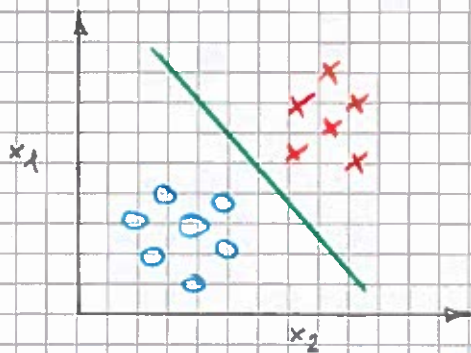


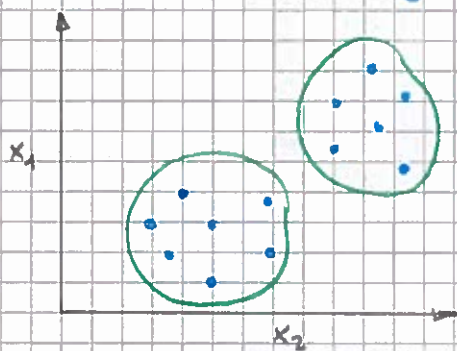
13. Clustering

Supervised learning:



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

Unsupervised learning:



Find some structure in the data

Clustering algorithm

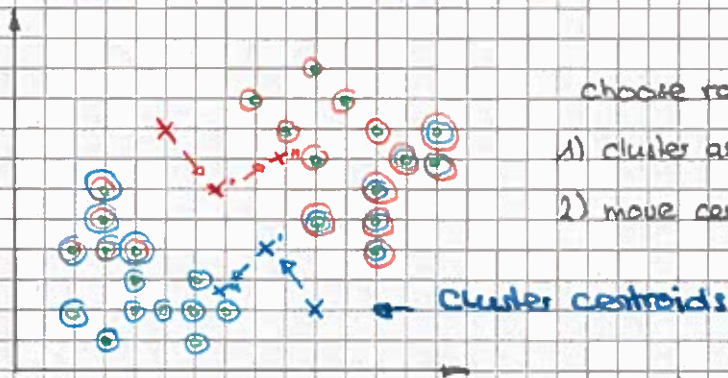
Training set: $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$

Application of clustering:

- Market segmentation
- Social network analysis ⇒ find coherent groups
- Organize computing clusters
- Astronomical data analysis

K-Means algorithm

[2]



choose randomly two cluster centroids

1) cluster assignment step \Rightarrow cluster points next to centroid

2) move centroid step \Rightarrow move to new means

set new color of data points

\Rightarrow find coherent groups

K-Means algorithm:

K-Means \rightarrow iterative 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)

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

Repeat $\{$

Cluster assignment step

For $i = 1:m$

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

closest to $x^{(i)}$ $\min_k \Rightarrow \|x^{(i)} - \mu_k\|^2 \Rightarrow$ distance between data point & centroid

move centroid step

For $k = 1:K$

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

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

\hookrightarrow means all points were assigned to cluster centroid 2

$\mu_2 = \frac{1}{4} [x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}] \in \mathbb{R}^n \Rightarrow$ compute the average of these points

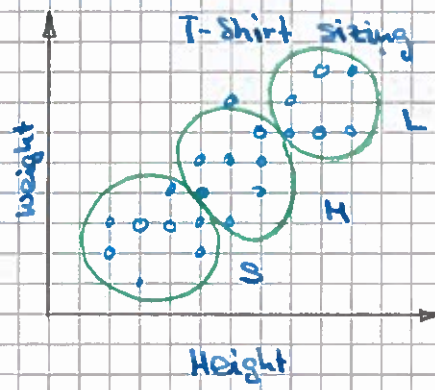
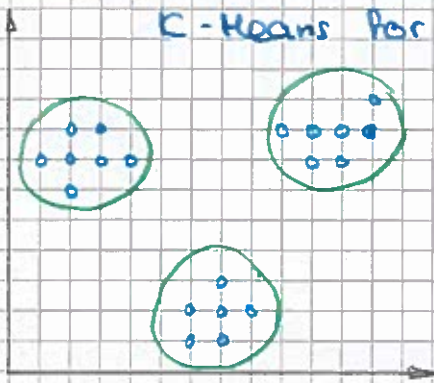
\hookrightarrow move μ_2 to average

if cluster centroid has no points assigned to it \rightarrow eliminate that $\Rightarrow K-1$ cluster

