Assignment1

Question1: Frequent Itemsets

a

Total Number of Baskets (transactions): 7

Support Threshold: 0.4 * 7 = **3** (meaning an itemset must appear at least 3 times to be considered frequent).

Summary of Frequent Itemsets

- **Frequent 1-itemsets**:
- {a}: 5, {b}: 6, {d}: 3
- **Frequent 2-itemsets**:
- {a, b}: 4, {b, d}: 3
- **Frequent 3-itemsets**: No frequent 3-itemsets exist in this dataset.

b

Information

- **Support**:
  - (Support(b) = 6) (Item (b) appears 6 times)
  - (Support(b, d) = 3) (Itemset {b, d} appears 3 times)

Calculating Support and Confidence

1. **Support**:
   - Support(b, d) = 3
2. **Confidence**:
   - Confidence(b —> d) = 3 / 6 = 0.5

Summary of Results

- **Support for (b —> d)**: 3
- **Confidence for (b —> d)**: 0.5 (or 50%)

Conclusion

From the previous steps, we have sufficient information to compute the support and confidence of the rule (b —> d). All necessary counts can be derived from the previously identified frequent itemsets, so there is no missing information.

c

Summary of Frequent Itemsets

- **Frequent 1-itemsets**:
- {1}: 2, {2}: 2, {3}: 2, {4}: 2, {7}: 2
- **Frequent 2-itemsets**:
- {2, 7}: 2

Question2: K-means

a

Initial Setup

- **Points**: P1=(0, 0), P2=(0, 0.5), P3=(1, 0.5), P4=(1, 1), P5=(4, 0), P6=(4, 1), P7=(5, 1)
- **Initial Centroids**: C1=(0, 0) (P1), C2=(1, 1) (P4)

Step 1: Assign Points to the Nearest Centroid

Clusters:

- **Cluster 1 (C1)**: P1=(0, 0), P2=(0, 0.5)
- **Cluster 2 (C2)**: P3=(1, 0.5), P4=(1, 1), P5=(4, 0), P6=(4, 1), P7=(5, 1)

Step 2: Calculate New Centroids

New Centroids:

- **C1**: Mean of P1 and P2 = (0, 0.25)
- **C2**: Mean of P3, P4, P5, P6, P7 = (3, 0.7)

Step 3: Reassign Points to the Nearest Centroid

Clusters:

- **Cluster 1 (C1)**: P1=(0, 0), P2=(0, 0.5), P3=(1, 0.5), P4=(1, 1)
- **Cluster 2 (C2)**: P5=(4, 0), P6=(4, 1), P7=(5, 1)

Step 4: Calculate New Centroids Again

New Centroids:

- **C1**: Mean of P1, P2, P3, P4 = (0.5, 0.5)
- **C2**: Mean of P5, P6, P7 = (4.33, 0.67)

Step 5: Check for Convergence

Since the centroids have not changed from the previous step, the algorithm terminates.

Final Clustering

- **Cluster 1 (C1)**: P1=(0, 0), P2=(0, 0.5), P3=(1, 0.5), P4=(1, 1)
- **Cluster 2 (C2)**: P5=(4, 0), P6=(4, 1), P7=(5, 1)

Final Centroids:

- **C1**: (0.5, 0.5)
- **C2**: (4.33, 0.67)

b

Initial Setup

- **Points**: P1=(-8.5, 0), P2=(-2, 0), P3=(-1, 0), P4=(0, 0), P5=(8, 0), P6=(9, 0)
- **Initial Centroids**: C1=(-8.5, 0) (P1), C2=(8, 0) (P5), C3=(9, 0) (P6)

Step 1: Assign Points to the Nearest Centroid

Clusters:

- **Cluster 1 (C1)**: P1=(-8.5, 0), P2=(-2, 0), P3=(-1, 0)
- **Cluster 2 (C2)**: P4=(0, 0), P5=(8, 0)
- **Cluster 3 (C3)**: P6=(9, 0)

