CS 484: Introduction to Machine Learning

Autumn 2021 Assignment 2 Answer Key

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}?

It is known that adding more items to an itemset may lower the Support, therefore, Support of {A, C, E} ▯ Support of {A, B, C, D, E} and Support of {B, D} ▯ Support of {A, B, C, D, E}. Since the question says that Support of {A, B, C, D, E} = 1, we have Support of {A, C, E} ▯ 1 and Support of {B, D} ▯ 1. As Support cannot exceed 1, therefore, Support of {A, C, E} = 1 and Support of {B, D} = 1.

The Confidence value of the rule {B, D} ➔ {A, C, E} is Support of {A, B, C, D, E} / Support {B, D} = 1. The Expected Confidence value of the rule {B, D} ➔ {A, C, E} is Support of {A, C, E} = 1. Finally, the Lift is Confidence / Expected Confidence = 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, Wings

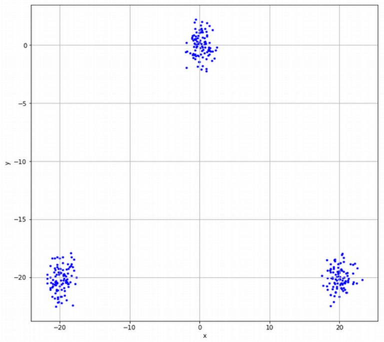
You noticed that many of your friends brought Cheese, Soda, and Wings together. Since you rather want to spend your money on food than Soda, you want to study how likely your friends will also bring Soda if they are going to bring Cheese and Wings. Therefore, please tell me the Lift of this association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

Friend	Items	{Cheese, Wings}	{Soda}	{Cheese, Wings, Soda}
Andrew	Cheese, Cracker, Soda, Wings	Yes	Yes	Yes
Betty	Cheese, Soda, Tortilla	No	Yes	No
Carl	Cheese, Ice Cream, Soda, Wings	Yes	Yes	Yes
Danny	Cheese, Ice Cream, Salsa, Tortilla	No	No	No
Emily	Salsa, Tortilla, Wings	No	No	No
Frank	Cheese, Cracker, Ice Cream, Wings	Yes	No	No

Support of {Cheese, Wings} = 3/6. Support of {Soda} = 3/6. Support of {Cheese, Wings, Soda} = 2/6. Confidence of {Cheese, Wings} ➔ {Soda} = Support of {Cheese, Wings, Soda} / Support of {Cheese, Wings} = (2/6) / (3/6) = 2/3. Expected Confidence of the rule is Support of {Soda} = 3/6. Therefore, the Lift of the rule is (2/3) / (3/6) = 4/3 = 1.3333.

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



The Silhouette value should be (C) Close to one. The scatterplot clearly shows three non-overlapping and compact clusters of observations. Moreover, the distances between the clusters are much larger than the distances within the clusters. Thus, the silhouette value for the 3-cluster solution should be close to one. Indeed, the silhouette value with the Euclidean distance is 0.9368.

Question 4 (15 points)

We have nine observations on one interval variable. Suppose we discovered two clusters using the Euclidean distance. Cluster 0 contains five observations {-2, -1, 1, 2, 3} and Cluster 1 contains four observations {4, 5, 7, 8}.

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

Cluster	Value	Distance from -1		Silhouette Width
0	-2	1	$a_{i;k}=\frac{1+2+3+4}{4}=2.5$	$s=\frac{(7-2.5)}{max(2.5,7)}$ ¿0.6429
	-1	N/A		
	1	2		
	2	3		
	3	4		
1	4	5	$d_{i;C_l}=\frac{5+6+8+9}{4}=7$ $b_{i;k}=7$	
	5	6		
	7	8		
	8	9		

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

Cluster	Value	Size n_k	Centroid	Intra-Distance	S_k	M_k
0	-2	5	0.6	2.6	1.68	5.4
	-1			1.6		
	1			0.4		
	2			1.4		
	3			2.4		
1	4	4	6	2	1.5	
	5			1		
	7			1		
	8			2		

Therefore, $R_{01}=\frac{S_0+S_1}{M_{01}}=\frac{1.68+1.5}{5.4}=0.5888889$. Similarly,

$$R_{10}=\frac{S_1+S_0}{M_{10}}=\frac{1.68+1.5}{5.4}=0.5888889$$.

Davies-Bouldin value of Cluster 0, $R_0=0.5889$ and of Cluster 1, $R_1=0.5889$.

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

The Davies-Bouldin Index of this two-cluster solution is $R=\left(R_0+R_1\right)/2=0.5889$.

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.

