1. a) Support threshold : 0.4 7 buckets

$\Rightarrow$ Support $= \lceil 0.4 \times 7 \rceil = \lceil 2.8 \rceil = 3$

$\Rightarrow$ Frequent items(s) : support ≥ 3

ID	Baskets
1	a,b,c,e
2	a,d,b
3	c,b
4	a,b,d,e
5	b,d
6	a,b
7	a

Pass 1

Counters in Memory

1. a. b. c. e — a:1 b:1 c:1 e:1
2. a. d. b — a:2 b:2 c:1 d:1 e:1
3. c. b — a:2 b:3 c:2 d:1 e:1
4. a. b. d. e — a:3 b:4 c:2 d:2 e:2
5. b. d — a:3 b:5 c:2 d:3 e:2
6. a. b — a:4 b:6 c:2 d:3 e:2
7. a — a:5 b:6 c:2 d:3 e:2

∴ Frequent items with cardinality 1 : {a}, {b}, {d}

ID	Baskets
1	a,b,c,e
2	a,d,b
3	c,b
4	a,b,d,e
5	b,d
6	a,b
7	a

Pass 2. Frequent items: a. b. d

Counter in Memory

1. a.b.c.e : ab:1
2. a. d. b : ab:2 ad:1 bd:1
3. c.b : None.
4. a.b.d.e : ab:3 ad:2 bd:2
5. b. d : ab:3 ad:2 bd:3
6. ab : ab:4 ad:2 bd:3
7. a : None

∴ Frequent items with cardinality 2 : {ab}, {b.d}

ID	Baskets
1	a,b,c,e
2	a,d,b
3	c,b
4	a,b,d,e
5	b,d
6	a,b
7	a

Pass 3. Frequent items: {a.b} . {b.d} (⇒ only {a.b.d} can be frequent)

1. a.b.c.e : None

2. a.d.b : abd : 1

3. cb : None

4. a.b.d.e : a.b.d : 1

5. b.d : None

6. a.b 、 7. a : None.

So Frequent items: Cardinality 1 : {a.} {b} {d}

Cardinality 2 : {a b} . {b.d}

1.b) Total Number of buckets: 7

Suppor {b} : 6. Suppor {b.d} : 3

∴ The Suppor of {b.d} = $\frac{3}{7}$

And The Confidence of {b.d} = $\frac{\text{Suppor of } \{b.d\}}{\text{Suppor of } \{b\}} = \frac{3}{6} = 0.5$

It's enough to answer with the answer in 1 a) because We {b} and {b.d} are all frequent items. and we have counters for both of them. So we can Calculate the Suppor and also Confidence.

1.c) threshold = 0.33 i.e Suppor of frequent items ≥ $\lceil 6 \times 0.33 \rceil = 2$

ID	Baskets
1	1,3,4
2	4, 5
3	2,7
4	1,6
5	2,7
6	3

Pass 1.	Counters for items	Counters for buckets
1) 1.3.4	1 : 1 3 : 1 4 : 1	B_1: 2 B_2: 1
2) 4, 5	1 : 1 2 : 1 4 : 2 5 : 1	B_0: 1 B_1: 2 B_2: 1
3) 2.7	1 : 1 2 : 1 3 : 1 4 : 2 5 : 1 7 : 1	B_0: 2 B_1: 2 B_2: 1
4) 1.6	1 : 2 2 : 1 3 : 1 4 : 2 5 : 1 6 : 1 7 : 1	B_0: 2 B_1: 3 B_2: 1
5) 2.7	1 : 2 2 : 2 3 : 1 4 : 2 5 : 1 6 : 1 7 : 2	B_0: 3 B_1: 3 B_2: 1

6) 3 1:2 2:2 3:2 4:2 5:1 6:1 7:2 B_0:3 B_1:3 B_2:1

Bucket 6 only has 1 item so no need to calculate Hash function to map to 2-cardinality candidate set.

