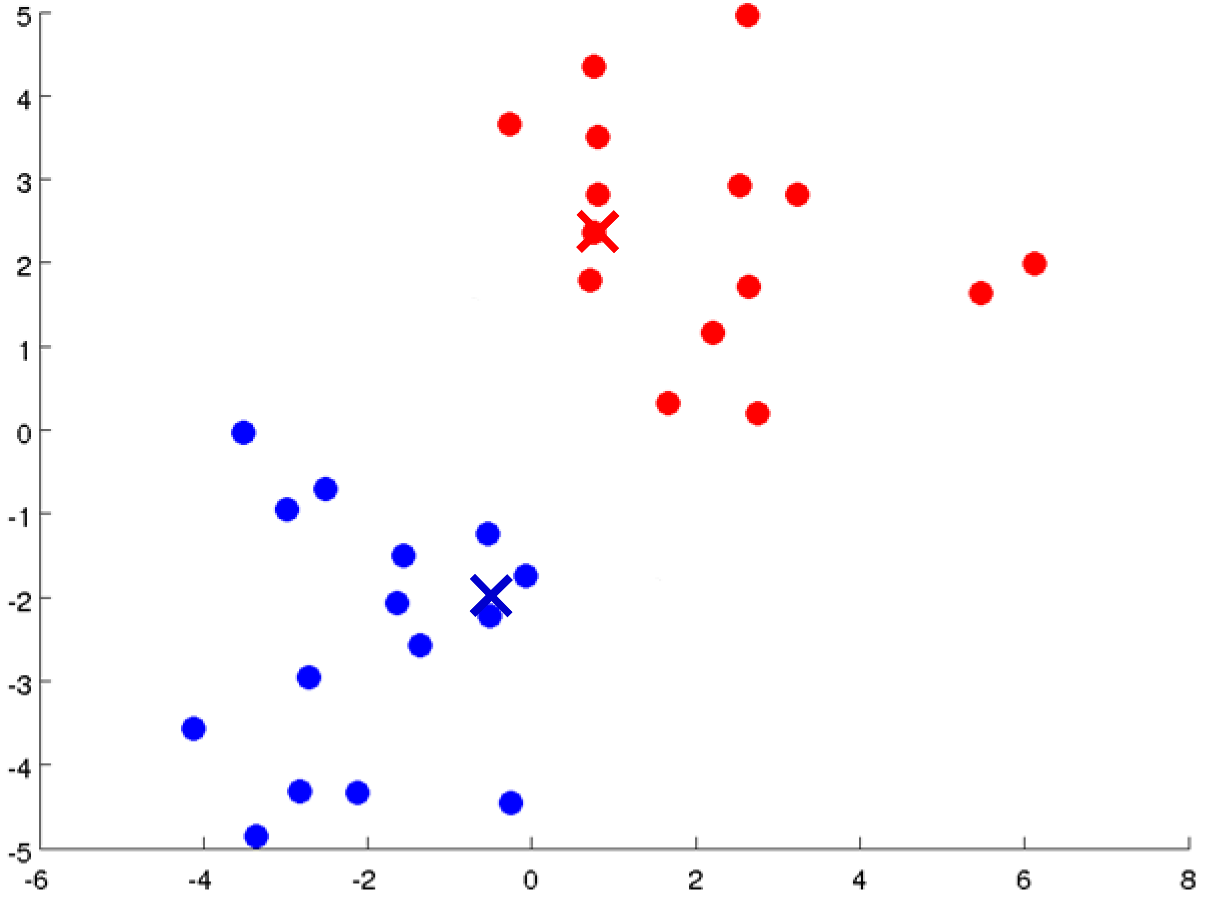
- **Cluster assignment**



- **Move centroids**

