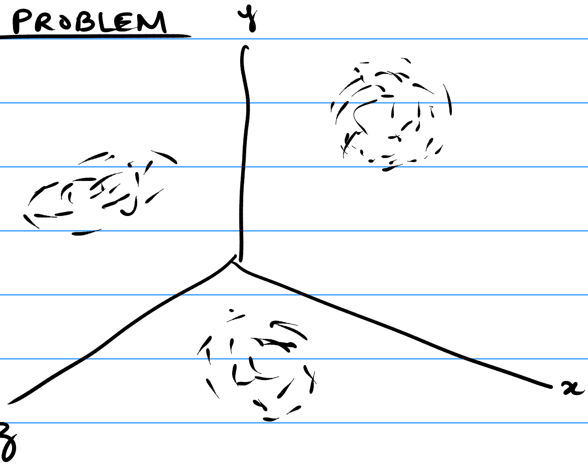
— K-MEANS CLUSTERING

1. THE PROBLEM — 13:17 – 13:29
2. ILLUSTRATION ON A 1-D EXAMPLE — 16:15 – 16:42
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

DATA

1. THE PROBLEM



FINDING CLUSTERS IN DATA

— GIVEN A, a matrix of DATA
— GIVEN k – NUMBER OF CLUSTERS
— FIND THE BEST CLUSTERS

→ APPLICATION: EG, MOVIE CLUSTERING USING RATINGS
— MANY OTHERS

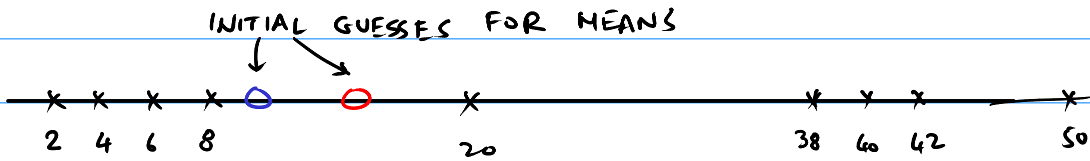
→ PROBLEM VERY DIFFICULT (IN GENERAL): NP-HARD

→ A SOLUTION: K-MEANS CLUSTERING
→ OFTEN FINDS GOOD SOLUTIONS QUICKLY
→ NO GUARANTEE OF GOOD SOLUTION

2. 1D-EXAMPLE OF K-MEANS

- 9 DATA POINTS
- GIVEN $k=2$ (find 2 best clusters)

- STEP 0: PICK $k=2$ GUESSES for CLUSTER CENTERS ("MEANS")

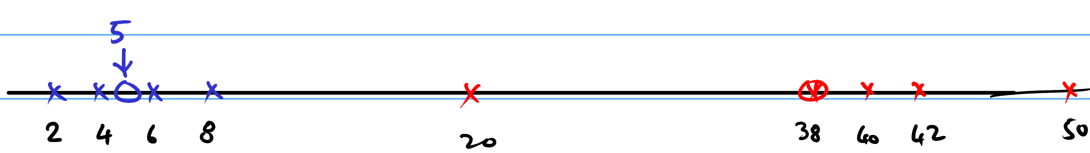


- ROUND 1

STEP 1: ASSIGN DATA TO NEAREST MEAN (CLUSTER)

STEP 2: UPDATE THE MEANS

- TO THE CENTROID/MEAN OF EACH CLUSTER'S DATA



$\frac{190}{5} = 38$

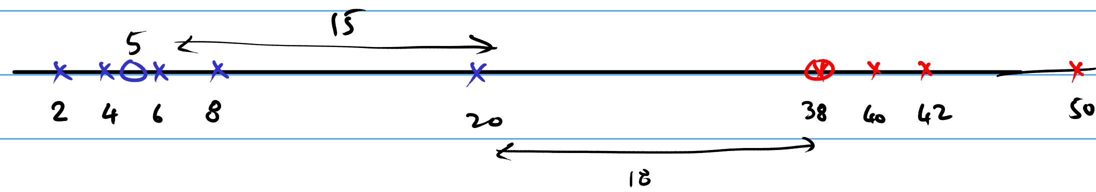
STEP 3 → HAVE CLUSTERS/MEANS CHANGED?

✓ → YES: GO BACK TO STEP 1, NEW ROUND

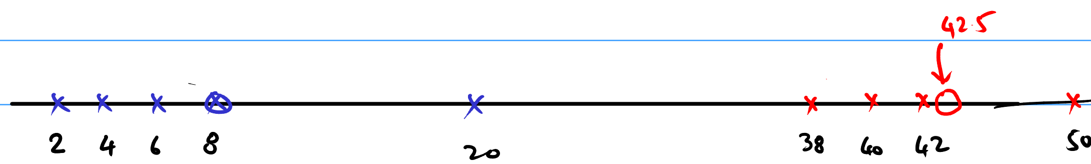
→ NO: DONE!

