Tania Soutonglang

CS 484: Introduction to Machine Learning

Autumn 2022 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}?

$$\frac{\text{Support}\{B, D\}}{\text{Support}\{A, B, C, D, E\}} = 1$$

$$\frac{\text{Support}\{A, C, E\}}{\text{Support}\{A, B, C, D, E\}} = 1$$

$$\text{Confidence}\{B, D\} \Rightarrow \{A, C, E\} = \frac{\text{Support}\{A, B, C, D, E\}}{\text{Support}\{B, D\}} = 1$$

$$Lift\{B, D\} \Rightarrow \{A, C, E\} = \frac{\text{Confidence}\{B, D\} \Rightarrow \{A, C, E\}}{\text{Support}\{A, C, E\}} = \mathbf{1}$$

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}.

$$\text{Lift}\{\text{Cheese, Wings}\} \Rightarrow \{\text{Soda}\} = \frac{\text{Confidence}\{\text{Cheese, Wings}\} \Rightarrow \{\text{Soda}\}}{\text{Support}\{\text{Soda}\}}$$

$$\text{Confidence}\{\text{Cheese, Wings}\} \Rightarrow \{\text{Soda}\} = \frac{\text{Support}\{\text{Cheese, Wings, Soda}\}}{\text{Support}\{\text{Cheese, Wings}\}}$$

$$\text{Support}\{\text{Cheese, Wings, Soda}\} = \frac{2 \text{ friends}}{6 \text{ friends}} = \frac{1}{3}$$

$$\text{Support}\{\text{Cheese, Wings}\} = \frac{3 \text{ friends}}{6 \text{ friends}} = \frac{1}{2}$$

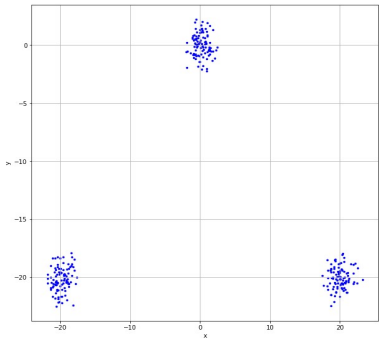
$$\text{Confidence}\{\text{Cheese, Wings}\} \Rightarrow \{\text{Soda}\} = \frac{1}{3} / \frac{1}{2} = \frac{2}{3}$$

$$\text{Support}\{\text{Soda}\} = \frac{3 \text{ friends}}{6 \text{ friends}} = \frac{1}{2}$$

$$\frac{\text{Confidence}\{\text{Cheese, Wings}\} \Rightarrow \{\text{Soda}\}}{\text{Support}\{\text{Soda}\}} = \frac{\frac{2}{3}}{\frac{1}{2}} = \mathbf{\frac{4}{3}}$$

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



C) Close to one

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, 0, 1, 2} and Cluster 1 contains four observations {6, 7, 8, 9}.

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

Cluster	Value	Distance from -1	
0	-2	1	$a_{ij} = \frac{1+1+2+3}{5-1} = \frac{7}{4} = 1.75$
	-1	-	
	0	1	

	1	2	
	2	3	
1	6	7	$d_{ij,c_r} = \frac{7+8+9+10}{4} = \frac{34}{4}$ $= \frac{17}{2} = 8.5 = b_{ij}$
	7	8	
	8	9	
	9	10	

$$s_{ij} = \frac{8.5 - 1.75}{\max(1.75, 8.5)} = \mathbf{0.79}$$

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	Centroid	Distance From Centroid	Intra-Cluster Distance	Inter-Cluster Distance
0	-2	5	0	2	$\frac{1}{5}(2+1+1+2)$ $= 1.2$	0.2
	-1			1		
	0			0		
	1			1		
	2			2		
1	6	4	7.5	1.5	$\frac{1}{4}(1.5+0.5+0.5$ $+1.5)$ $= 1$	
	7			0.5		
	8			0.5		
	9			1.5		

$$R_0 = R_1 = \frac{1 + 1.2}{0.2} = \mathbf{11}$$

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

The Davies-Bouldin Index is 11.

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?

