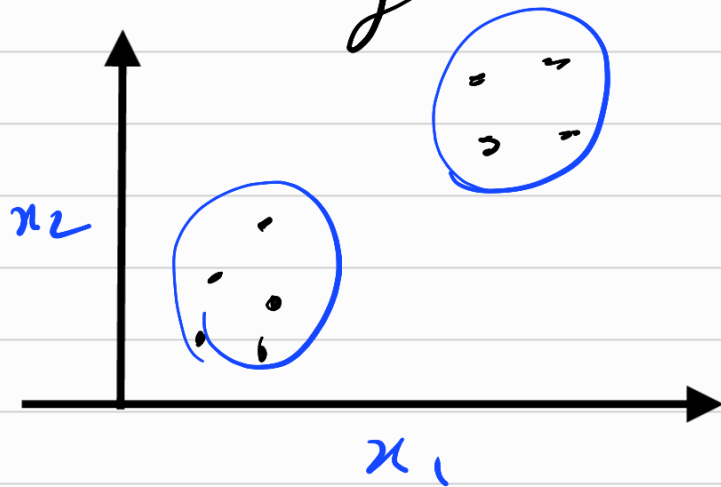
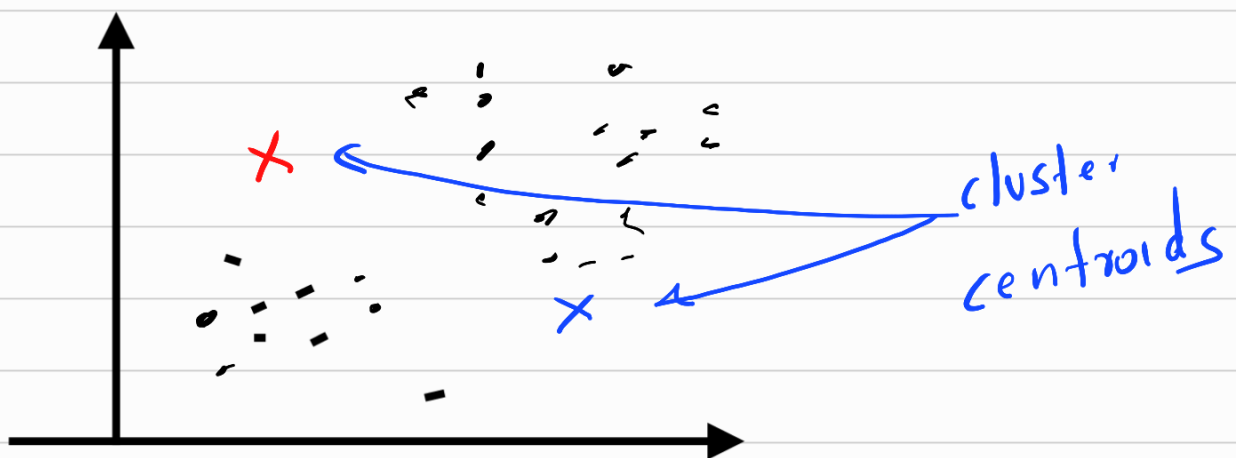


1) Clustering



eg: Group similar news
Market segmentation
DNA analysis
Astronomical analysis

i) K-means clustering algorithm

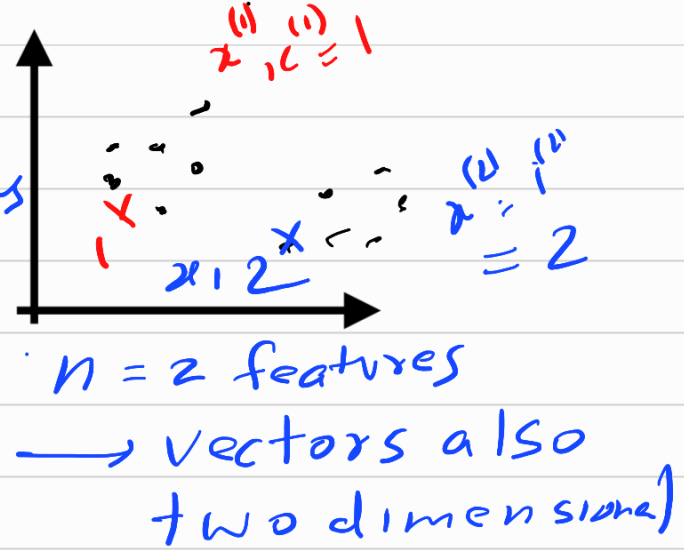


- ① Assign each point to the closest centroid
- ② Recompute centroids - assign each point closest too
(avg / point assigned)

