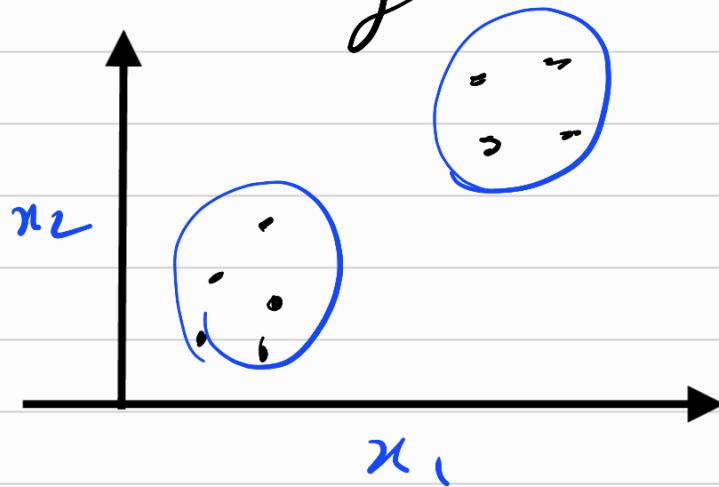
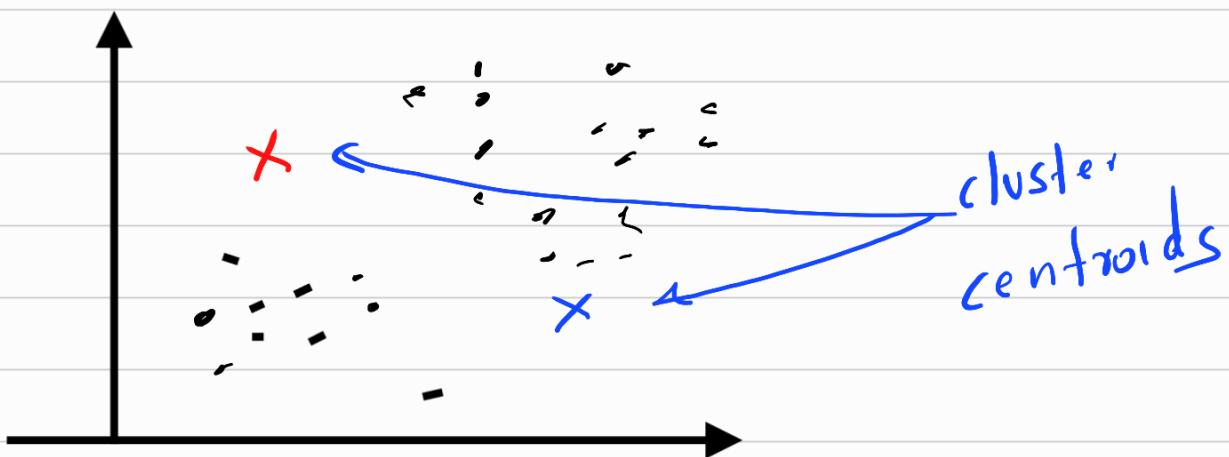


i) Clustering



e.g.: 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 to (avg / point assigned)

