2. HAC using euclidean

A : (0, 3, 12)
B : (2, 8)
C : (15)

Dist between A & B

$\Rightarrow \min\left(\sqrt{(0-2)^2}, \sqrt{(0-8)^2}, \sqrt{(3-2)^2}, \sqrt{(3-8)^2}, \sqrt{(12-2)^2}, \sqrt{(12-8)^2}\right)$

$= \min(2, 8, 1, 5, 10, 4) = 1$

Dist between A & C

$\Rightarrow \min\left(\sqrt{(0-15)^2}, \sqrt{(3-15)^2}, \sqrt{(12-15)^2}\right)$

$= \min(15, 12, 3) = 3$

Dist between B & C

$\Rightarrow \min\left(\sqrt{(2-15)^2}, \sqrt{(8-15)^2}\right)$

$= \min(13, 7) = 7$

The smallest is 1 so A & B is going to merge in next iteration

4. K-means using Euclidean

(i) Data points: (-2, 1, 9, 18)

Cluster centres: $C_1 = 3$

$C_2 = 21$

K = 2

$\sqrt{(x_2 - x_1)^2}$

Data points (x_1)	Centres $x_1 = 3$ (C_1)	$x_2 = 21$ (C_2)	Cluster
-2	5	529	C_1
1	4	400	C_1
9	36	144	C_1
18	225	9	C_2

(ii) New clusters

For C_1: $\frac{-2+1+9}{3} = 2.66 = 2.67$

For C_2: $\frac{18}{1} = 18$

Q.5 KNN Classifier using euclidean

(a.)

X	Y	Class
0	0	-1
3	7	1
7	3	1
6	3	-1
2	5	-1
2	4	X

Euclidean : $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Dist between:

(2,4) & (0,0) $= \sqrt{2^2 + 4^2} = \sqrt{20} = 4.47$

(2,4) & (7,7) $= \sqrt{(2-7)^2 + (4-7)^2} = \sqrt{10} = 3.16$

(2,4) & (3,3) $= \sqrt{1^2 + 1^2} = \sqrt{2} = 1.41$

(2,4) & (6,7) $= \sqrt{4^2 + 1^2} = \sqrt{17} = 4.12$

(2,4) & (2,5) $= \sqrt{0 + 1^2} = 1 = 1$

Sort with ascending & choosing first 3 as $k=3$

(2,4) & (2,5) → -1
(2,4) & (7,3) → 1
(2,4) & (7,7) → 1

So we classify as 1

(b)

knn = KNeighborsClassifier(n_neighbors=10, metric='cosine')

n_neighbor: Number of neighbors to use for kneighbors queries. Here 10 neighbors are assumed

Metric: Used for distance computation. Here cosine is used.

Here classifier model will choose 10 nearest point using cosine similarity and assign class label that is most common among those 10 neighbors.

6. Choose number of eigen values

(a.)

1. Total variance method

→ Calculate a % of all eigen values in total & set a % of a cutoff like 80%.

2. Scree plot

→ Choose the components where elbow is observed.

3. Pick eigen values >1.

(b.) For PC1 & PC2

	PC 1	PC 2
~~Piped~~ (a)	✓	
~~Satisfied~~ (b)	✓	
~~Satisfaction~~ (c)		✓
(d)		✓
(e)	✓	
(f)	✓	
(g)	✓	
(h)	✓	
(i)		✓
(j)		✓
(k)		✓
(l)		✓

for	P(1) & P(2)	P(2) & P(3)	P(3)
	P(1)	P(2)	P(3)
(a)	✓		
(b)			✓
(c)			✓
(d)		✓	
(e)			✓
(f)	✓		
(g)	✓		
(h)			✓
(i)			✓
(j)		✓	
(k)		✓	
(l)			✓

Yes, it changes as now we look at P(3) along with P(1) & P(2) to look for latent variables being assigned.

7. Association

Support = 50% & confidence = 70%.

Transaction ID	Items purchased
1	BK1, BK2, BK3
2	BK2, BK3, BK4
3	BK2, BK5, BK6
4	BK2, BK4, BK6
5	BK1, BK3, BK6

Step 1
Freq table

Item	frequency
BK1	2
BK2	4
BK3	3
BK4	2
BK5	1
BK6	3

min support count = min support x item count
= 50% x 6
= 3

Step 2:
Support check:

Item	Freq	Support	
BK1	2	3	x
BK2	4	3	✓
BK3	3	3	✓
BK4	2	3	x
BK5	1	3	x
BK6	3	3	✓

Step 3:
Two item set frequency

Item	Freq
BK2, BK3	2
BK2, BK6	2
BK3, BK6	1

Step 4:

Two items support check

2 BK / 3 / 6 BK

Item	freq	Support 50% x3
BK2, BK3	2	✓
BK2, BK6	2	✓
BK3, BK6	1	x

Step 5:

Three item Support check:

Item	freq	Support
BK2, BK3, BK6	0	X

50% X 3 = 1.5

Min confidence = 70%

Confidence (A → B) = P(Y|X) = P(X ∪ Y) / P(X)

P(X ∪ Y): number of times both items were bought together

P(X): no of times this item was bought

Candidate rules are

→ (BK2, BK3), (BK2, BK6)

BK2 > BK3 ⇒ 2/4 ⇒ 50% (count of BK2, BK3 / count of BK2) X

BK3 > BK2

⇒ 2/3 = 66.67% X

BK2 > BK6 ⇒ 2/4 ⇒ 50% X

BK6 > BK2 ⇒ 2/3 = 66.67% X

8. Collaborative filtering

	BK1	BK2	BK3	BK4	BK5
Alvin	4	3	4	2	
Bob	5		1		2
Calvin	3	5	3	3	1
Davy	1	1	5	4	
Edwin	2		2	3	3

(A) Alvin & (B) Calvin

A	B	AB	A^2	B^2
4	3	12	16	9
3	5	15	9	25
4	3	12	16	9
2	3	6	4	9
0	1	0	0	1
Σ 13	15	45	45	

$\Rightarrow \sqrt{45} = 6.708$

$\sqrt{53} = 7.280$

$\times$ 48.834

Cosine similarity $= (A, B) = \dfrac{(\Sigma AB)}{\sqrt{\Sigma A^2}\sqrt{\Sigma B^2}}$

$= \dfrac{45}{48.834} = 0.921$

Value close to 0 : orthogonal or perpendicular

Value close to 1 : the angle is small & images are more similar

Item - Item Collaborative filtering

	BK1	BK3	BK4
Alwin	4	4	2
Bob	5	1	x
Calvin	3	2	3
Davy	1	5	4
Edwin	2	2	3

Cosine for BK4 & BK1

	A	B	AB	A^2	B^2
Alwin	4	2	8	16	4
Bob	5	0	0	25	0
Calvin	3	3	9	9	9
Davy	1	4	4	1	16
Edwin	2	3	6	4	9
Σ	15	15	27	55	38
				7.416	6.164

× 45.71

Cosine = 0.59

Cosine for BK4 & BK3

A	B	AB	A^2	B^2
4	2	8	16	4
1	0	0	1	0
3	3	9	9	9
5	4	20	25	16
2	3	6	4	9
		43	7.416	6.164

× 45.71

Cosine = 0.94

Weighting

For BK1 = 0.59 × 5 = 2.95

For BK3 = 0.94 × 1 = 0.94

estimated rating

$$= \frac{(\text{Weight of BK1} + \text{Weight of BK3})}{\text{Sum of similarities (positive val)}}$$

$$= \frac{2.95 + 0.94}{0.94 + 0.59} = 2.54 \approx 3$$