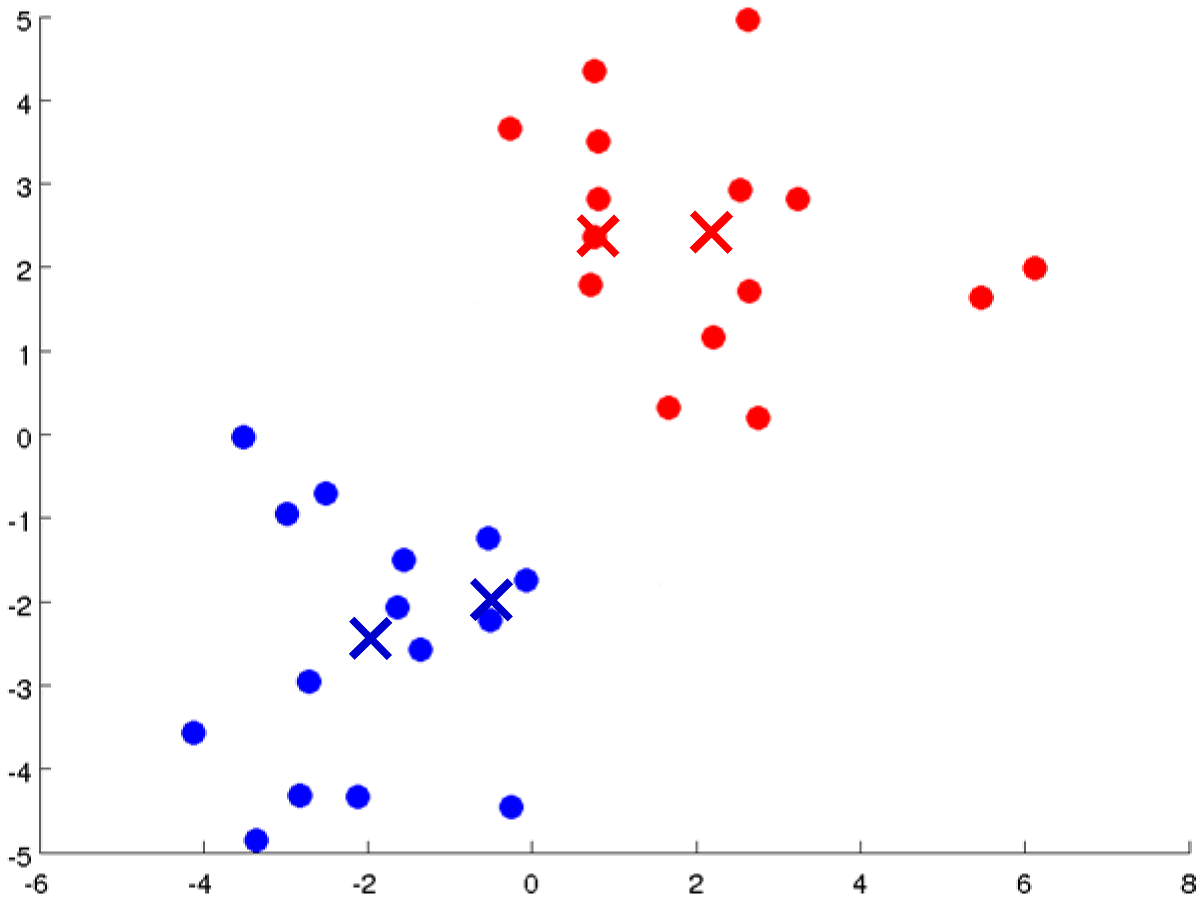


- **Cluster assignment**



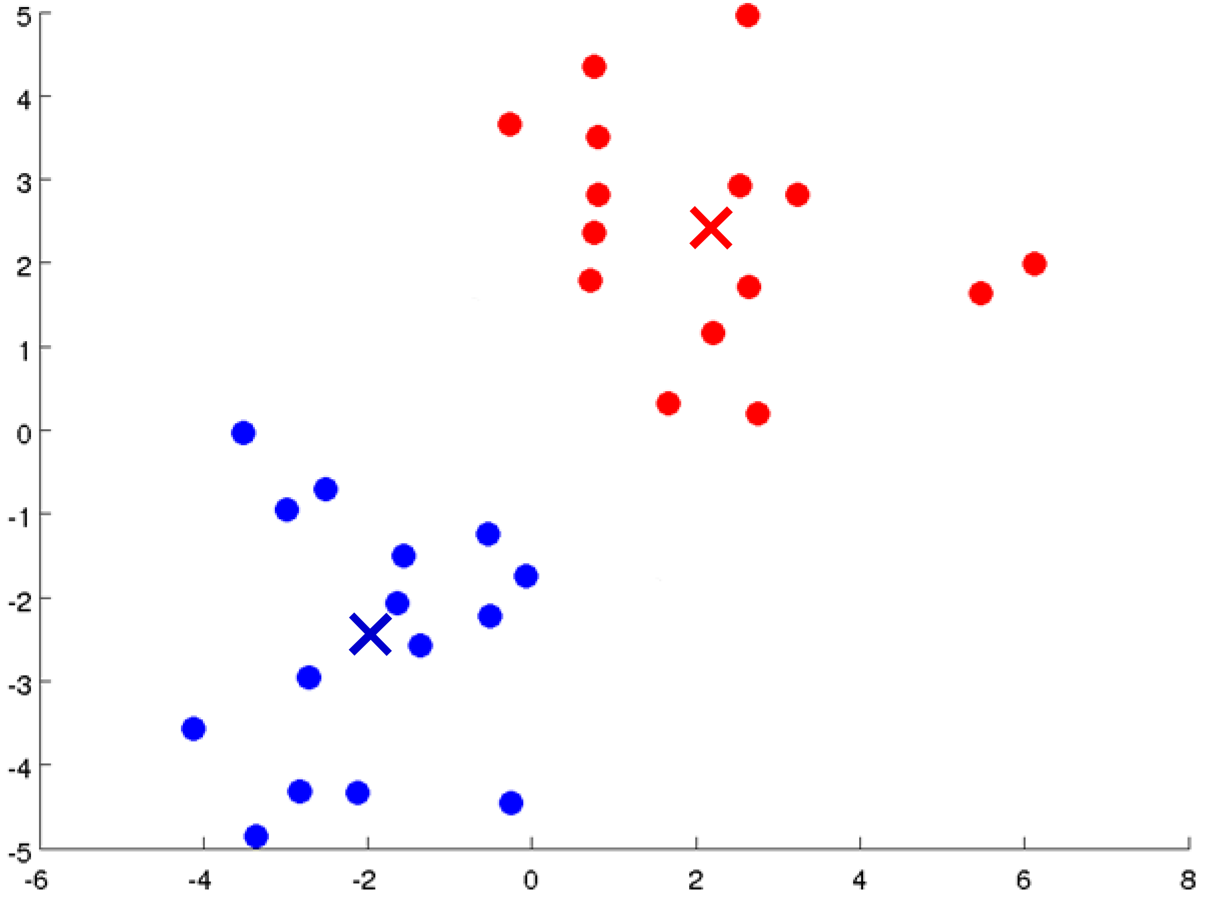
K-means algorithm

- **Move centroids**



K-means algorithm

- **Cluster assignment**



K-means algorithm

Input:

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

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

K-means algorithm

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

Repeat {

for = 1 to *n*

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

Cluster assignment

$c^{(i)} = \mathrm{Argmin}\|x^{(i)} - \mu_k\|^2$

for *k* = 1 to *K*

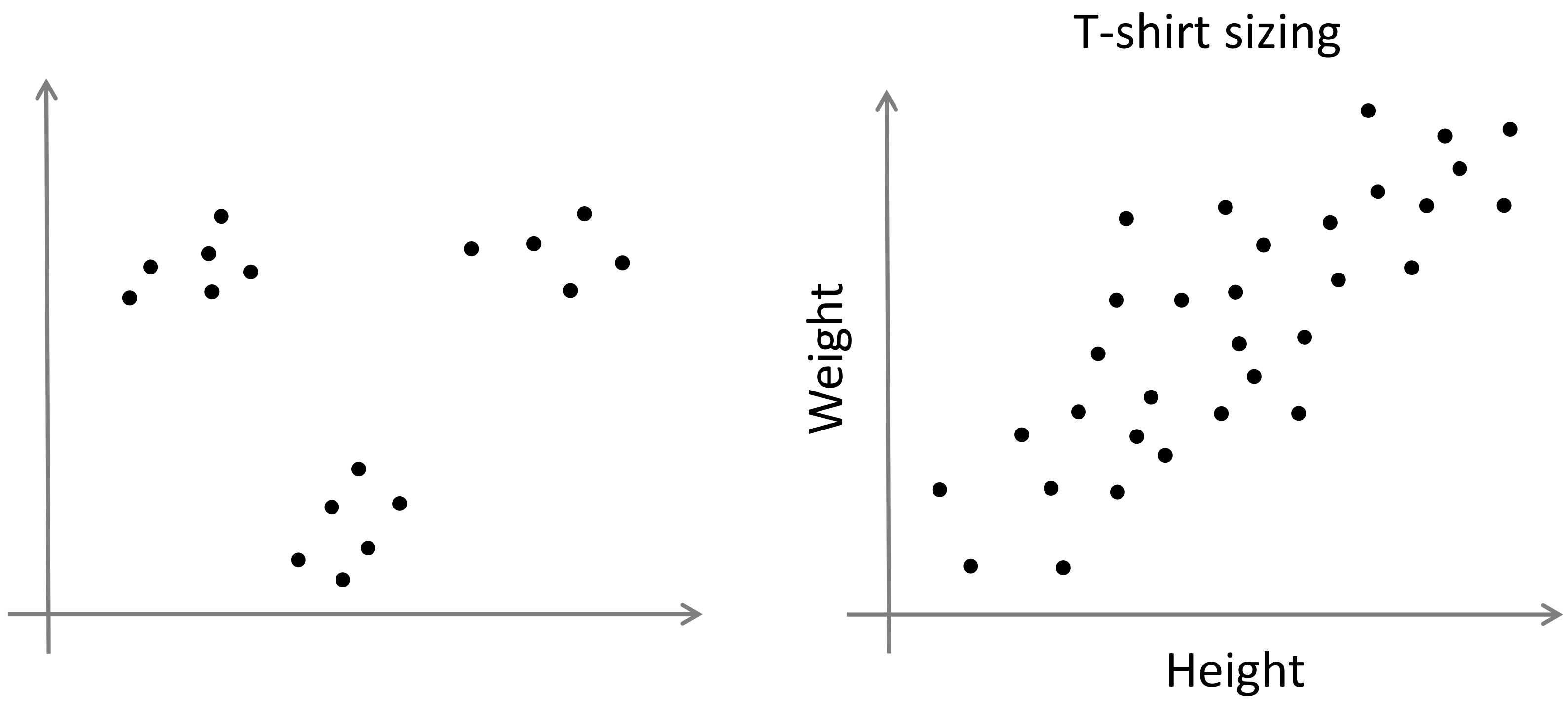
μ_k := average (mean) of points assigned to cluster

Move centroid

$\mu_k = \underset{c^{(i)}=k}{\mathrm{Mean}}\, x^{(i)}$

}

K-means for non-separated clusters



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 ($\mu_k \in \mathbb{R}^p$)

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

Optimization objective:

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

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

K-means algorithm

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

Repeat {

for = 1 to *n*

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

for *k* = 1 to *K*

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

}

Cluster assignment

Minimize *J(.)* w.r.t. $c^{(i)}$ (holding μ_k fixed)