K-means algorithm

Assign points to centroids
for $i = 1$ to m

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



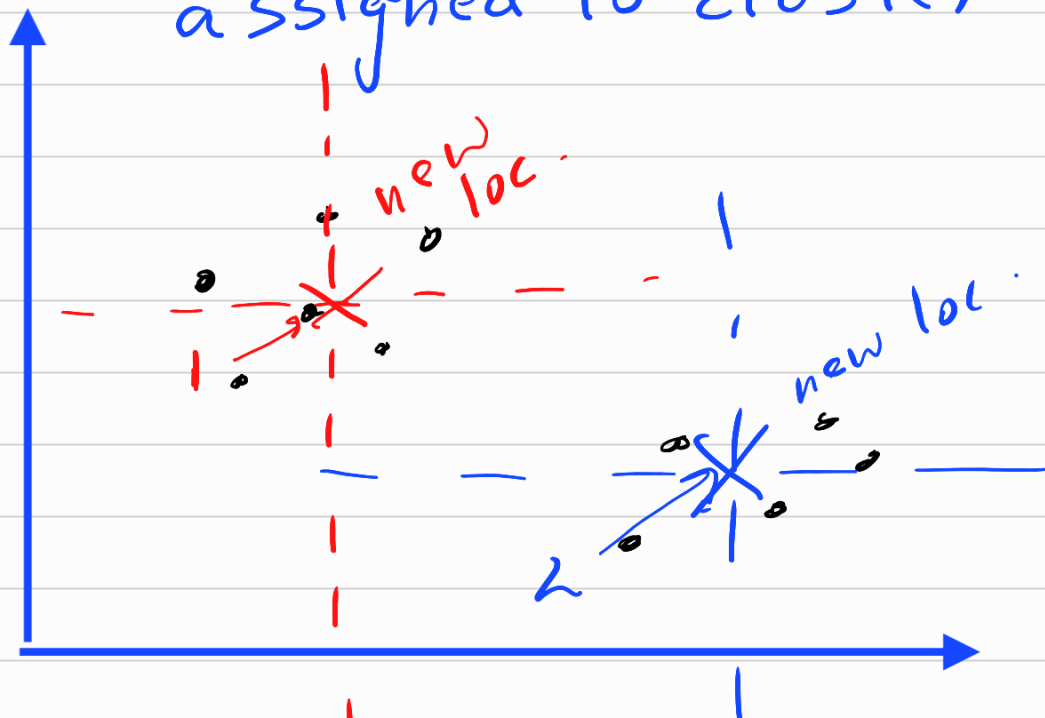
lowest distance to centroid closest to $x^{(i)}$