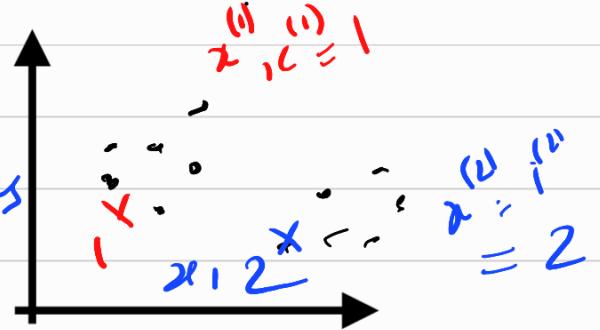
K -means algorithm

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 $x^{(i)}$

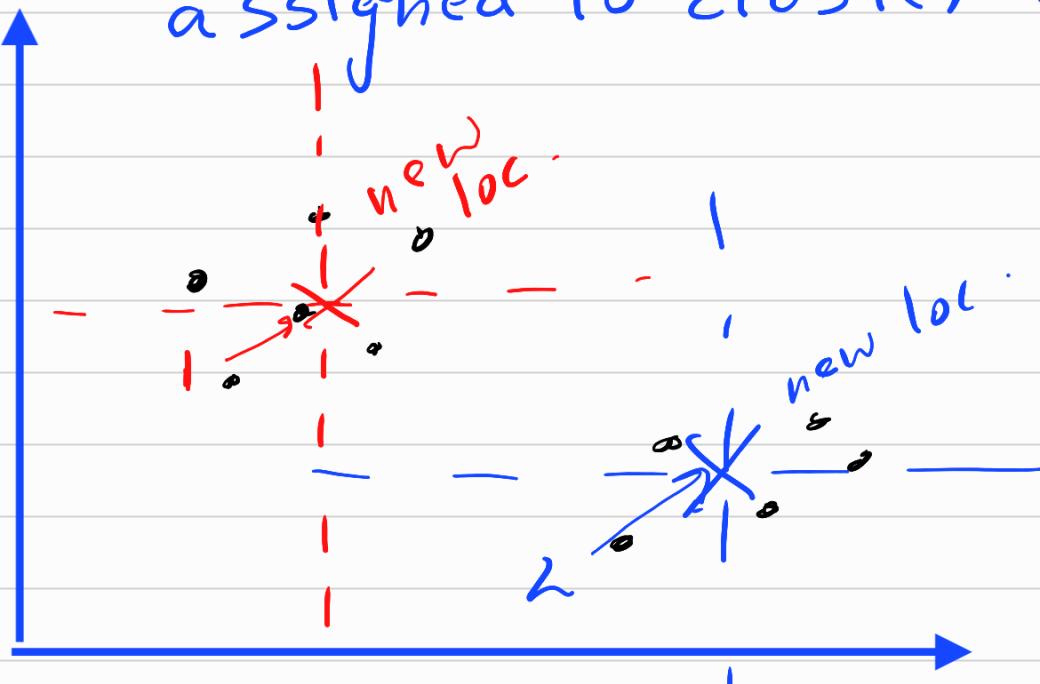


$$= \min_k \|x^{(i)} - \mu_k\|^2$$

Move cluster centroids

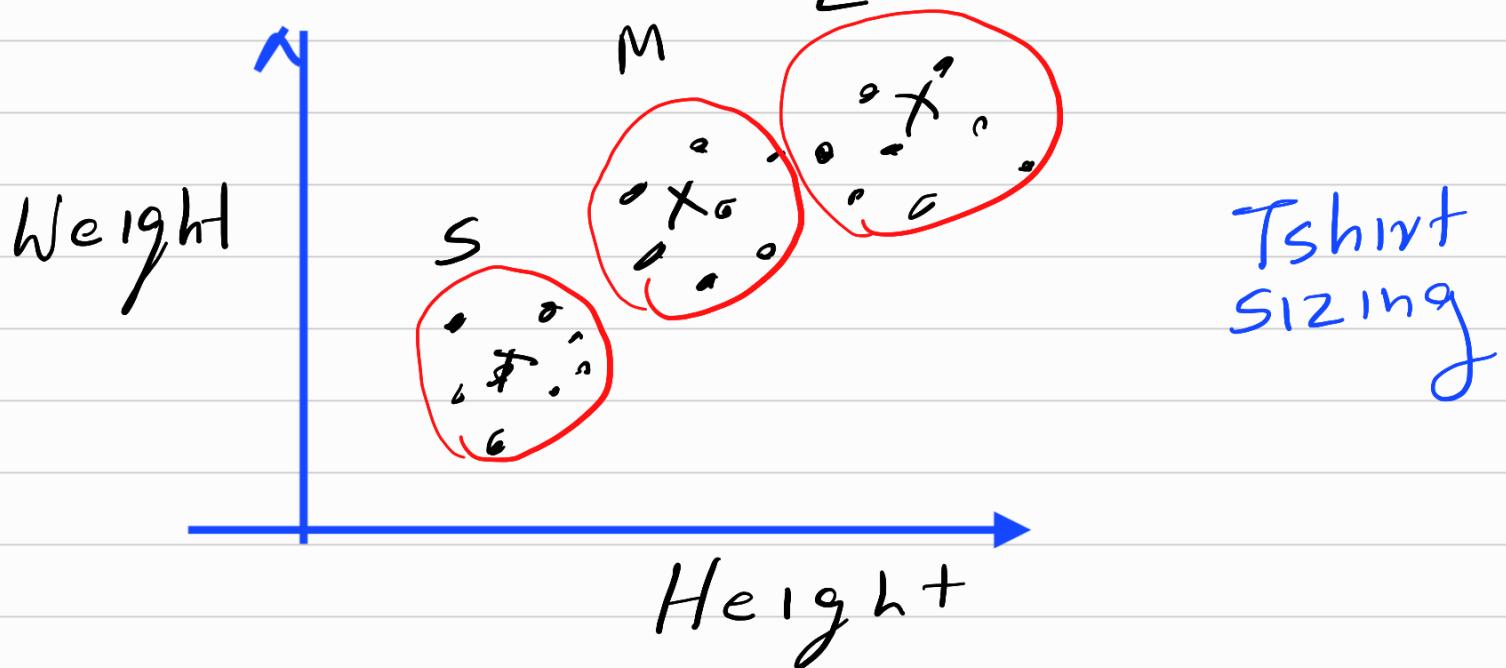
for $k = 1$ to K ; \rightarrow no. of clusters