— ROUND 2

→ STEP 1: ASSIGN DATA TO NEAREST MEAN (CLUSTER)



→ STEP 2: RE-CALCULATE CLUSTER MEANS

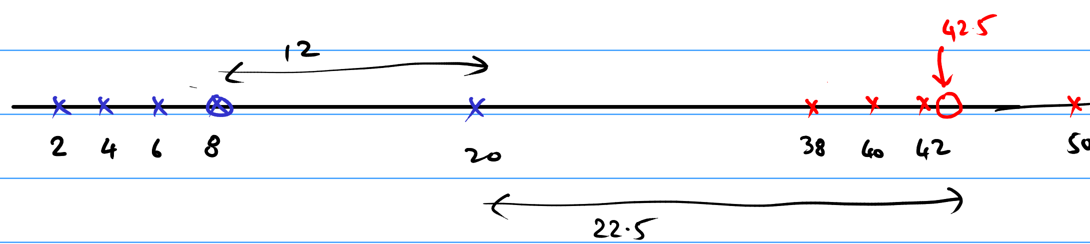


$\frac{40}{5} = 8$ $\frac{170}{4} = 42.5$

→ STEP 3: CLUSTERS CHANGED? YES. → ROUND 3, STEP 1

— ROUND 3:

→ STEP 1: RE-ASSIGN DATA TO NEAREST MEAN



→ NO CHANGE TO CLUSTERS

→ ⇒ (STEP 2) NO CHANGE TO MEANS.

→ DONE!

→ BLUE DATA IS ONE CLUSTER: MEAN = 8

→ RED DATA IS 2nd CLUSTER: MEAN = 42.5

3. THE ALGORITHM FOLLOWED ABOVE

$S_1 \cdots S_k$ ← CLUSTER NAMES

A. PICK k GUESSES FOR MEANS : $m_1, \ldots, m_k$

START ROUND:

B. ASSIGN DATA TO NEAREST MEAN (CLUSTER)

$$\rightarrow \underbrace{x_i \mapsto S_j}_{\text{assigned to}} \quad \text{s.t.} \quad \underbrace{|x_i - m_j|}_{\text{CLOSEST MEAN FOR } x_i} \le |x_i - m_\ell|, \ \ell = 1, \ldots, k$$

C. RE-CALCULATE MEANS

$$\longrightarrow m_i = \frac{1}{|S_i|} \sum_{j=1}^{|S_i|} x_{ij}$$

x_{ij}: jth data point in cluster i

$|S_i|$ → # of elements in S_i (cardinality of S_i)

D. IF ANY CHANGES TO CLUSTERS/MEANS → NEW ROUND, STEP B

→ ELSE: DONE, k CLUSTERS ASSIGNED

---

4. THE DISTORTION METRIC AND ITS MINIMIZATION

— MEASURE OF HOW GOOD THE CLUSTER ASSIGNMENT IS

— THE SMALLER THE BETTER.

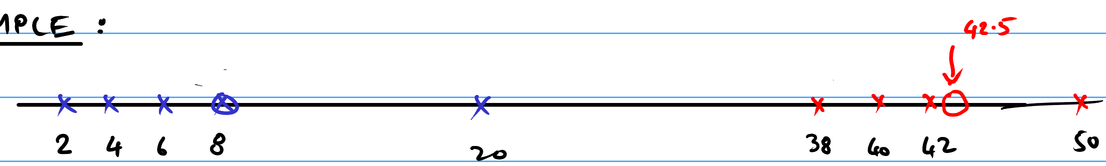
→ DISTORTION OF i^{th} CLUSTER

— $D_i = \sum_{j=1}^{|S_i|} (x_{ij} - m_i)^2$ (SUM OF SQUARES OF DISTANCES TO m_i)

— $D = \sum_{i=1}^{k} D_i$ (SUM OF DISTORTIONS OF ALL CLUSTERS)

D → TOTAL DISTORTION

→ **EXAMPLE:**



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

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

← MEASURE OF SPREAD OF CLUSTER = Variance $\times |S_i|$

$D = D_1 + D_2 = 283$

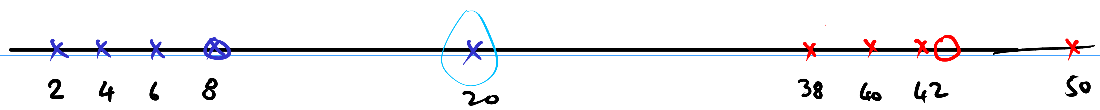
→ EACH STEP OF K-MEANS **REDUCES DISTORTION**

→ **STEP 1**: ASSIGN DATA TO NEAREST MEAN.

→ CHANGING THE ASSIGNMENT OF **ANY ONE** **INCREASES** D_i

↳ i.e., K-MEANS STEP 1 MINIMIZES D OVER CLUSTER ASSIGNMENTS (WITH MEANS FIXED)

— **EXAMPLE**:



→ CONTRIBUTION OF 20 TO D_1 : $(20-8)^2 = 144$

→ LESS THAN $(20-42.5)^2 = 506.25 > 144$

→ MAKING 20 RED WILL **INCREASE** D_2

→ SAME FOR EVERY OTHER POINT

→ HENCE STEP 1 ASSIGNMENTS MINIMIZED D (for MEANS FIXED)

→ STEP 2: ASSIGN MEANS TO AVERAGE OF CLUSTER.