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

Move cluster centroids

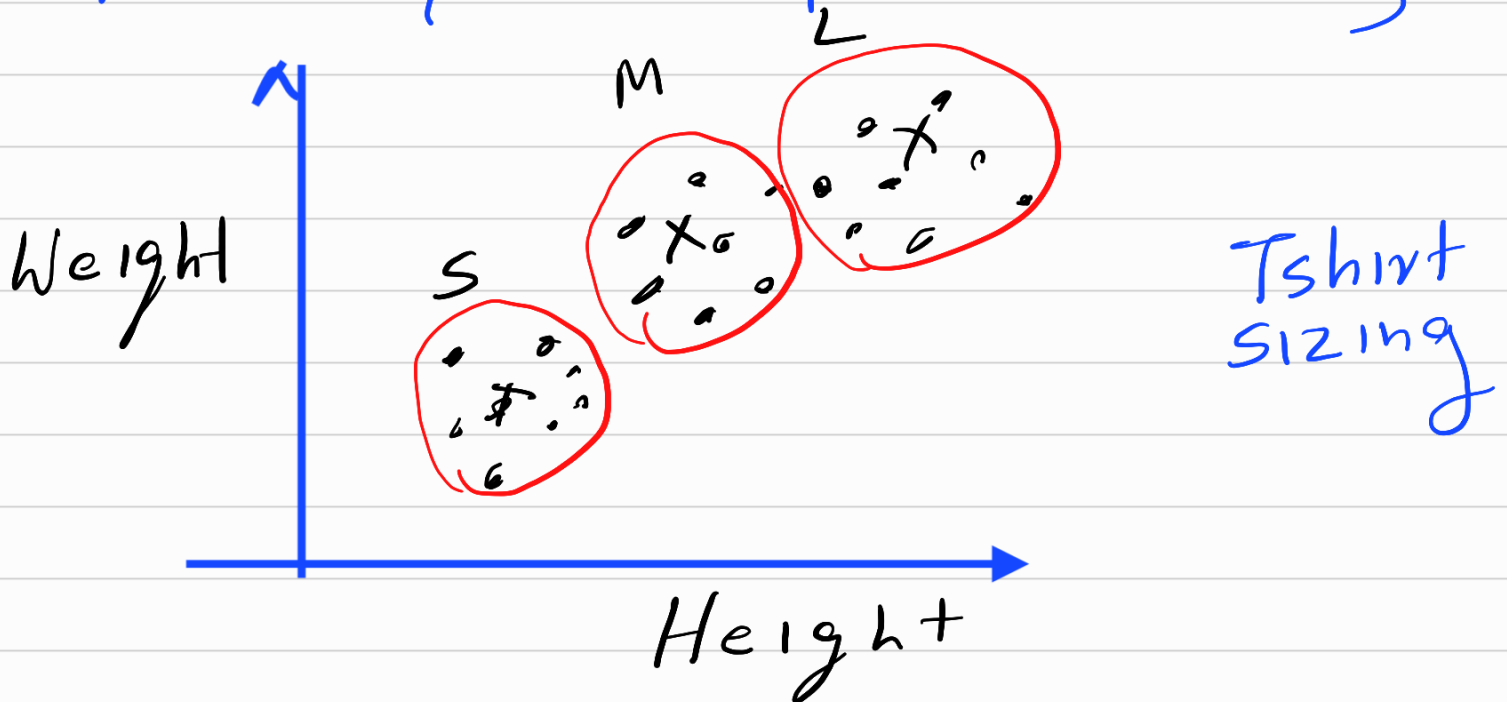
for $k = 1$ to K ; → no. of clusters

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



1
If no points assigned to cluster
→ eliminate cluster

Not only for separated ones;



K-means optimization - objective

$c^{(i)}$ - index of cluster $x^{(i)}$ currently assigned.

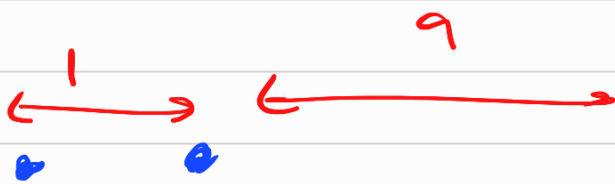
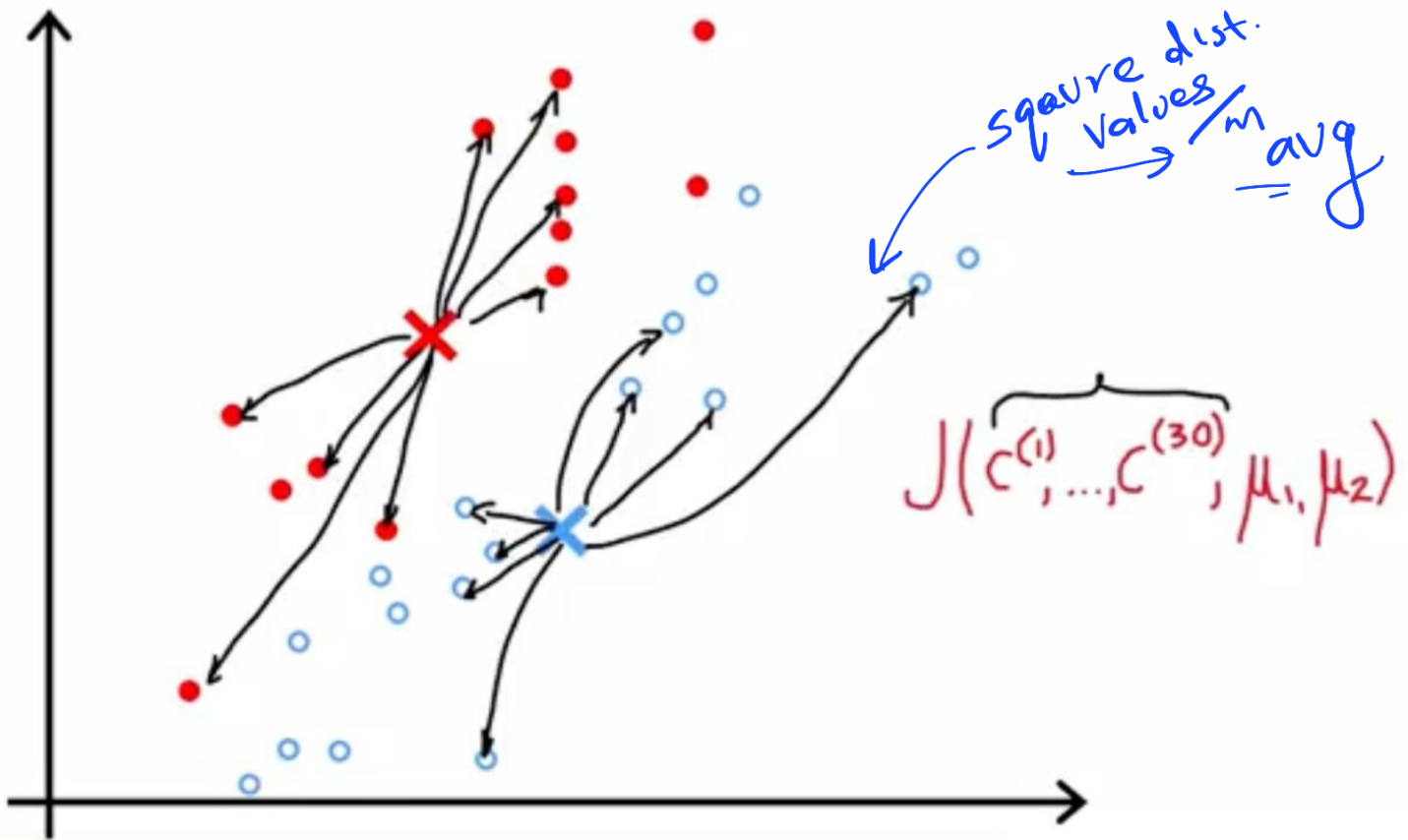
μ_k - cluster centroid k