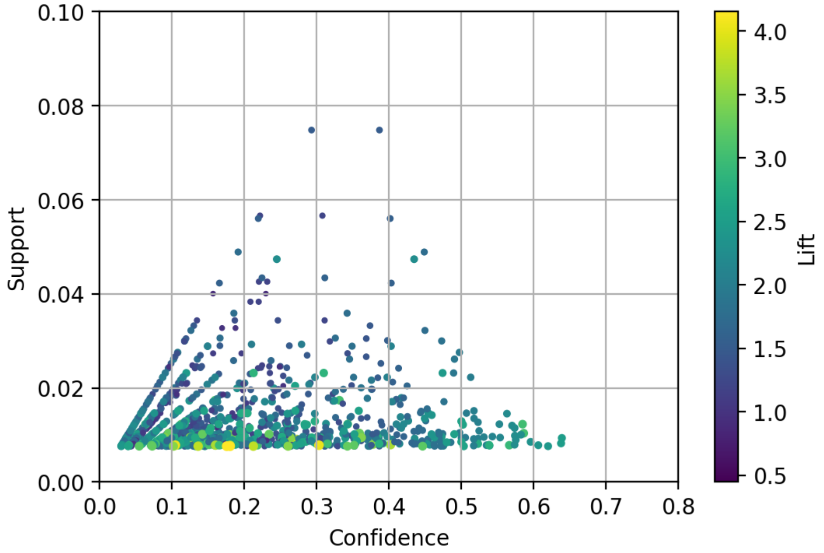
a) (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy-five (75) customers. How many itemsets in total can we find? Also, what is the largest *k* value among our itemsets?

Since there are 9,835 customers in the data, we should specify the minimum support as MIN_SUPPORT = 75/9835 = 0.0076258261311642. Since the maximum number of items ever purchased by a customer is 32, we will specify the maximum length of an itemset as MAX_LEN = 32 accordingly. Based on our specifications, we found 524 itemsets and the highest *k* value 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.

We specified MIN_THRESHOLD = 0.01 and found 1,228 association 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.



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 in a table that shows the Antecedent, the Consequent, the Support, the Confidence, the Expected Confidence, and the Lift.

Antecedent	Consequent	Support	Confidence	Expected Confidence	Lift
{'butter', 'root vegetables'}	{'whole milk'}	0.00824	0.63780	0.25552	2.49611
{'yogurt', 'butter'}	{'whole milk'}	0.00935	0.63889	0.25552	2.50039
{'yogurt', 'root vegetables', 'other vegetables'}	{'whole milk'}	0.00783	0.60630	0.25552	2.37284
{'yogurt', 'tropical fruit', 'other vegetables'}	{'whole milk'}	0.00763	0.61983	0.25552	2.42582

Question 6 (40 points)

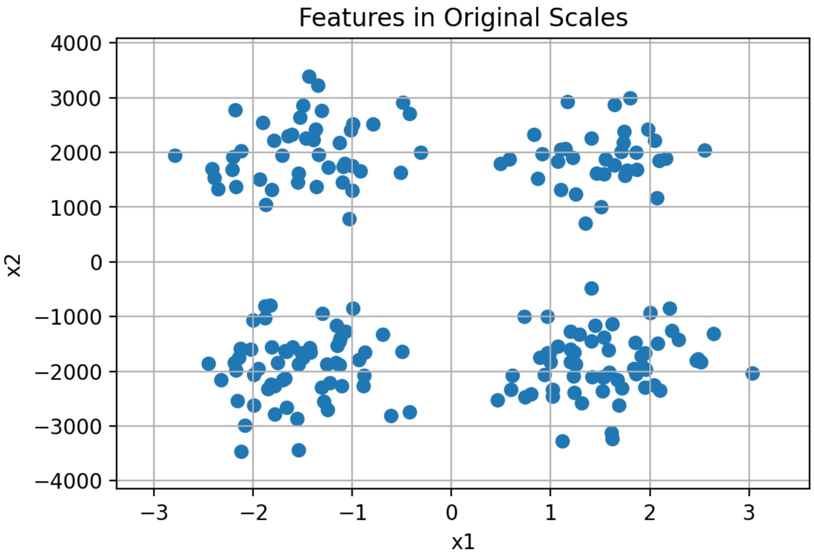
You are asked to discover the optimal clusters in the TwoFeatures.csv. This data has 200 observations and two interval variables, namely, x1 and x2. Here are the analysis specifications.

- The input interval variables are x1 and x2
- The metric is the Manhattan distance
- The minimum number of clusters is 1
- The maximum number of clusters is 8
- Use the Elbow value for choosing the optimal number of clusters

Since the sklearn.cluster.KMeans class assumes only the Euclidean distance, you will need to implement the K-Means algorithm with the Manhattan distance in Python. However, you may consider calling the sklearn. metrics.pairwise.manhattan_distances function for calculating the Manhattan distance.

Please answer the following questions.

a) (5 points) Plot x2 (vertical axis) versus x1 (horizontal axis). Add gridlines to both axes. Based on this scatterplot, how many clusters do you think are there?

