CS 484 - ASSIGNMENT 02

Sukanta Sharma

Spring 2021

CS 484: Introduction to Machine Learning

Spring 2021 Assignment 2

Question 1 (5 points)

Suppose the itemset {A, B, C, D, E} has a Support value of 1, then what is the Lift value of this association rule {B, D} ➔ {A, C, E}?

Ans:

The itemset $\{A, B, C, D, E\}$ has a Support value of 1 that means in all the transaction all five items are present. Now for the association rule $\{B, D\}$ ➔ $\{A, C, E\}$, it does not matter A, C, E is dependable of B, D or not, because in every transaction they all five items are present. Thus, the occurrence of itemset $\{A, C, E\}$ is not dependent on itemset $\{B, D\}$, So the Lift value of this association rule is 1.

Question 2 (5 points)

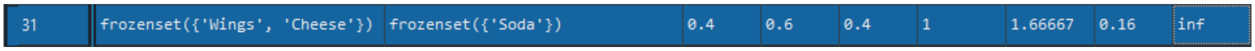
You invited your six friends to your home to watch a basketball game. Your friends brought snacks and beverages along. The following table lists the items your friends brought.

Friend	Items
Andrew	Cheese, Cracker, Soda, Wings
Betty	Cheese, Soda, Tortilla
Carl	Cheese, Ice Cream, Soda, Wings
Danny	Cheese, Ice Cream, Salsa, Tortilla
Emily	Salsa, Tortilla, Wings
Frank	Cheese, Cracker, Ice Cream, Soda, Wings

You noticed that many of your friends brought Cheese, Soda, and Wings together. Since you rather want to spend your money on food than Soda, you want to study how likely your friends will also bring Soda if they are going to bring Cheese and Wings. Therefore, please tell me the Lift of this association rule {Cheese, Wings} ==> {Soda}.

Ans:

We have created a csv data file named 'q2data.csv' using the above information and running the python code gives us the following information, where 8th column is Lift Value:

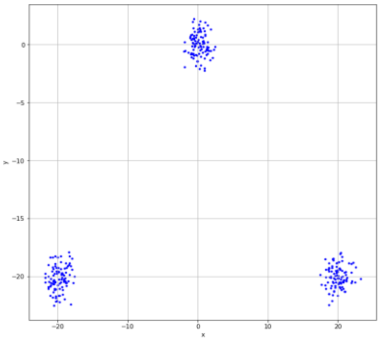


```
31    frozenset({'Wings', 'Cheese'})  frozenset({'Soda'})    0.4    0.6    0.4    1    1.66667    0.16    inf
```

The lift of this association rule $\{Cheese, Wings\}$ ➔ $\{Soda\}$ is 1.6666666666666667

Question 3 (5 points)

You are provided with the following scatterplot of two interval variables, namely, x and y. Without accessing the data, what do you think the Silhouette value will be for the 3-cluster K-mean solution? (A) Close to negative one, (B) About zero, (C) Close to one, (D) Close to three, or (E) Cannot be determined



Ans:

The Silhouette Index has a range of [-1, 1], where **+1 indicates a perfect clustering result**. Without accessing the data, for the above figure and 3-cluster K-mean solution it seemed to be almost perfect clustering, so I think the Silhouette value will be (C) Close to one.

Question 4 (15 points)

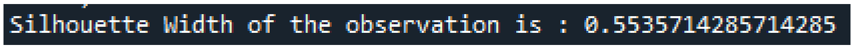
Suppose Cluster 0 contains observations {-2, -1, 1, 2, 3} and Cluster 1 contains observations {4, 5, 7, 8}.

a) (5 points) Calculate the Silhouette Width of the observation 2 (i.e., the value -1) in Cluster 0.

Ans:

x	Cluster	Distance (-1)	Sum	
-2	0	1	10	Divide Sum by (5 - 1) gives $a_{ij} = 2.5$
-1	0	0		
1	0	2		
2	0	3		
3	0	4		
4	1	5	28	Divide Sum by (5 - 1) gives $b_{ij} = 7$
5	1	6		
7	1	8		
8	1	9		

Silhouette Width of the observation, $S_{ij} = \frac{b_{ij}- a_{ij}}{\max(a_{ij},b_{ij})} = \frac{7-2.5}{7} = 0.6428571428571429$

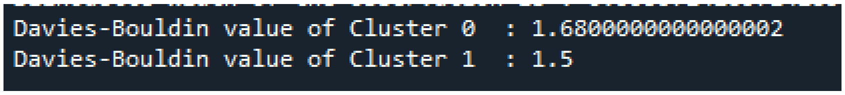


```
Silhouette Width of the observation is : 0.5535714285714285
```

b) (5 points) Calculate the cluster-wise Davies-Bouldin value of Cluster 0 (i.e., R_0) and Cluster 1 (i.e., R_1).

