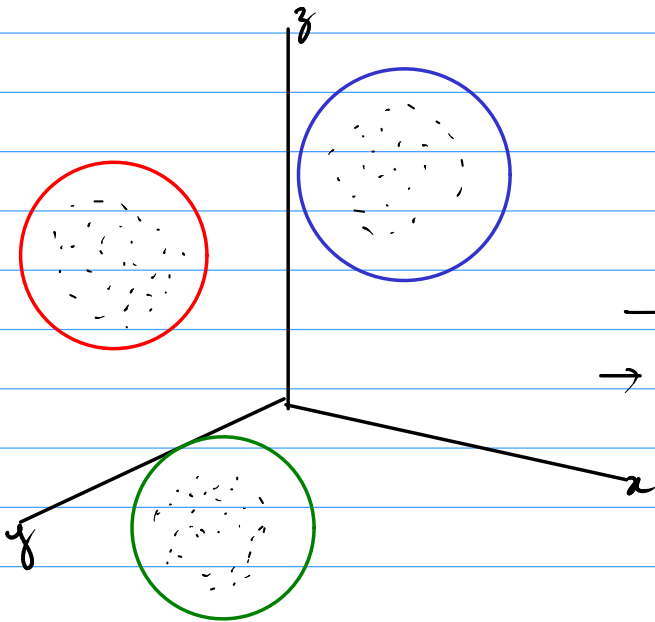


— K-MEANS CLUSTERING

1. THE PROBLEM
2. ILLUSTRATION ON A 1-D EXAMPLE
- ✓ 3. THE 1-D ALGORITHM
- ✓ 4. DISTORTION AND ITS MINIMIZATION BY K-MEANS
- ✓ 5. THE K-MEANS ALGORITHM FOR HIGHER-D DATA
- ✓ 6. "FAILURE" OF K-MEANS
- ✓ 7. EXAMPLES AND APPLICATIONS

1. PROBLEM



— DATA MATRIX A

— GIVEN k : DESIRED # OF CLUSTERS

— GOAL : FIND BEST k CLUSTERS

— APPLICATIONS

→ Eg: (SUDS): MOVIE RANKING DATA CLUSTERING

— VERY DIFFICULT: NP-HARD

— K-MEANS CLUSTERING

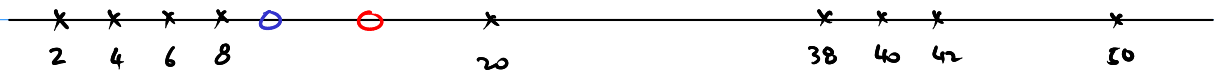
— GOOD SOLUTIONS QUICKLY: USUALLY

— BUT NO GUARANTEE

2. 1-D EXAMPLE



→ STEP 0: PICK $k=2$ CLUSTER CENTERS ("MEANS")

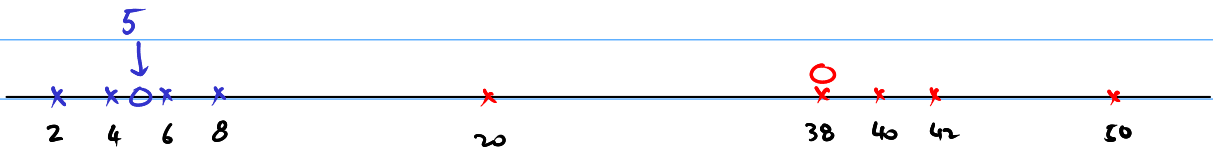


— ROUND 1

— STEP 1: ASSIGN DATA TO NEAREST CLUSTER MEAN



— STEP 2: UPDATE MEANS TO AVG OF CLUSTER DATA



$$\frac{20}{4} = 5$$

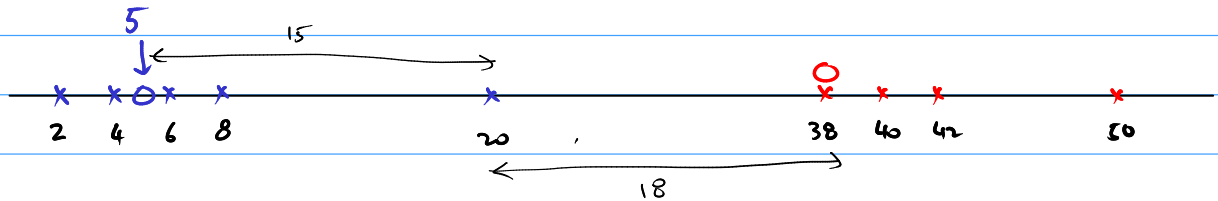
$$\frac{20 + 120 + 50}{5} = \frac{190}{5} = 38$$