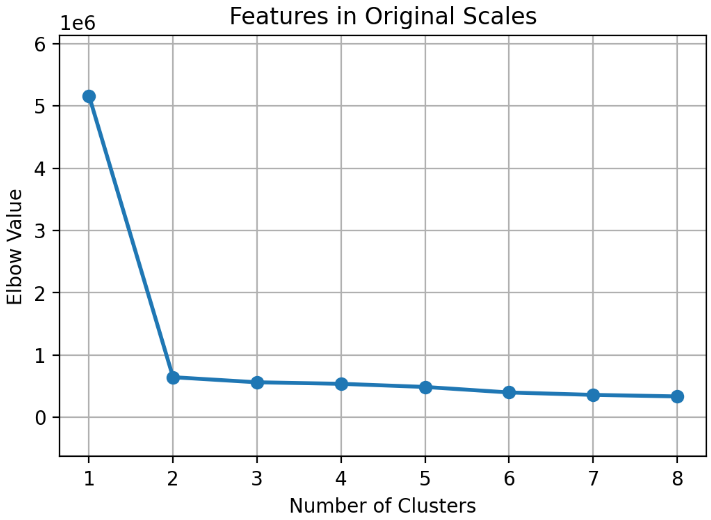


The above scatterplot clearly shows four clusters of points.

b) (10 points) Calculate the Total Within-Cluster Sum of Squares (TWCSS) values and the Elbow values for solutions for the number of clusters from 1 to 8. Please remember to calculate the TWCSS values using the Manhattan distance. Please list your results in a table.

Number of Clusters	Total WCSS	Elbow Value
1	1,032,560,056.57	5,162,800.28
2	65,089,654.39	642,801.09
3	39,338,640.78	561,286.75
4	32,452,819.79	536,759.32
5	28,276,984.15	486,275.71
6	12,845,481.32	397,771.48
7	8,880,156.13	358,928.70
8	7,487,198.05	334,308.89

c) (5 points) Plot the Elbow Values from part (b) versus the number of clusters. Based on the Elbow chart, what is your choice for the optimal number of clusters?

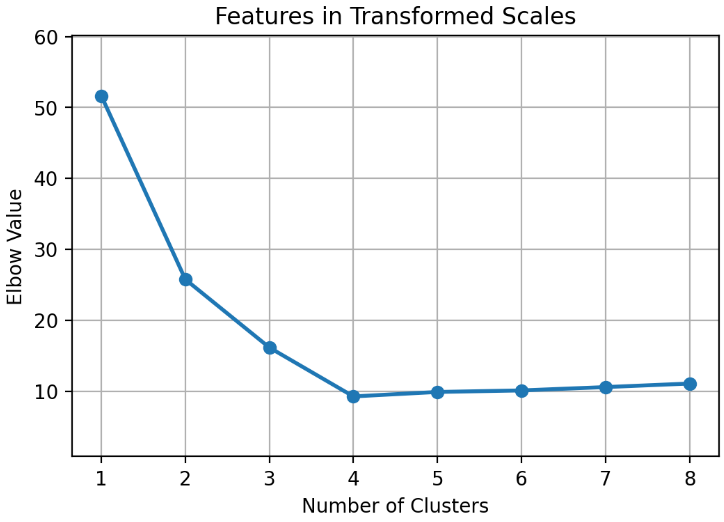


The elbow appears when the number of clusters is two. Therefore, the optimal number of clusters is two for the data in the original scale. Since x2 values are in thousands but x1 values are in units, the distance is dominated by x2. If the tick marks on both axes in the scatterplot in Part (a) are thousands, then the gaps along the horizontal axis will disappear.

d) (10 points) Linearly rescale x1 such that the resulting variable has a minimum of zero and a maximum of ten. Similarly, rescale x2. Using the rescaled observations, calculate the Total Within-Cluster Sum of Squares (TWCSS) values and the Elbow values for solutions for the number of clusters from 1 to 8. Please remember to calculate the TWCSS values using the Manhattan distance. Please list your results in a table.

Number of Clusters	Total WCSS	Elbow Value
1	10,326.63	51.63
2	2,584.09	25.75
3	1,149.05	16.15
4	472.77	9.25
5	416.47	9.88
6	388.68	10.10
7	340.08	10.57
8	312.85	11.07

e) (5 points) Plot the Elbow Values from part (d) versus the number of clusters. Based on the Elbow chart, what is your choice for the optimal number of clusters?



The elbow appears when the number of clusters is four. Therefore, the optimal number of clusters is four for the data in the transformed scale.

f) (5 points) Between your choices in Part (c) and Part (e), what will you recommend for the number of clusters? Please state your reasons.

Since the scatterplot in Part (a) clearly shows four clusters of observations, we will go with the choice in Part (e). This exercise shows that when the features are on very different scales, rescaling the features may help us choose the more sensible number of clusters.