K-Means algorithm

3

K-Means for non separated clusters



S, H, L

=> Market segmentation

K-Means optimization objective

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

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

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

K = # number of clusters

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

$x^{(i)} = 5$ $c^{(i)} = 5$
 \downarrow
 assigned to cluster #5

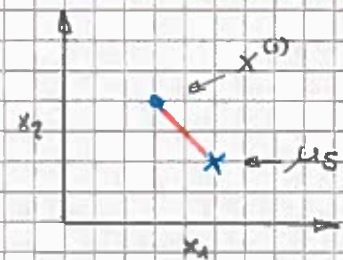
$\mu_{c^{(i)}} = \mu_5$

Optimization Objective:

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

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

also called Distortion of cost function



K-Means algorithm

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

Repeat $\{$

Cluster assignment step

For $i = 1:m$

$\Rightarrow \min J(\dots)$ wrt $c^{(1)}, c^{(2)}, \dots, c^{(m)}$

(holding μ_1, \dots, μ_K fixed)

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

move centroid step

For $k = 1:K$

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

$\}$

$\Rightarrow \min J(\dots)$ wrt μ_1, \dots, μ_K

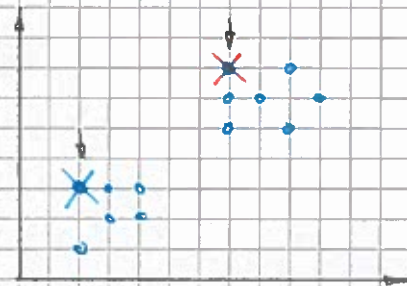
Random initialization:

Should have $K < m$

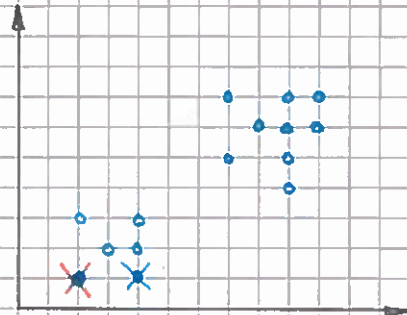
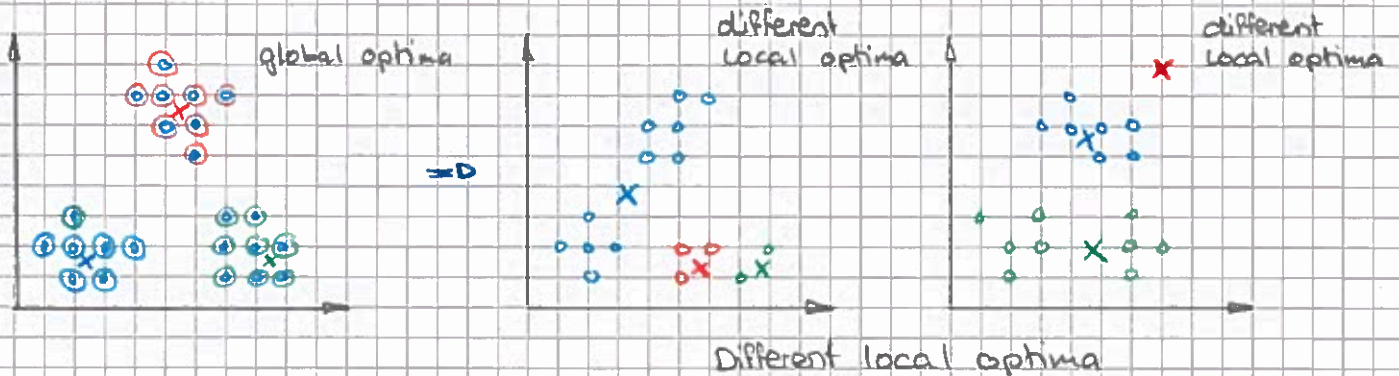
Randomly pick K training examples

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

$K=2$



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

Local optima:

to prevent K-Means getting stuck at local optima, try multiple initializations

Random initialization:

for $i = 1 : 100$ \rightarrow typically 50 - 1000

Random initialize K-Means

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

Compute cost function (distortion)

$$J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$$

}

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

if we have $K = 2-10$ random-init can make difference, but not so much a problem with K large.