We can find 524 itemsets. The largest *k* value among our itemsets is 4.

```
5.a
       support              itemsets
0     0.008033                   (0)
1     0.033452                   (1)
2     0.017692                   (7)
3     0.052466                   (9)
4     0.033249                  (10)
..         ...                   ...
519   0.007931       (162, 166, 158)
520   0.015150       (166, 158, 167)
521   0.010880       (162, 166, 167)
522   0.007829  (167, 124, 166, 103)
523   0.007626  (167, 166, 158, 103)
```

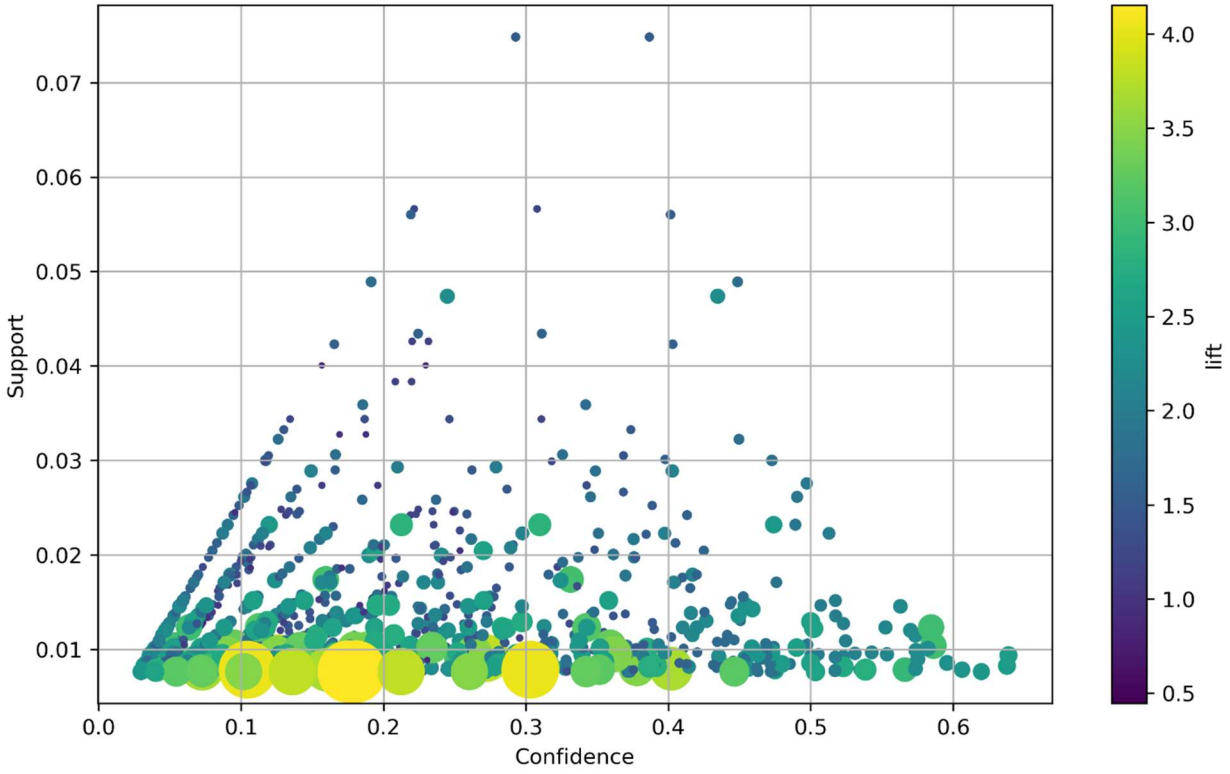
(5 points) Use the largest *k* value you found in (a) and find out the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent. Please **do not** display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

We can find 1,227 association rules.

5.b

	antecedents	consequents	...	leverage	conviction
0	(1)	(103)	...	0.001662	1.065626
1	(103)	(1)	...	0.001662	1.008964
2	(1)	(139)	...	0.001793	1.069409
3	(139)	(1)	...	0.001793	1.010750
4	(166)	(7)	...	0.004732	1.019216
...	...	...	...	...	...
1223	(158, 103)	(166, 167)	...	0.005615	1.198645
1224	(167)	(158, 166, 103)	...	0.005243	1.039756
1225	(166)	(103, 158, 167)	...	0.004482	1.018081
1226	(158)	(103, 166, 167)	...	0.005289	1.054357
1227	(103)	(158, 166, 167)	...	0.004694	1.025257

b) (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.



c) (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.

Antecedents	Consequents	Support	Confidence	Expected Confidence	Lift
(16, 124)	(166)	0.008236	0.637795	0.255516	2.496107
(16, 167)	(166)	0.009354	0.638889	0.255516	2.500387
(124, 103, 167)	(166)	0.007829	0.606299	0.255516	2.372842
(103, 158, 167)	(166)	0.007626	0.619835	0.255516	2.425816

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.