Ans:

x	Cluster	Mean	Distance = x - Mean	Count	Sum	Davies-Bouldin value = Sum/Count
-2	0	0.6	2.6	5	8.4	1.68
-1	0		1.6			
1	0		0.4			
2	0		1.4			
3	0		2.4			
4	1	6	2	4	6	1.5
5	1		1			
7	1		1			
8	1		2			



```
Davies-Bouldin value of Cluster 0  : 1.6800000000000002
Davies-Bouldin value of Cluster 1  : 1.5
```

c) (5 points) What is the Davies-Bouldin Index of this two-cluster solution?

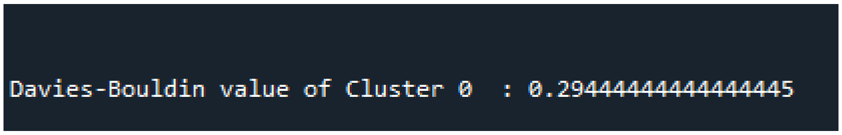
Ans:

From (b), we have means as, $c_k = 0.6\,, c_l = 6, S_k = 8.4,\ S_l = 6, K = 9$

Therefore, $M_{kl} = d(c_k,\ c_l) = 5.4$

Now, $R_{kl} = \frac{S_k + S_l}{M_{kl}} = 2.6667$ and $R_k = \max_{1\le l<k;l\neq k} R_{kl} = 2.6667\ (as\ there\ are\ only\ two\ clusters)$

So, Davies-Bouldin Index, $DB = \frac{1}{K}\sum_{k=1}^{K} R_k = \frac{2.6667}{9} = 0.295\ (approx)$



```
Davies-Bouldin value of Cluster 0  : 0.29444444444444445
```

Question 5 (30 points)

The file Groceries.csv contains market basket data. The variables are:

1. Customer: Customer Identifier
2. Item: Name of Product Purchased

After you have imported the CSV file, please discover association rules using this dataset. For your information, the observations have been sorted in ascending order by Customer and then by Item. Also, duplicated items for each customer have been removed.

a) (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy-five (75) customers. How many itemsets in total can we find? Also, what is the largest *k* value among our itemsets?

Ans:

From the python code we can find the following information:

- 524 item-sets we can find in total.
- The largest k value among our item-sets is 4

b) (5 points) Use the largest *k* value you found in (a), find out the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent. Please **do not** display those rules in your answer.

Ans:

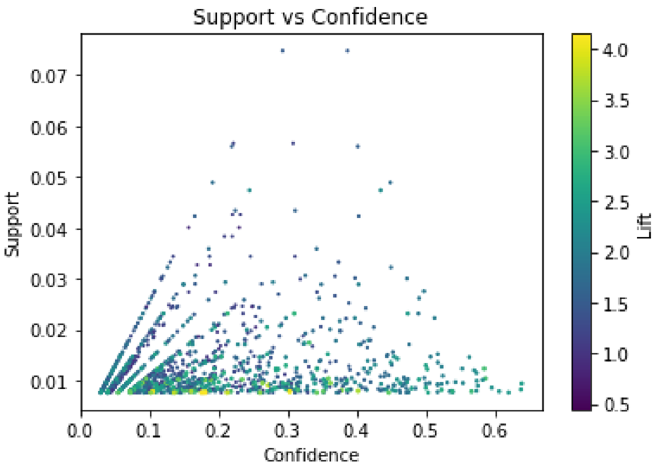
From the python code we can find the following information:

- We can found 1228 rules

c) (10 points) Plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (b). Please use the Lift metrics to indicate the size of the marker. You must add a color gradient legend to the chart for the Lift metrics.

Ans:

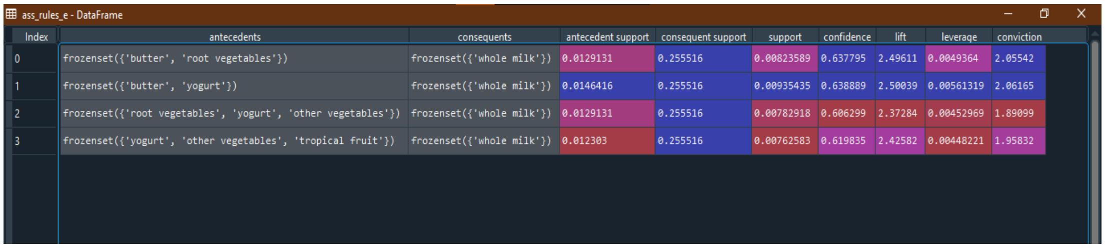
The plot is given below:



d) (5 points) Among the rules that you found in (b), list the rules whose Confidence metrics are greater than or equal to 60%. Please show the rules, including the Support, the Confidence, and the Lift metrics, in a table.