— STEP 3: ANY CHANGE → GO TO STEP 1 (NEXT ROUND) ✓

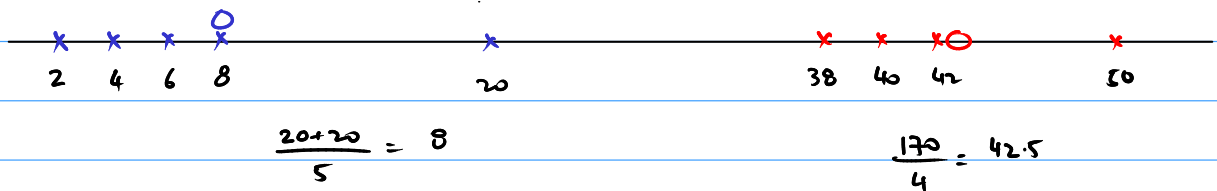
— ELSE: DONE (STOP)

ROUND 2

STEP 1: ASSIGN DATA TO NEAREST CLUSTER MEAN



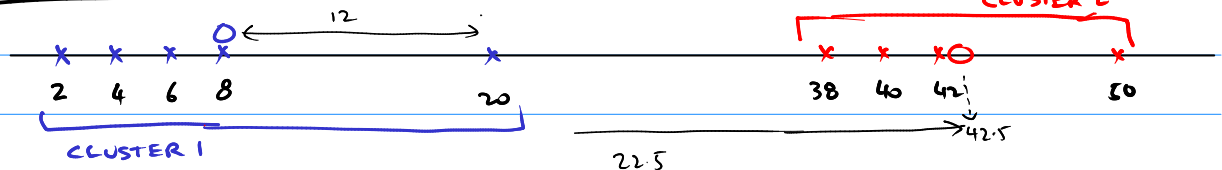
STEP 2: UPDATE MEANS TO AVG OF CLUSTER DATA



STEP 3: CHANGES? YES $\rightarrow$ ROUND 3

ROUND 3

STEP 1: ASSIGN DATA TO NEAREST CLUSTER MEAN



— NO CHANGES TO CLUSTER ASSIGNMENT

STEP 2: NO CHANGES

STEP 3: DONE

3 : 1-D ALGORITHM

S_1 ... S_k ← CLUSTER NAMES
↓ ↓

A: CHOOSE k MEANS (CLUSTER CENTERS) : $m_1, \dots, m_k$

ROUND :

B: ASSIGN DATA TO NEAREST CLUSTER:

$$x_i \xrightarrow{\text{ASSIGNED TO}} S_j, \text{ s.t. } |x_i - m_j| \leq |x_i - m_l|, \text{ for } l=1, \dots, k.$$

↓
mean for S_j

C: RE-ASSIGN MEANS.

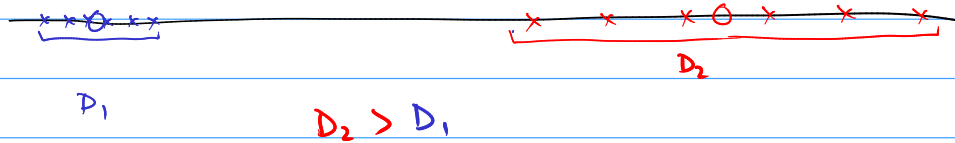
$|S_i|$ ← the number of data points in S_i

$$\rightarrow m_i \triangleq \frac{1}{|S_i|} \sum_{j=1}^{|S_i|} x_{ij} \leftarrow \text{the } j^{\text{th}} \text{ data point in } S_i$$

D. If any changes from prev. round: START NEW ROUND

→ ELSE : STOP. CLUSTERS FOUND.

4. DISTORTION AND ITS MINIMIZATION

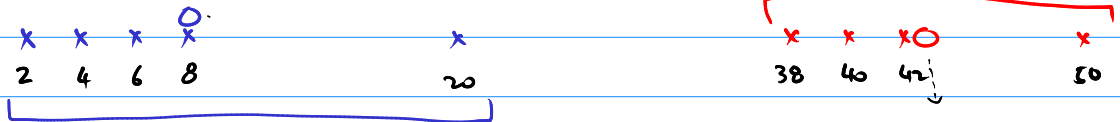


→ DISTORTION OF A CLUSTER i (S_i)

$$D_i = \sum_{j=1}^{|S_i|} (x_{ij} - m_i)^2$$

$$\rightarrow \text{TOTAL DISTORTION: } D = \sum_{i=1}^k D_i$$

→ EXAMPLE:



$$D_1 = (2-8)^2 + (4-8)^2 + (6-8)^2 + (8-8)^2 + (20-8)^2 = 200$$

$$D_2 = (38-42.5)^2 + (40-42.5)^2 + (42-42.5)^2 + (50-42.5)^2 = 83$$

$$D = D_1 + D_2 = 283$$

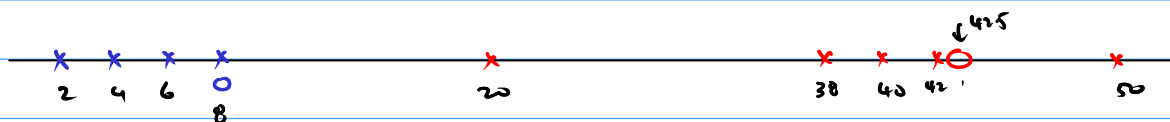
$$\rightarrow \text{Variance of } S_i = \frac{D_i}{|S_i|}$$