Move centroid

$\mu_k = \underset{c^{(i)}=k}{\mathrm{Mean}}\, x^{(i)}$

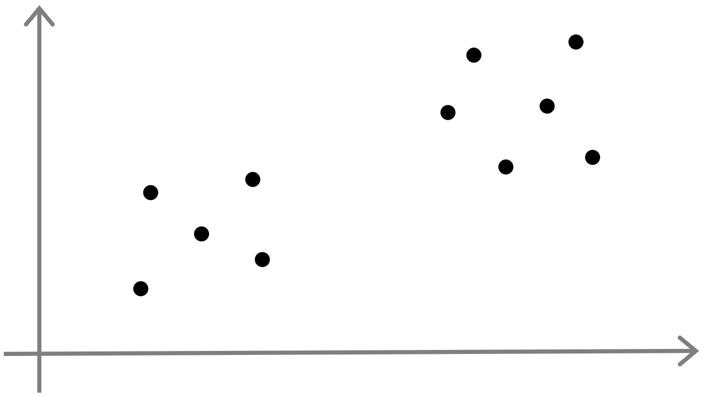
Minimize *J(.)* w.r.t. μ_k (holding $c^{(i)}$ fixed)

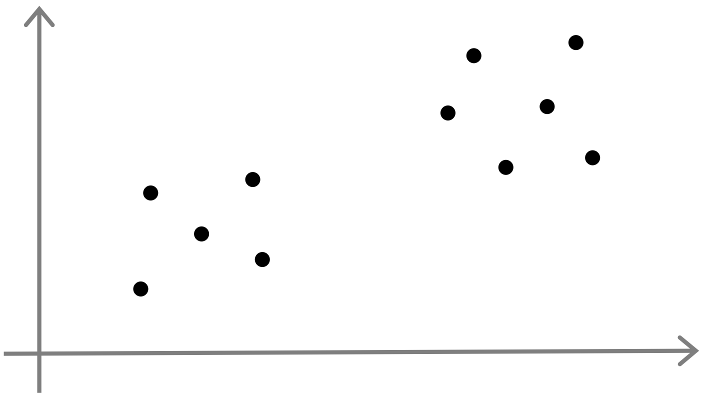
Random initialization

Should have $K < n$

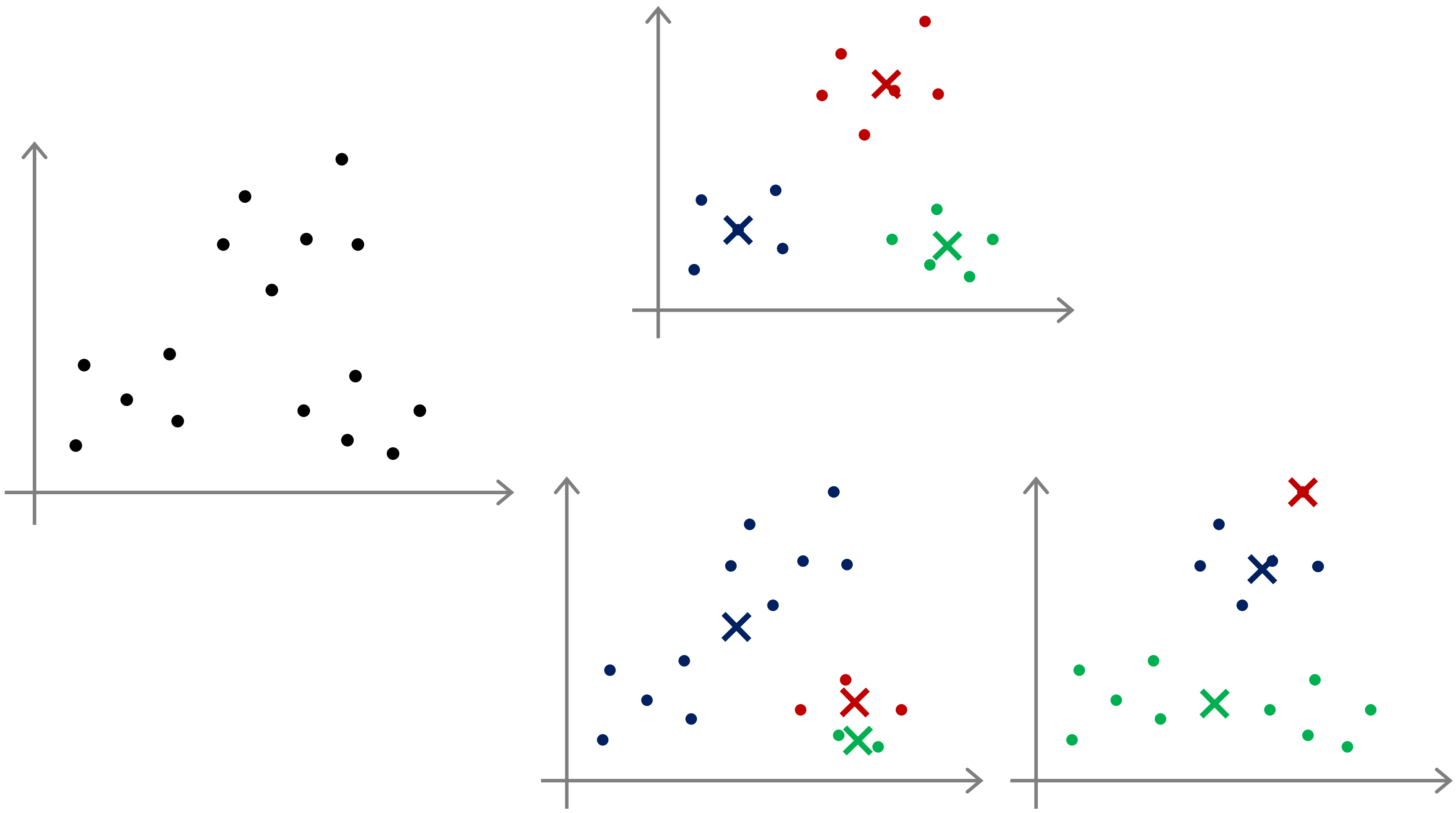
Randomly pick K training examples.