Ans:

From the python code we can get the following information:



ass_rules_e - DataFrame

Index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	frozenset({'butter', 'root vegetables'})	frozenset({'whole milk'})	0.0129131	0.255516	0.00823589	0.637795	2.49611	0.0049364	2.05542
1	frozenset({'butter', 'yogurt'})	frozenset({'whole milk'})	0.0146416	0.255516	0.00935435	0.638889	2.50039	0.00561319	2.06165
2	frozenset({'root vegetables', 'yogurt', 'other vegetables'})	frozenset({'whole milk'})	0.0129131	0.255516	0.00782918	0.606299	2.37284	0.00452969	1.89099
3	frozenset({'yogurt', 'other vegetables', 'tropical fruit'})	frozenset({'whole milk'})	0.012303	0.255516	0.00762583	0.619835	2.42582	0.00448221	1.95832

Question 6 (40 points)

You are asked to discover the optimal clusters in the cars.csv. Here are the specifications.

- The input interval variables are Weight, Wheelbase, and Length
- Scale each input interval variable such that the resulting variable has a range of 0 to 10
- The distance metric is Manhattan
- The minimum number of clusters is 2
- The maximum number of clusters is 10
- Specify random_state = 60616 in calling the KMeans function in scikit-learn library

Please answer the following questions.

a) (20 points) List the Elbow values, the Silhouette values, the Calinski-Harabasz Scores, and the Davies-Bouldin Indices for your 2-cluster to 10-cluster solutions.

Ans:

From the python code we can get the following information:

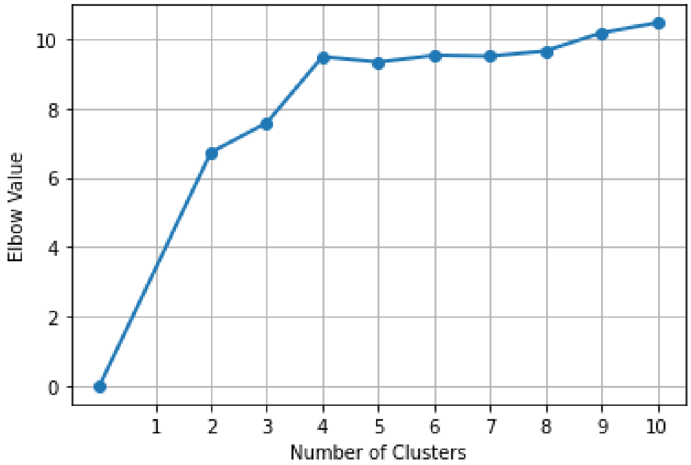
N Clusters	Elbow values	Silhouette values	Calinski-Harabasz Scores	Davies-Bouldin Indices
2	6.7306	0.4374	454.6607	0.8405
3	7.5829	0.4038	465.7613	0.8294
4	9.4929	0.3801	437.0228	0.8853
5	9.3369	0.3218	411.1514	0.971
6	9.5309	0.3274	402.2028	0.9744
7	9.5052	0.3453	413.6565	0.9377
8	9.6593	0.3292	397.4171	0.9491

9	10.18	0.3315	378.9413	0.9386
10	10.4693	0.3224	365.9016	0.9835

b) (10 points) Based on the values in (a), what is your suggested number of clusters?

Ans:

From the python code we can get the following information:



N Clusters	Elbow values	Slope	Acceleration
2	6.7306		
3	7.5829	-0.8523	
4	9.4929	-1.91	-1.0577
5	9.3369	0.156	2.066
6	9.5309	-0.194	-0.35
7	9.5052	0.0257	0.2197
8	9.6593	-0.1541	-0.1798
9	10.18	-0.5207	-0.3666
10	10.4693	-0.2893	0.2314

As we have got the highest value of acceleration for 4 Clusters, I will suggest 4 clusters.

c) (10 points) What are the cluster centroids of your suggested cluster solution? Please show the centroids in their original scales.

Ans:

The cluster centroids of the suggested cluster solution as 4 are given below in their original scales (values we can get from the python code).

#Cluster	Centroid List
0	[3378.13978495, 106.24731183, 184.53225806]
1	[2637.79545455, 100.27272727, 171.93181818]
2	[4073.8974359, 113.09401709, 194.8034188]
3	[5250.21621622, 120.86486486, 203.18918919]