Frequent items: {1} {2} {3} {4} {7}

ID	Baskets
1	1,3,4
2	4,5
3	2,7
4	1,6
5	2,7
6	3

Pass 2: Frequent items: 1. 2. 3. 4. 7

Counters for Buckets: B_0:3 B_1:3 B_2:1

$\Rightarrow$ B_0 and B_1 are frequent buckets B_2 is not

1. 1.3. 4 : $f_{C(1,3)} = f_{C(3,4)} = 1 \Rightarrow$ Counters: (1.3):1 (3.4):1

$f_{C(1,4)} = 2 \Rightarrow$ No frequent.

2. 4.5 : $f_{(4,5)} = 0 \Rightarrow$ Counters: (1.3):1 (3.4):1 (4.5):1

3. 2.7 : $f_{(2,7)} = 0 \Rightarrow$ Counters: (1.3):1 (3.4):1 (4.5):1 (2.7):1

4. 1.6 : $f_{(1,6)} = 1 \Rightarrow$ Counters: (1.3):1 (3.4):1 (4.5):1 (2.7):1

(1.6):1

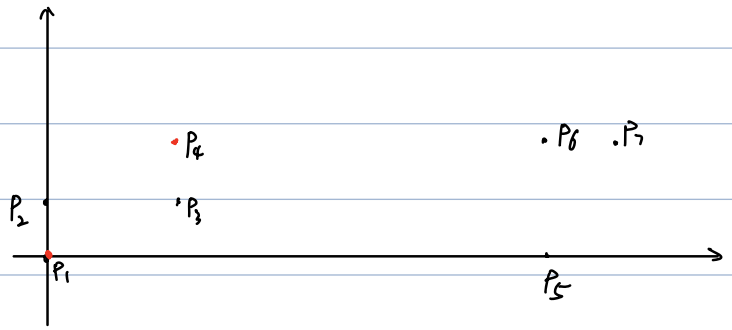
5. 2.7 : $f_{(2,7)} = 0 \Rightarrow$ Counters: (1.3):1 (3.4):1 (4.5):1 (2.7):2

(1.6):1

C_0, Frequent items with cardinality 2: {2.7}

Frequent items: {1} {2} {3} {4} {7} {2.7}

1. We are given the following points in the 2-dimensional euclidean space. P1=(0,0), P2=(0,1/2), P3=(1,1/2), P4=(1,1), P5=(4,0), P6=(4,1), P7=(5,1). Suppose that P1 = (0, 0) and P4 = (1, 1) are chosen as initial centroids for the K-means algorithm, K=2. Show step by step the clustering you would obtain by running K-Means on the previous set of points, while specifying for each clustering the current set of centroids. Recall that the algorithm terminates when the current set of centroids does not change.



1. $C_1 = (0,0)$ $C_2 = (1,1)$

Iteration 1:

P_1: $d(P_1, C_1) = 0$ $d(P_1, C_2) = \sqrt{2}$ $\Rightarrow$ P_1 to C_1

P_2: $d(P_2, C_1) = 0.5$ $d(P_2, C_2) = \frac{\sqrt{5}}{2}$ $\Rightarrow$ P_2 to C_1

P_3 $d(P_3, C_1) = \frac{\sqrt{5}}{2}$ $d(P_3, C_2) = 0.5$ $\Rightarrow$ P_3 to C_2

P_4: $d(P_4, C_1) = \sqrt{2}$ $d(P_4, C_2) = 0$ $\Rightarrow$ P_4 to C_2

P_5: $d(P_5, C_1) = 4$ $d(P_5, C_2) = \sqrt{10}$ $\Rightarrow$ P_5 to C_2

P_6: $d(P_6, C_1) = \sqrt{17}$ $d(P_6, C_2) = 3$ $\Rightarrow$ P_6 to C_2

P_7: $d(P_7, C_1) = \sqrt{26}$ $d(P_7, C_2) = 4$ $\Rightarrow$ P_7 to C_2

$\therefore$ C_1 cluster: P_1, P_2

C_2 cluster: P_3, P_4, P_5, P_6, P_7

C_1 update: $(0, 0.25)$ C_2 update: $(3, 0.7)$

Iteration 2.