→ K-MEANS and DISTORTION

→ STEPS 1 & 2 OF K-MEANS EACH MINIMIZE DISTORTION

→ STEP 1: (MEANS FIXED) DATA ASSIGNED TO NEAREST MEAN

→ CHANGE DATA POINT ASSIGNMENT → INCREASE D



→ $D = 283$; CONTRIBUTION OF 20: $(20 - 8)^2 = 12^2 = 144$

→ Re-assign 20 → RED

→ CONTRIB OF 20: $(20 - 42.5)^2 = 22.5^2 = 506.25 > 144$

→ K-MEANS' STEP 1 MINIMIZES TOTAL DISTORTION ← W MEANS FIXED

→ STEP 2: MEANS SET TO CLUSTER AVGS (CLUSTER ASSIGNMENTS FIXED)

→ THE CHOICE OF MEAN FOR EACH S_i MINIMIZES D_i

→ Proof.

$$\rightarrow D_i = \sum_{j=1}^{|S_i|} (x_{ij} - m_i)^2$$

↳ change this to $p \neq$ avg. of data necessarily

$$\rightarrow \hat{D}_i = \sum_{j=1}^{|S_i|} (x_{ij} - p)^2$$

→ FIND THE MINIMUM OF $\hat{D}_i$ as p changes

$$\rightarrow \frac{\partial \hat{D}_i}{\partial p} = \sum_{j=1}^{|S_i|} -2(x_{ij} - p) = 0 \text{ for a minimum}$$

$$\rightarrow \sum_{j=1}^{|S_i|} x_{ij} - \underbrace{\sum_{j=1}^{|S_i|} p}_{|S_i| p} = 0$$

$$\rightarrow p = \frac{1}{|S_i|} \sum_{i=1}^{|S_i|} x_{ij} \leftarrow \text{avg of the data, i.e., } m_i$$

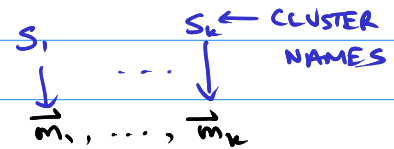
→ (IF THE CLUSTERS ARE FIXED)

→ STEP 2 MINIMIZES D_i for each cluster ✓

HENCE MINIMIZES $\sum D_i = D$

5. K-MEANS FOR GENERAL HIGHER DIMENSIONAL DATA

A: CHOOSE k MEANS (CLUSTER CENTERS): $\vec{m}_1, \dots, \vec{m}_k$



ROUND:

(STEP 1) B: ASSIGN DATA TO NEAREST CLUSTER:

$$\vec{x}_i \mapsto S_j, \text{ s.t. } \|\vec{x}_i - \vec{m}_j\| \leq \|\vec{x}_i - \vec{m}_l\|, \text{ for } l=1, \dots, k.$$

Diagram illustrating the assignment of data point $\vec{x}_i$ to cluster S_j . The label "ASSIGNED TO" points to the arrow from $\vec{x}_i$ to S_j . The label "mean for S_j " points to $\vec{m}_j$.

(STEP 2) C: RE-ASSIGN MEANS.

$$\vec{m}_i \triangleq \frac{1}{|S_i|} \sum_{j=1}^{|S_i|} \vec{x}_{ij}$$

Diagram illustrating the calculation of the new mean $\vec{m}_i$ for cluster S_i . The label " $|S_i|$ " points to the denominator, with a note "the number of data points in S_i ". The label " $\vec{x}_{ij}$ " points to the summation term, with a note "the j th data point in S_i ".

D. If any changes from prev. round: START NEW ROUND)

→ ELSE: STOP. CLUSTERS FOUND.

→ DISTORTION OF A CLUSTER i (S_i)

$$D_i \triangleq \sum_{j=1}^{|S_i|} \|\vec{x}_{ij} - \vec{m}_i\|^2$$

$$\rightarrow \text{TOTAL DISTORTION: } D = \sum_{i=1}^k D_i$$

→ MINIMIZATION

- STEP 1 MINIMIZED D w.r.t CLUSTER ASSIGNMENTS (MEANS CONST)
 - STEP 2 " " " CHOICE OF MEANS (CLUSTERS CONST)
-

6. "FAILURE" OF K-MEANS

