A20474413
Jorge Gonzalez Lopez

Assignment 2

CS484 Spring 2021
February 15, 2021

Question 1

- **Answer**

  Lift = 1

- **Explanation**

  The itemset {A, B, C, D, E} has a Support value of 1. That means that all transactions contain the whole itemset.

  Now, splitting the itemset into X = {B, D} and Y = {A, C, E}, the association rules are as follows:

  - Confidence: 1 (as all the transactions contain all the elements, all transactions that contain itemset X also contains itemset Y).
  - Expected Confidence: 1 (all transactions contain all elements of Y)

  Finally, Lift can be described as the ratio of Confidence to Expected Confidence:

  - Lift: 1

  What was expected as the occurrence of both item sets does not have a determined dependency.

Question 2

- **Answer**

  Lift = 1.5

- **Explanation**

  The itemset are defined as X = {Cheese, Wings} and Y = {Soda} and, with the list of items defined in the Assignment, the association rules are as follows:

  - Support of X: 3/6 = 0.5 (3 out of the 6 transactions include Cheese and Wings)
  - Support of Y: 4/6 = 2/3 (4 out of the 6 transactions include Soda)
  - Confidence: $\Pr(Y|X) = \Pr(X \cup Y)/\Pr(X)$ = (3/6) / (3/6) = 1 (the transactions that include Cheese and Wings, also always include Soda)

  Finally, Lift can be described as the ratio of Confidence to Expected Confidence:

  - Lift: 1 / (2/3) = 1.5

  The occurrence of a friend bringing Cheese and Wing, also increases the occurrence of him bringing Soda.

Question 3

- **Answer**

  (C): Close to one

- **Explanation**

First of all, the Silhouette Index has a range of [-1, 1], so the option (D) is discarded. Then, as it can be seen in the Figure from the Assignment questions, all the observations within every cluster are close together but far away from other clusters. Therefore, it is a pretty ideal situation, which suggests a larger value of the Silhouette Index. Specifically, close to +1, which indicates a perfect clustering result.

Question 4

a)

- **Answer**

  0.6428571428571429

- **Explanation**

  $a_{ij} = \sum_{x_{ij}, x_{is} \in C_i, j \neq s} d(x_{ij}, x_{is}) / (n_{C_i} - 1)$ This is calculated between the value -1 and all other values in Cluster 0.

  $d_{ij,C_r} = \sum_{x_{ij} \in C_i, x_{rs} \in C_r} d(x_{ij}, x_{rs}) / n_{C_r}$ This is calculated between the value -1 and all other values in Cluster 1.

  As there are only 2 clusters: $b_{ij} = d_{ij}$, and the width is then computed as: $s_{ij} = \frac{b_{ij} - a_{ij}}{\max(a_{ij}, b_{ij})}$

b)

- **Answer**

  $R_0 = 0.5889$
  $R_1 = 0.5889$

- **Explanation**

  The centroids of both clusters are computed. Then, the intra-clusters distances are calculated as: $S_k = \frac{1}{n_k} \sum_{x_i \in C_k} d(x_i, c_k)$.

  Then the inter-cluster distance as: $M_{kl} = d(c_k, c_l)$

  And finally, the Davies-Bouldin value as: $R_k = \max_{\substack{1 \leq l \leq K \\ l \neq k}} R_{kl}$ where R_{kl} stands for $\frac{S_k + S_l}{M_{kl}}$.

c)

- **Answer**

  0.5889

- **Explanation**

  The Davis-Bouldin Index is calculated $DB = \frac{1}{K}\sum_{k=1}^{K} R_k$.

Question 5

a)

- **Answer**

  - 524 itemsets.
  - 4 is the largest k value among them.

- **Explanation**

The dataset is converted to a list of itemsets and then to a Item Indicator format. Finally, with the function apriori, the itemsets that appear at least in 75 % of the market baskets (support = 75 / total_number_of_itemsets) are calculated and displayed ordered. Therefore, the total number of itemsets is the length of this list and the largest k value, the length of the last itemset.

b)

- **Answer**

  1228

- **Explanation**

Applying the function association_rules to the list with the itemsets that appear at least in 75 % of the market baskets and with the argument min_threshold defined as 0.01 (the minimum coincidence), gives out a list of all the association rules. Its length is 1228.

c)

- **Image**

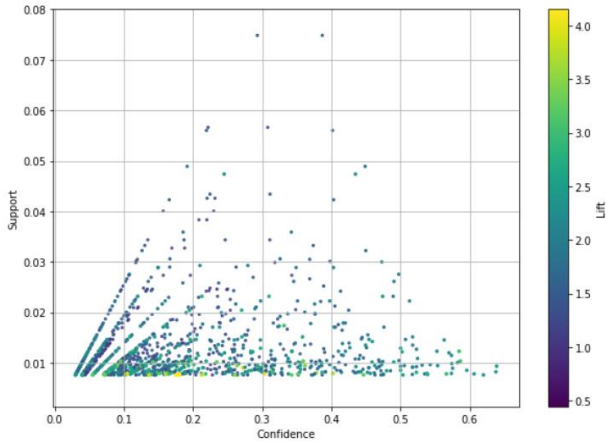

Figure 1: Scatter plot of Confidence vs Support

d)

- **Command Prompt Output**

	antecedents	consequents	support	confidence	lift
727	(root vegetables, butter)	(whole milk)	0.008236	0.637795	2.496107
734	(yogurt, butter)	(whole milk)	0.009354	0.638889	2.500387
1200	(root vegetables, yogurt, other vegetables)	(whole milk)	0.007829	0.606299	2.372842
1214	(yogurt, other vegetables, tropical fruit)	(whole milk)	0.007626	0.619835	2.425816

Figure 2: Rules with a confidence of at least 60 %.

Question 6

a)

- **Command Prompt Output**

K	WCSS	Elbow Value	Silhouette values	Calinski-Harabasz Scores	Davies-Bouldin Index
2	1360.8074	6.7306	0.4374	454.6607	0.8405
3	881.4518	7.5802	0.4040	465.6973	0.8275
4	687.5853	9.4943	0.3796	436.9141	0.8836
5	575.5305	9.3369	0.3218	411.1514	0.9710
6	487.9367	9.5308	0.3274	402.2028	0.9744
7	408.7366	9.5114	0.3459	412.7620	0.9410
8	369.0070	9.6593	0.3292	397.4171	0.9491
9	341.3922	10.2850	0.3308	379.2098	0.9603
10	317.3215	10.1014	0.3299	365.3021	0.9415

Figure 3: List of WCSS, Elbow value, Silhouette values, Calinski-Harabasz Scores and Davies-Bouldin Indexes for a series of numbers of clusters.

b)

- **Answer**

  3

- **Explanation**

Based on the values in answer a), the number of clusters that minimize the Davies-Bouldin Index and also maximizes the Calinski-Harabasz Scores is 3 clusters.

Looking at the Silhouette value, the best number of clusters is 2. However, 3 clusters is the second highest Silhouette value and is pretty similar to the highest value.

Finally, the Elbow value increases drastically after the third cluster and the WCSS decreases more slowly after the third cluster.

c)

- **Command Prompt Output**

```
Cluster Centroid 0: [3635.36206897  108.89655172  188.81896552]
Cluster Centroid 1: [4874.36764706  118.97058824  201.42647059]
Cluster Centroid 2: [2785.1796875  101.0625     173.90625  ]
```

Figure 4: Cluster Centroids in the original scales.