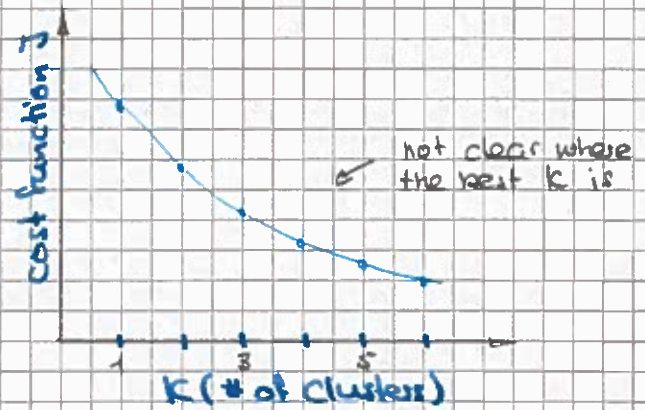
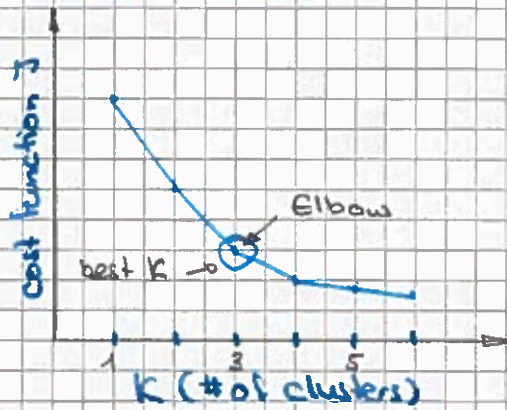
Choosing the number of clusters

6

What is the write value of k ?

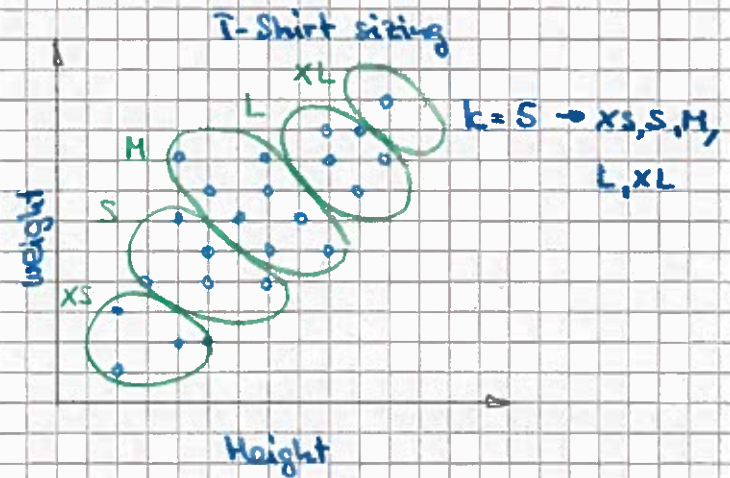
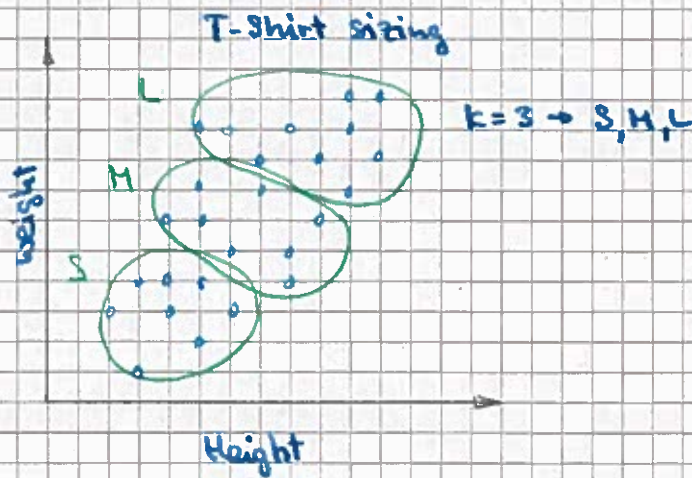
Choosing the value of k

Elbow method



\Rightarrow no high expectation with this approach.

Sometimes you are running K-Means to get clusters to use for some later/downstream purposes. Evaluate K-Means based on a metric for how well it performs for that later purpose.



Choosing k deciding by looking at clustered output \Rightarrow chosen by human inside