COMP7103B Data Mining Assignment 1

Lam Wun Yin (3035372505)

Question (1a)

Pass 1

In the first pass, count the occurrences of each individual item.

Itemset	Occurrences
a	5
b	6
c	2
d	3
e	2

Frequent itemsets in Pass 1 (at least three times): {a}, {b}, {d}

Pass 2

Generate candidate itemsets of size 2 and count their occurrences. Since {c}, {e} is not frequent so none of itemsets with them can be frequent.

Itemset	Occurrences
{a, b}	4
{a, d}	2
{b, d}	3

Frequent itemsets in Pass 2 (at least three times): {a, b}, {b, d}

Pass 3

There are only 2 frequent candidate itemsets of size 2. However, for a possible frequent candidate itemset of size 3, there must be at least 3 candidate itemsets of size 2. Therefore, it should marks the end of the algorithm.

Final Result

The frequent itemsets with a support threshold of 0.4 are:
{a}, {b}, {d}, {a, b}, {b, d}

Question (1b)

To compute the support and confidence of the rule b → d:

Basket containing {b, d} = 3
Basket containing {b} = 6

Confidence {b → d} = 3 / 6 = 0.5
Support {b → d} = 3 / Total number of baskets = 3 / 7 = 0.4286

Therefore, the confidence of the rule b → d is 0.5, and the support is 0.4286.

Question (1c)

Pass 1

In the first pass, count the occurrences of each individual item and counters for the buckets by hash function.

Itemset	Occurrences
1	2
2	2
3	2
4	2
5	1
6	1
7	2

Bucket	Occurrences
B0	3
B1	3
B2	1

Frequent itemsets in Pass 1 (at least two times): {1}, {2}, {3}, {4}, {7}
Frequent buckets: {B0}, {B1}

Pass 2

Generate candidate itemsets of size 2 and count their occurrences. Since {5}, {6} and hash function returns 2 are not frequent so none of itemsets with them can be frequent.

Itemset	Occurrences
{1, 3}	1
{3, 4}	1
{2, 7}	2

Frequent itemsets in Pass 2 (at least two times): {2, 7}

Final Result

The frequent itemsets with a support threshold of 0.33 are:
{1}, {2}, {3}, {4}, {7}, {2, 7}

Question (2) (1)

Step 1: Assigning Points to Clusters

Points	Distance to C1 (0, 0)	Distance to C2 (1, 1)	Clusters
P1 = (0, 0)	0	1.41	C1
P2 = (0, 1/2)	0.5	1.12	C1
P3 = (1, 1/2)	1.12	0.5	C2
P4 = (1, 1)	1.41	0	C2
P5 = (4, 0)	4	3.16	C2
P6 = (4, 1)	4.12	3	C2
P7 = (5, 1)	5.10	4	C2

Step 2: Updating Centroids

Clusters	Points	Centroid
C1	P1 = (0, 0) P2 = (0, 1/2)	(0, 0.25)
C2	P3 = (1, 1/2) P4 = (1, 1) P5 = (4, 0) P6 = (4, 1) P7 = (5, 1)	(3, 0.7)

Step 3: Checking Centroid Stability

Points	Distance to C1 (0, 0.25)	Distance to C2 (3, 0.7)	Clusters
P1 = (0, 0)	0.25	3.08	C1
P2 = (0, 1/2)	0.25	3.01	C1
P3 = (1, 1/2)	1.03	2.01	C1
P4 = (1, 1)	1.25	2.02	C1
P5 = (4, 0)	4.01	1.22	C2
P6 = (4, 1)	4.07	1.04	C2
P7 = (5, 1)	5.06	2.02	C2

Step 4: Updating Centroids Again

Clusters	Points	Centroid
C1	P1 = (0, 0) P2 = (0, 1/2) P3 = (1, 1/2) P4 = (1, 1)	(0.5, 0.5)
C2	P5 = (4, 0) P6 = (4, 1) P7 = (5, 1)	(13/3, 2/3)

Step 5: Checking Centroid Stability

Points	Distance to C1 (0.5, 0.5)	Distance to C2 (13/3, 2/3)	Clusters
P1 = (0, 0)	0.71	4.38	C1
P2 = (0, 1/2)	0.5	4.34	C1
P3 = (1, 1/2)	0.5	3.34	C1
P4 = (1, 1)	0.71	3.35	C1
P5 = (4, 0)	3.54	0.75	C2
P6 = (4, 1)	3.54	0.47	C2
P7 = (5, 1)	4.53	0.75	C2

Since the assignments are the same as before, we can conclude that the centroids have stabilized, and the algorithm terminates.

Question (2) (2)

Let's consider an example with K = 3 and 6 points in a one-dimensional space.

Points: P1 = 1, P2 = 10, P3 = 11, P4 = 12, P5 = 22, P6 = 23

Step 1: Initial Centroid Selection

Randomly select K initial centroids from the input points.

Initial Centroids: C1 = 1 (P1), C2 = 22 (P5), C3 = 23 (P6)

Step 2: Assigning Points to Clusters

Points	Distance to C1 (1)	Distance to C2 (22)	Distance to C3 (23)	Clusters
P1 = 1	0	21	22	C1
P2 = 10	9	12	13	C1
P3 = 11	10	11	12	C1
P4 = 12	11	10	11	C2
P5 = 22	21	0	1	C2
P6 = 23	22	1	0	C3

Step 3: Updating Centroids

Clusters	Points	Centroid
C1	P1 = 1 P2 = 10 P3 = 11	7.33
C2	P4 = 12 P5 = 22	17
C3	P6 = 23	23

Step 4: Checking Centroid Stability

Points	Distance to C1 (7.33)	Distance to C2 (17)	Distance to C3 (23)	Clusters
P1 = 1	6.33	16	22	C1
P2 = 10	2.67	7	13	C1
P3 = 11	3.67	6	12	C1
P4 = 12	4.67	5	11	C1
P5 = 22	14.67	5	1	C3
P6 = 23	15.67	6	0	C3

Step 5: Updating Centroids Again

Clusters	Points	Centroid
C1	P1 = 1 P2 = 10 P3 = 11 P4 = 12	8.5
C2	/	Undefined
C3	P5 = 22 P6 = 23	22.5

Step 6: Checking Centroid Stability

Points	Distance to C1 (8.5)	Distance to C2 (Undefined)	Distance to C3 (22.5)	Clusters
P1 = 1	7.5	Undefined	21.5	C1
P2 = 10	1.5	Undefined	12.5	C1
P3 = 11	2.5	Undefined	11.5	C1
P4 = 12	3.5	Undefined	10.5	C1
P5 = 22	13.5	Undefined	0.5	C3
P6 = 23	14.5	Undefined	0.5	C3

Since the assignments are the same as before, we can conclude that the centroids have stabilized, and the algorithm terminates.

The final clustering is as follows:

Clusters	Points	Centroid
C1	P1 = 1 P2 = 10 P3 = 11 P4 = 12	8.5
C2	/	Undefined
C3	P5 = 22 P6 = 23	22.5

In this example, the K-means algorithm produces an empty cluster (Cluster 2) since it does not have any assigned points.