→ THIS CHOICE OF MEANS MINIMIZES D OVER ALL OTHER CHOICES (IF CLUSTERS ARE FIXED)

→ SUPPOSE p IS THE NEW MEAN **INSTEAD**.

$$D_i = \sum_{j=1}^{|S_i|} (x_j - p)^2$$

→ MINIMUM OF D_i WRT p : $\frac{\partial D_i}{\partial p} = 0$

$$\rightarrow \frac{\partial D_i}{\partial p} = \sum_{j=1}^{|S_i|} -2(x_j - p) = 0$$

$$\Longrightarrow |S_i|\, p = \underbrace{\sum_{j=1}^{|S_i|} x_j}_{\text{this is } m_i}$$

→ HENCE $p = m_i$ minimizes D_i

→ TRUE FOR EVERY CLUSTER, HENCE $D = \sum_{i=1}^{k} D_i$ IS MINIMIZED.

5. THE GENERAL k-MEANS ALGORITHM (HIGHER-D DATA)

A. PICK k GUESSES FOR MEANS : $\underset{S_1}{\vec{m}_1}, \ldots, \underset{S_k}{\vec{m}_k} \rightarrow$ q-d

S_1 ... S_k ← CLUSTER NAMES

→ START ROUND

B. ASSIGN DATA TO NEAREST MEAN (CLUSTER)

$$\rightarrow \quad \vec{x}_i \underset{\text{assigned to}}{\longmapsto} S_j \quad \text{s.t.} \quad \underbrace{\|\vec{x}_i - \vec{m}_j\|}_{\text{CLOSEST MEAN FOR } \vec{x}_i} \leq \|\vec{x}_i - \vec{m}_\ell\|, \quad \ell = 1, \ldots, k$$

C. RE-CALCULATE MEANS

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

$\vec{x}_{ij}$ ← jth data point in cluster i

$|S_i|$ → # of elements in S_i (cardinality of S_i)

D. IF ANY CHANGES TO CLUSTERS / MEANS → NEW ROUND, STEP B.

→ ELSE: STOP. k CLUSTERS ASSIGNED.

→ DISTORTION :

— D_i = SUM OF SQUARES OF DISTANCES TO MEAN (FOR i^{th} CLUSTER)

D_i → DISTORTION OF i^{th} CLUSTER

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

— $D = \sum_{i=1}^{k} D_i$ (SUM OF DISTORTIONS OF ALL CLUSTERS)

D → TOTAL DISTORTION

→ SAME MINIMIZATION PROPERTIES FOR STEPS 1 & 2.

— NUMERICAL DEMOS ⟶ SLIDES

---

6. "FAILURE" OF K-MEANS

→ K-MEANS DOESN'T ALWAYS DO A GOOD JOB

$k=3$

⟶ DEMOS OF BAD K-MEANS: ⟶ SLIDES

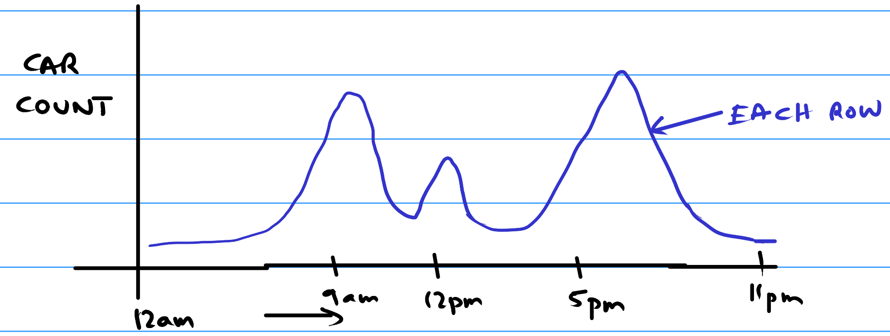
---

7. EXAMPLE APPLICATION: TRAFFIC PATTERNS

⟶ MONITOR TRAFFIC ACROSS AN INTERSECTION EVERY HOUR

	12-1am	1-2am	- - - -	11pm-12am
M	CAR COUNT	CAR COUNT		CAR COUNT
T			...	
W				
⋮			⋮	
F				
Sa			...	
Su				
			⋮	

— EACH ROW CAN BE PLOTTED:



→ TRY CLUSTERING THE DATA (A PAPER DID SO)

— 164 DAYS; INTERSECTION IN S. CAROLINA

→ WITH $k=4$, THE CLUSTERS WERE:

M–Th, F, Sa, Su

→ (SHOW SLIDES)

→ HOW TO CHOOSE k?

→ ELBOW CURVE METHOD:

— RUN K-MEANS FOR $k = 2, 3, 4, \ldots$

— PLOT DISTORTION AGAINST k