μ_c - cluster centroid of $x^{(i)}$ assigned cluster

Cost function

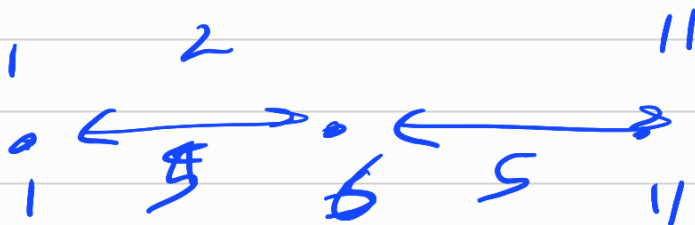
$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k) = \frac{1}{m} \sum_{i=1}^m \|p c^{(i)} - \mu_c\|^2$$

distortion



$$\frac{1}{2} (1^2 + 9^2) = 41$$

✓



$$\frac{1}{2} (5^2 + 5^2) = 25$$

Cost function works

Initialize k-Mean

$k < m \leftarrow$ train examples

- Random initialization;

↳ Randomly pick k training examples (on top of a) ↗

Sometimes okay; but if a local minimum
→ gets stuck

for $i = 1$ to 100 {

Randomly initialize k -means

Run k -means

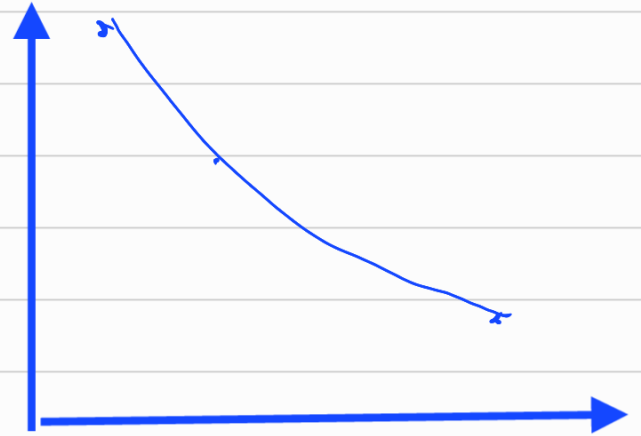
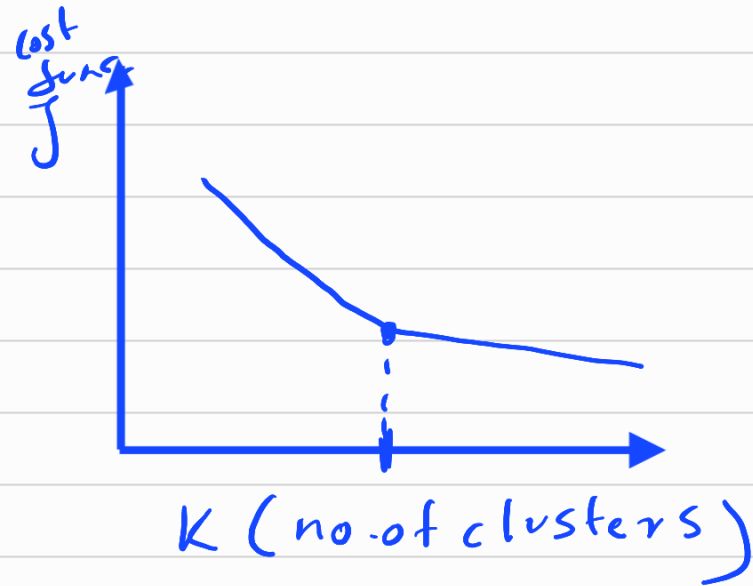
Compute cost function

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

pick clusters of lowest cost

Choosing No. of clusters;

Elbow method



Don't choose k to minimize J
→ Largest possible $k - \infty$

Best → Evaluate K-means on performance on later purpose (tradeoffs)