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





Local optima



Random initialization

For i = 1 to 100 {

Randomly initialize K-means.

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

Compute cost function (distortion)

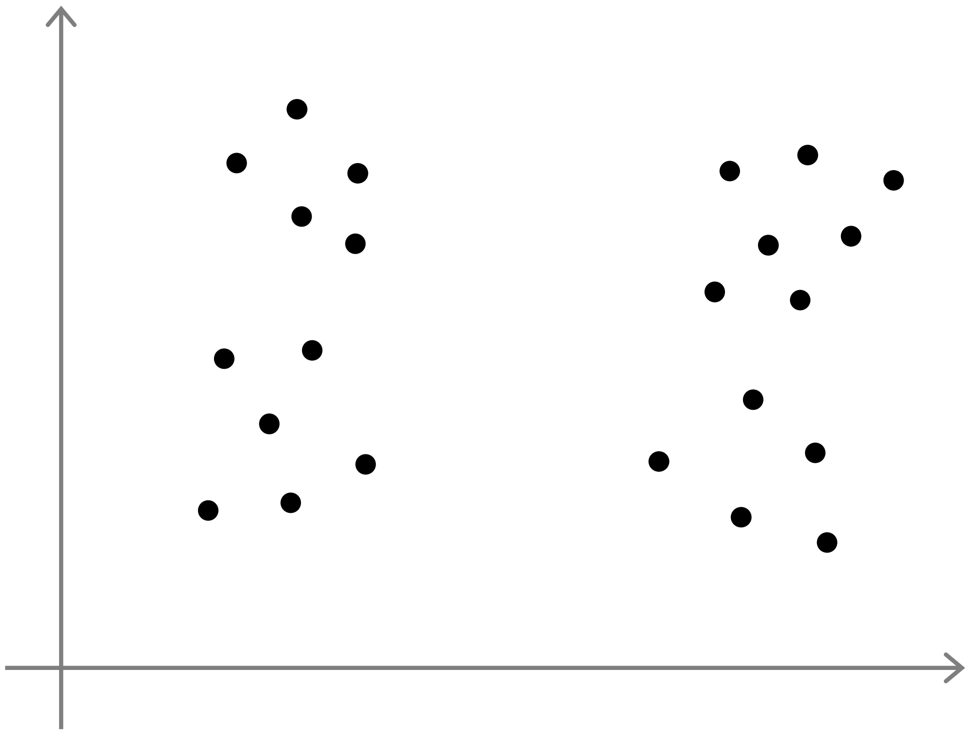
$$J(c^{(1)}, \ldots, c^{(n)}, \mu_1, \ldots, \mu_K)$$