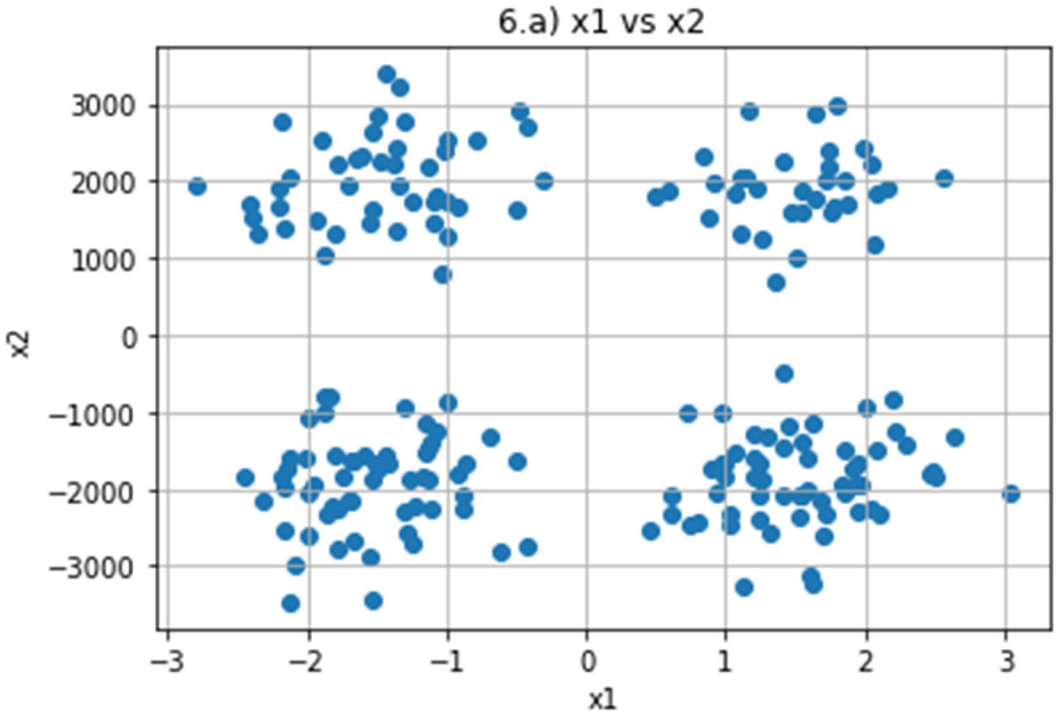
- The input interval variables are x1 and x2
- The metric is the Manhattan distance
- The minimum number of clusters is 1
- The maximum number of clusters is 8
- Use the Elbow value for choosing the optimal number of clusters

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?

I believe there are four clusters.