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



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.

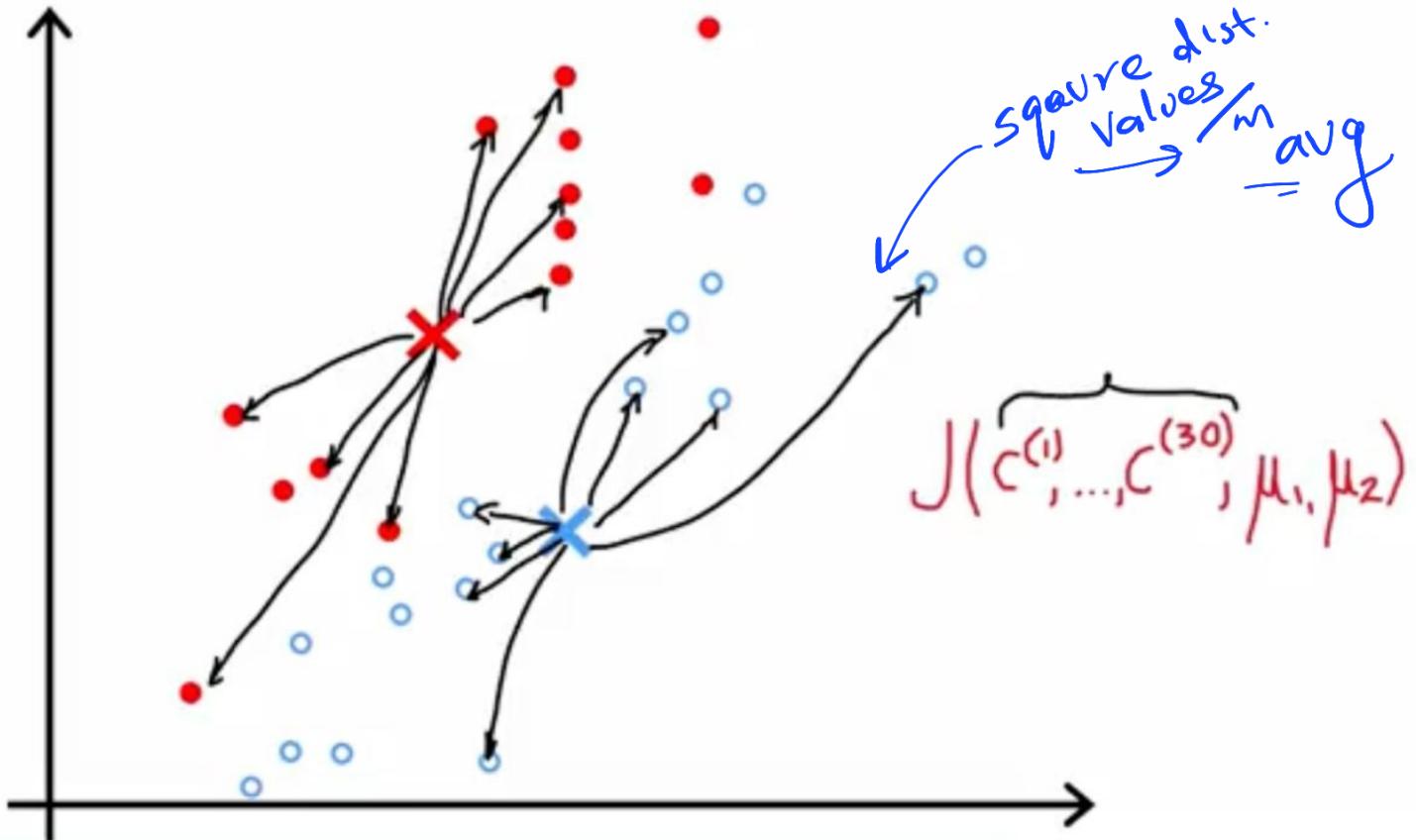
m_k - closer centroid k

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

Cost function

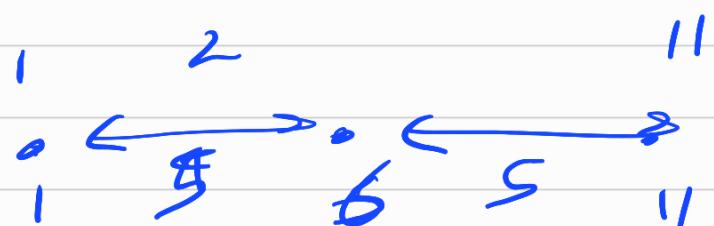
$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k) = \frac{1}{m} \sum_{j=1}^m \| \mathbf{x}^{(j)} - \mu_c^{(j)} \|^2$$

↓ distortion