P_1: $d(P_1, C_1) = 0.25$ $d(P_1, C_2) = \sqrt{3^2 + 0.7^2} \simeq 3.05$ $\Rightarrow P_1$ to C_1

P_2: $d(P_2, C_1) = 0.25$ $d(P_2, C_2) \simeq 3.01$ $\Rightarrow P_2$ to C_1

P_3 $d(P_3, C_1) \simeq 1.03$ $d(P_3, C_2) \simeq 2.02$ $\Rightarrow P_3$ to C_1

P_4: $d(P_4, C_1) \simeq 1.27$ $d(P_4, C_2) \simeq 2.02$ $\Rightarrow P_4$ to C_1

P_5: $d(P_5, C_1) \simeq 4.01$ $d(P_5, C_2) \simeq 1.22$ $\Rightarrow P_5$ to C_2

P_6: $d(P_6, C_1) \simeq 4.13$ $d(P_6, C_2) \simeq 1.04$ $\Rightarrow P_6$ to C_2

P_7: $d(P_7, C_1) \simeq 5.03$ $d(P_7, C_2) \simeq 2.02$ $\Rightarrow P_7$ to C_2

C_1 cluster: P_1, P_2, P_3, P_4

C_2 cluster: P_5 P_6 P_7

C_1 update: $(0.5, 0.5)$ $C_2(4.33, 0.67)$

Iteration 3: P_1: $d(P_1, C_1) \approx 0.71$ $d(P_1, C_2) \approx 4.38$ $\Rightarrow P_1$ to C_1

P_2: $d(P_2, C_1) = 0.5$ $d(P_2, C_2) \approx 4.33$ $\Rightarrow P_2$ to C_1

P_3: $d(P_3, C_1) \approx 0.5$ $d(P_3, C_2) \approx 3.34$ $\Rightarrow P_3$ to C_1

P_4: $d(P_4, C_1) \approx 0.71$ $d(P_4, C_2) \approx 3.35$ $\Rightarrow P_4$ to C_1

P_5: $d(P_5, C_1) \approx 3.58$ $d(P_5, C_2) \approx 0.73$ $\Rightarrow P_5$ to C_2

P_6: $d(P_6, C_1) \approx 3.61$ $d(P_6, C_2) \approx 0.47$ $\Rightarrow P_6$ to C_2

P_7: $d(P_7, C_1) \approx 4.53$ $d(P_7, C_2) \approx 0.75$ $\Rightarrow P_7$ to C_2

C_1 cluster: P_1, P_2, P_3, P_4

C_2 cluster: P_4, P_5, P_6

C_1 update: $(0.5, 0.5)$ C_2 update $(4.33, 0.67)$

No change! Algorithm Terminates.

2. Provide an example for which the K-means algorithm produces at least an empty cluster, that is, the number of non-empty clusters is $< K$. Your example should contain at most 6 points in a one-dimension Euclidean space, while the number of points should not be smaller than K. We recall that the initial centroids are always chosen among the input points. Show all the steps of the algorithm until it terminates.