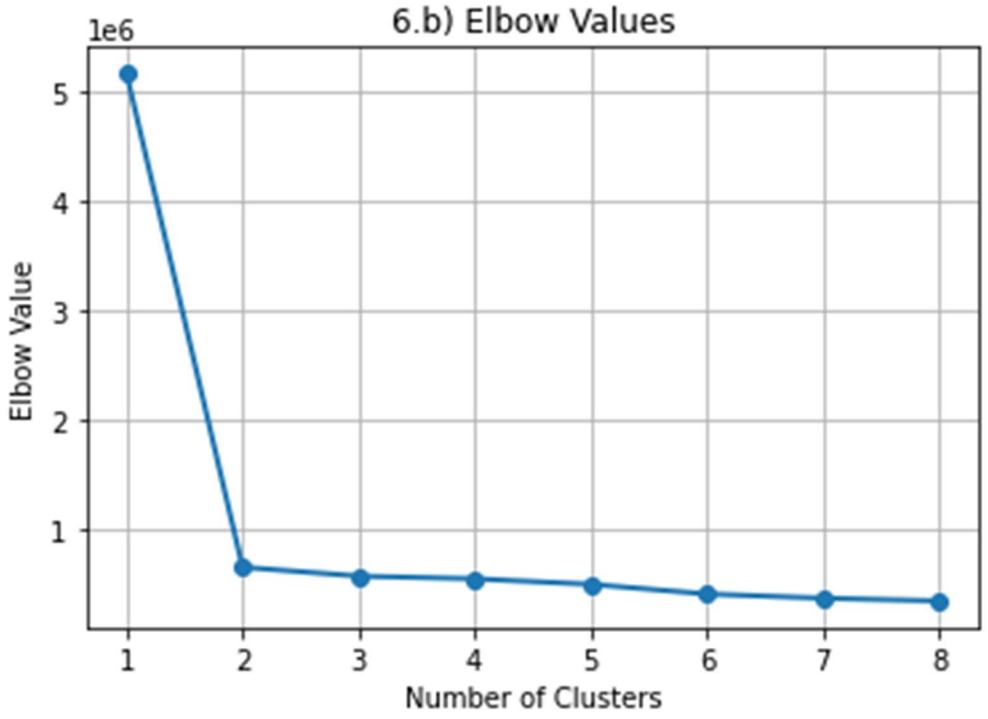


b) (10 points) Calculate the Total Within-Cluster Sum of Squares (TWCSS) values and the Elbow values for solutions for the number of clusters from 1 to 8. Please remember to calculate the TWCSS values using the Manhattan distance. Please list your results in a table.

Number of Clusters	TWCSS	Elbow Value
1	1,032,560,056.5653	5,162,800.2828
2	65,089,654.3897	642,801.0915
3	39,338,640.7771	561,286.7520
4	32,452,819.7934	536,759.3237
5	28,276,984.1469	486,275.7067
6	12,845,481.3165	397,771.4816
7	8,880,156.1293	358,928.7026
8	7,487,198.0489	334,308.8856

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?

From the graph, 2 is the most optimal number of clusters.



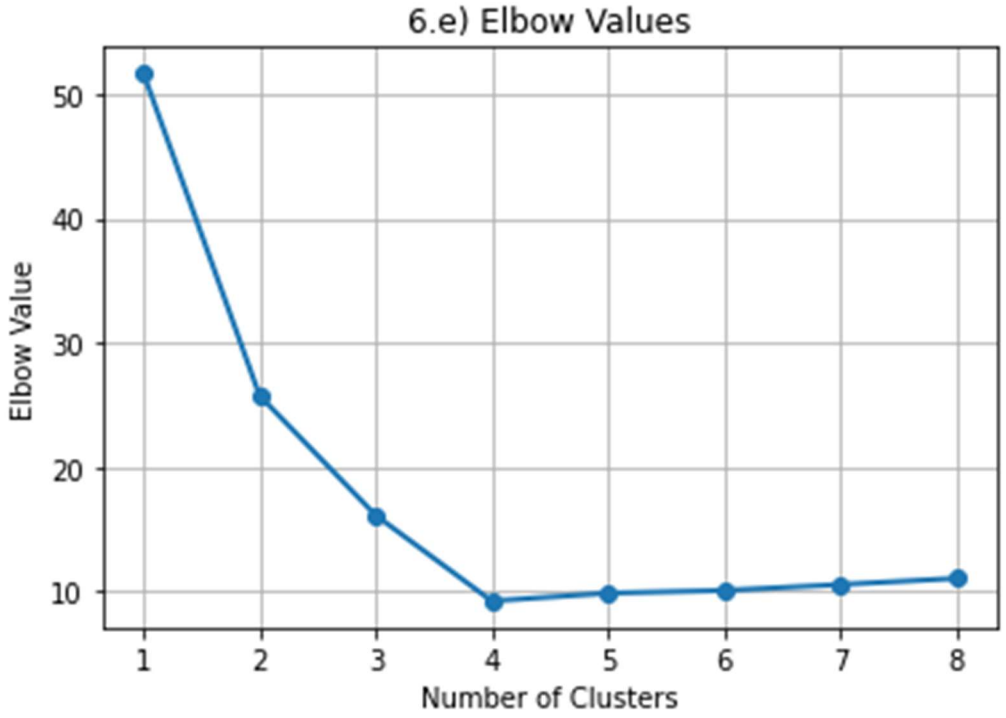
d) (10 points) Linearly rescale x1 such that the resulting variable has a minimum of zero and a maximum of ten. Similarly, rescale x2. Using the rescaled observations, calculate the Total Within-

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	TWCSS	Elbow Value
1	10,326.6276	51.6331
2	2,584.0854	25.7523
3	1,149.0505	16.1460
4	472.7691	9.2547
5	416.4666	9.8822
6	388.6763	10.0993
7	340.0794	10.5703
8	312.8536	11.0693

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?

In this graph, the elbow appears when the number of clusters is 4.



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.

I would go with 4 clusters since the graph in part (a) also had 4 clusters just like the graph in part (e). Rescaling the numbers showed a closer look at the data that might have been missed when the data was its original size.