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

✓



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

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

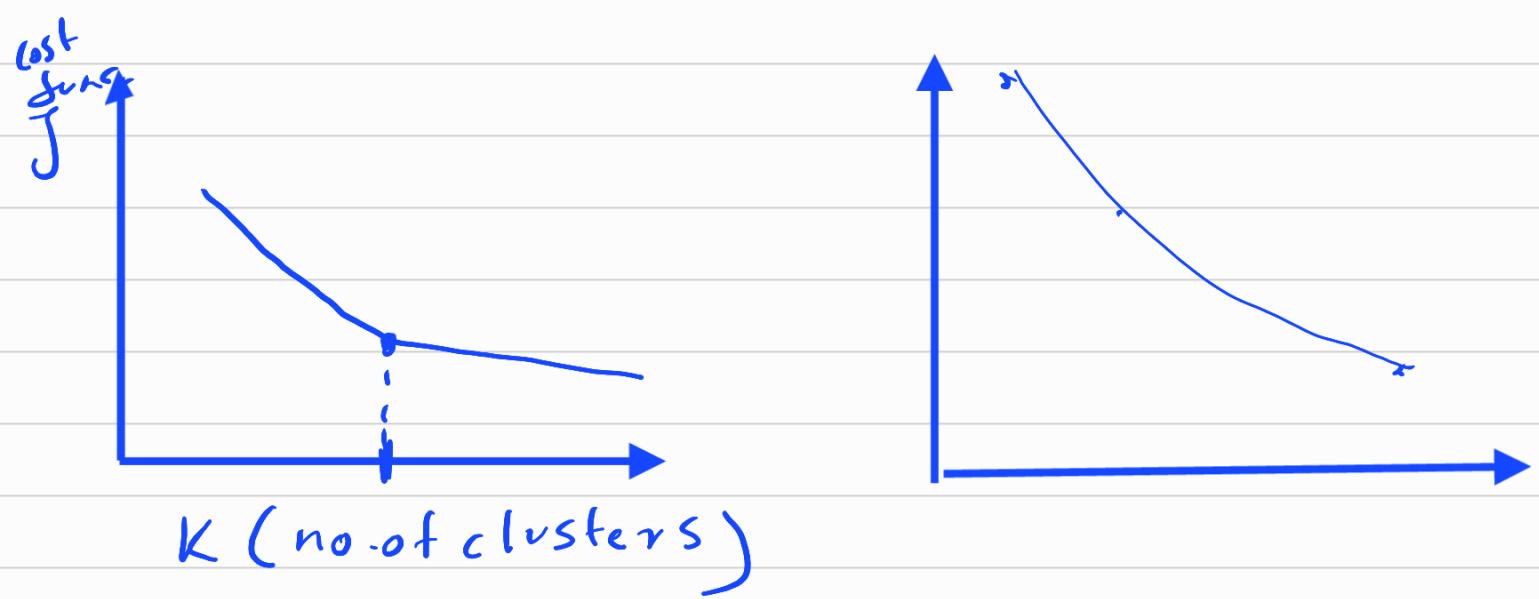
 Compute cost function

$J(c^{(1)}, \dots c^{(m)}, \mu_1, \mu_2, \dots \mu_k)$

 } pick clusters of lowest cost

Choosing No. of clusters;

Elbow method



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

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