Suppose: Data points:

$P_1 = 0$ $P_2 = 8$ $P_3 = 9$ $P_4 = 10$ $P_5 = 19$ $P_6 = 20$

And Centroids: $C_1 = P_1 = 0$ $C_2 = P_5 = 19$ $C_3 = P_6 = 20$ $(k = 3)$

Then: Iteration 1:

$d(P_1, C_1) = 0$ $d(P_1, C_2) = 19$ $d(P_1, C_3) = 20$ $\Rightarrow P_1$ to C_1

$d(P_2, C_1) = 8$ $d(P_2, C_2) = 11$ $d(P_2, C_3) = 12$ $\Rightarrow P_2$ to C_1

$d(P_3, C_1) = 9$ $d(P_3, C_2) = 10$ $d(P_3, C_3) = 11$ $\Rightarrow P_3$ to C_1

$d(P_4, C_1) = 10$ $d(P_4, C_2) = 9$ $d(P_4, C_3) = 10$ $\Rightarrow P_4$ to C_2

$d(P_5, C_1) = 19$ $d(P_5, C_2) = 0$ $d(P_5, C_3) = 1$ $\Rightarrow P_5$ to C_2

$d(P_6, C_1) = 20$ $d(P_6, C_2) = 1$ $d(P_6, C_3) = 0$ $\Rightarrow P_6$ to C_3

C_1 cluster: P_1, P_2, P_3 C_2 cluster: P_4, P_5, C_3 cluster: P_6

C_1 update: $\frac{8+9}{3} = \frac{17}{3} \approx 5.66$

C_2 update: $\frac{10+19}{2} = \frac{29}{2} = 14.5$

C_3 update: 20

Iteration 2:

$d(P_1, C_1) = 5.66$ $d(P_1, C_2) = 14.5$ $d(P_1, C_3) = 20$ $\Rightarrow P_1$ to C_1

$d(P_2, C_1) = 2.34$ $d(P_2, C_2) = 4.5$ $d(P_2, C_3) = 12$ $\Rightarrow P_2$ to C_1

$d(P_3, C_1) = 3.34$ $d(P_3, C_2) = 5.5$ $d(P_3, C_3) = 11$ $\Rightarrow P_3$ to C_1

$d(P_4, C_1) = 4.34$ $d(P_4, C_2) = 4.5$ $d(P_4, C_3) = 10$ $\Rightarrow P_4$ to C_1

$d(P_5, C_1) = 13.34$ $d(P_5, C_2) = 4.5$ $d(P_5, C_3) = 1$ $\Rightarrow P_5$ to C_3

$d(P_6, C_1) = 14.34$ $d(P_6, C_2) = 5.5$ $d(P_6, C_3) = 0$ $\Rightarrow P_6$ to C_3

Then: C_1 cluster: P_1 P_2 P_3 P_4

C_2 cluster: NULL

C_3 cluster: P_5, P_6

Then C_2 becomes an empty cluster.

C_1 update: $C_1 = \frac{8+9+10}{4} = 6.75$

$C_2 =$ NULL

$C_3 = \frac{19+20}{2} = 19.5$

Iteration 3:

$d(P_1, C_1) = 6.75$ $d(P_1, C_2) =$ NULL $d(P_1, C_3) = 19.5$ $\Rightarrow P_1$ to C_1

$d(P_2, C_1) = 1.25$ $d(P_2, C_2) =$ NULL $d(P_2, C_3) = 11.5$ $\Rightarrow P_2$ to C_1

$d(P_3, C_1) = 2.25$ $d(P_3, C_2) =$ NULL $d(P_3, C_3) = 10.5$ $\Rightarrow P_3$ to C_1

$d(P_4, C_1) = 3.25$ $d(P_4, C_2) =$ NULL $d(P_4, C_3) = 9.5$ $\Rightarrow P_4$ to C_1

$d(P_5, C_1) = 12.25$ $d(P_5, C_2) =$ NULL $d(P_5, C_3) = 0.5$ $\Rightarrow P_5$ to C_3

$d(P_6, C_1) = 13.25$ $d(P_6, C_2) =$ NULL $d(P_6, C_3) = 0.5$ $\Rightarrow P_6$ to C_3

C_1 cluster: P_1 P_2 P_3 P_4 C_2 cluster: P_5 P_6

C_1 update: No change C_2 update: No change

Then Algorithm Terminates