}

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

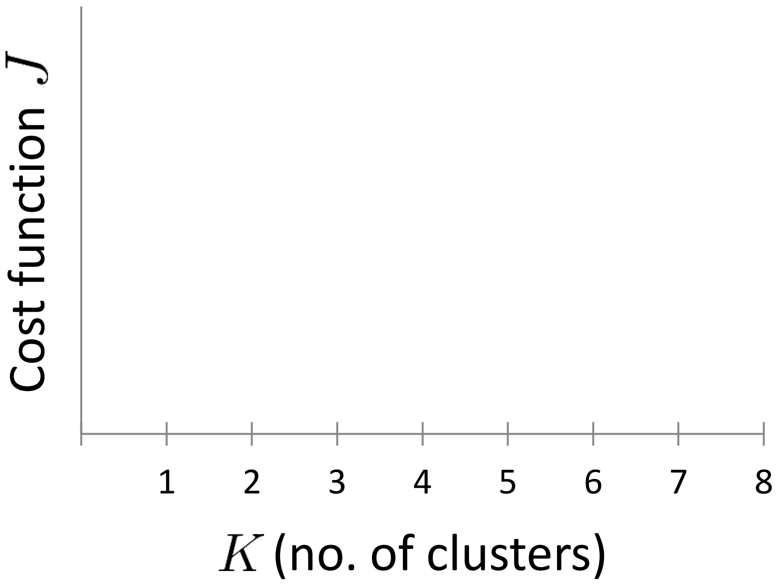
Choosing the number of clusters

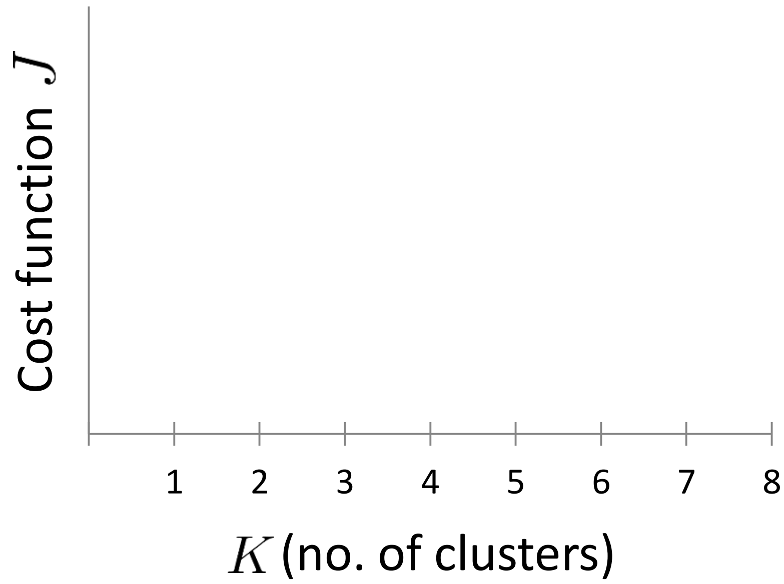
What is the right value of K?



Choosing the value of K

Elbow method:





Choosing the number of clusters

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.

E.g.