Step 2: Calculate New Centroids

New Centroids:

- **C1**: Mean of P1, P2, P3 = (-3.83, 0)
- **C2**: Mean of P4, P5 = (4, 0)
- **C3**: Mean of P6 = (9, 0)

Step 3: Reassign Points to the Nearest Centroid

Clusters:

- **Cluster 1 (C1)**: P1=(-8.5, 0), P2=(-2, 0), P3=(-1, 0), P4=(0, 0)
- **Cluster 2 (C2)**: Empty
- **Cluster 3 (C3)**: P5=(8, 0), P6=(9, 0)

Step 4: Calculate New Centroids Again

New Centroids:

- **C1**: Mean of P1, P2, P3, P4 = (-3.83, 0)
- **C2**: Empty

- **C3**: Mean of P5, P6 = (8.5, 0)

Step 5: Check for Convergence

Since the centroids have not changed from the previous step, the algorithm terminates.

Final Clustering

- **Cluster 1 (C1)**: P1=(-8.5, 0), P2=(-2, 0), P3=(-1, 0), P4=(0, 0)
- **Cluster 2 (C2)**: Empty
- **Cluster 3 (C3)**: P5=(8, 0), P6=(9, 0)
- **C1**: (-3.83, 0)
- **C2**: Empty
- **C3**: (8.5, 0)

Conclusion

The above example shows that after classification, there may be cases where a cluster is empty.

Question3

1

- When using default parameters, the SSE = 1697.893, and MSE = 56.596.

2

- init: This parameter determines the method for initializing the centroids. Using k-means++ can select initial centroids that are far apart, avoiding the problem of slow convergence or local optima that may result from random initialization.
- n_init: This parameter specifies the number of times the algorithm will run with different initial centroids, and the final result will be the one with the smallest total squared error (SSE) among these runs.
    - Increasing this value within a certain range will significantly reduce the SSE because the more times it runs, the higher the chance of finding the optimal value.
    - Changing this parameter's value does not necessarily lead to better results, as the number of initial centroids is limited. Beyond a certain limit, the optimal solution may have already been found, and there will be no significant changes.

3

- Cluster 0: Retail and Services
    - Kraft, Verizon, Procter & Gamble, AT&T, Merck, McDonalds, Coca-Cola

- Cluster 1: Traditional industry
  - DuPont, Caterpillar, Alcoa
- Cluster 2: Technology
  - Microsoft, IBM, The Home Depot, Intel, Wal-Mart, General Electric, United Technologies, Travelers, 3M, Johnson & Johnson
- Cluster 3: Telecommunication
  - Cisco Systems
- Cluster 4: Energy and Chemicals
  - Chevron, Pfizer, ExxonMobil
- Cluster 5: Bank
  - Bank of America
- Cluster 6: Electronics product
  - Hewlett-Packard
- Cluster 7: Travel
  - American Express, Boeing, Walt Disney, JPMorgan Chase

Question4

1

BI-RADS = 4, Margin=1 -> Severity = 0

- support = 0.299688
- confidence = 0.911392

BI-RADS = 5, Shape = 4 -> Severity = 1

- support = 0.246618
- confidence = 0.908046

BI-RADS = 5, Shape = 4, Density = 3 -> Severity = 1

- support = 0.224766
- confidence = 0.915254

2

Shape=2, Margin=1 -> Severity=0

Margin_1, Density_3, Shape_1 -> Severity_0

Insights: When both margin and shape are relatively normal, the severity of the disease tends to be lower.

3

BI-RADS=4, Shape=2 -> Severity=0

From the above examples, it can be seen that although the BI-RADS rating is 4, which indicates a higher possibility of malignancy, but the lesion is ultimately benign.

4

support: 0.0125

confidence: 0.9231

I think that even though the confidence of this association is very high, the support is too low, indicating that the sample size is too small and cannot be considered as valid information.

5

Age=0(< mean age), Shape=1 -> Severity=0

- support=0.140479
- confidence=0.924658