CS 484: Introduction to Machine Learning

Autumn 2022 Assignment 2

**Michal Malek**

# 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} 🡺 (when given){A, C, E}?

**{B, D} = X and {A, C, E} =Y and we know since support value for the entire itemset is 1 then the support value for any given subset is 1**

**Lift Value**

# 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.

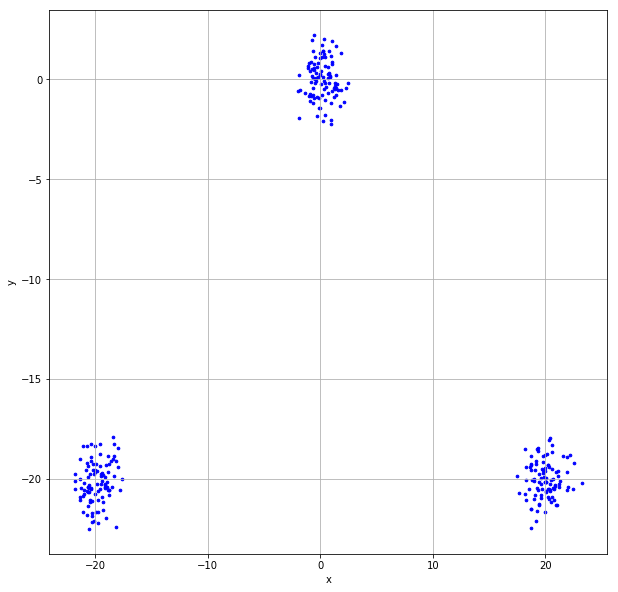
|  |  |
| --- | --- |
| Friend | Items |
| Andrew | Cheese, Cracker, Soda, Wings |
| Betty | Cheese, Soda, Tortilla |
| Carl | Cheese, Ice Cream, Soda, Wings |
| Danny | Cheese, Ice Cream, Salsa, Tortilla |
| Emily | Salsa, Tortilla, Wings |
| Frank | Cheese, Cracker, Ice Cream, Wings |

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}.

**Lift Value**

# Question 3 (5 points)

You are provided with the following scatterplot of two interval variables, namely, and . 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



**1**

**The silhouette value would be closer to 1 since the clusters are similar in size, and since answer D physically can’t make sense as a silhouette value is only between -1 and 1. And E is illogical as we have a scatter plot with 3 different clusters since only one cluster would give us undefined.**

# 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}.

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

**0.7941176470588235**

1. (5 points) Calculate the cluster-wise Davies-Bouldin value of Cluster 0 (i.e., ) and Cluster 1 (i.e., ).
2. (5 points) What is the Davies-Bouldin Index of this two-cluster solution?

**0.293333**

# 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.

1. (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 in total**

**The largest k value among our itemsets in a maximum of 4**

1. (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 1228 association rules**

1. (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.

Chart, scatter chart

Description automatically generated

1. (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** | **antecedent support** | **consequent support** | **support** | **confidence** | **lift** | **leverage** | **conviction** |
| **0** | frozenset({'root vegetables', 'butter'}) | frozenset({'whole milk'}) | 0.012913065 | 0.25551601423 | 0.008235892221 | 0.6377952755 | 2.4961068585 | 0.004936397172 | 2.055423178 |
| **1** | frozenset({'yogurt', 'butter'}) | frozenset({'whole milk'}) | 0.014641586 | 0.25551601423 | 0.009354346720 | 0.6388888888 | 2.5003868771 | 0.005613186980 | 2.061647960 |
| **2** | frozenset({'other vegetables', 'yogurt', 'root vegetables'}) | frozenset({'whole milk'}) | 0.012913065 | 0.25551601423 | 0.007829181494 | 0.60629921259 | 2.3728423222 | 0.004529686445 | 1.8909893238 |
| **3** | frozenset({'yogurt', 'other vegetables', 'tropical fruit'}) | frozenset({'whole milk'}) | 0.012302999 | 0.25551601423 | 0.007625826131 | 0.6198347107 | 2.4258155114 | 0.004482212737 | 1.9583165712 |

# 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.

1. (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?

**There looks to be four clusters**

**Chart, scatter chart

Description automatically generated**

1. (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.

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **WCSS** | **Elbow Value** |
| **1** | 7.82891e+08 | 3.91446e+06 |
| **2** | 6.50897e+07 | 642801 |
| **3** | 3.93365e+07 | 561191 |
| **4** | 2.3905e+07 | 472686 |
| **5** | 1.67531e+07 | 437485 |
| **6** | 1.27943e+07 | 401529 |
| **7** | 8.87993e+06 | 357822 |

1. (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?

**After graphing, I would say the optimal number of clusters is 2.**

Chart, line chart

Description automatically generated

1. (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.

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **WCSS** | **Elbow Value** |
| **1** | 5929.37 | 29.6469 |
| **2** | 2294.06 | 22.7438 |
| **3** | 1155.01 | 15.7564 |
| **4** | 472.769 | 9.25467 |
| **5** | 412.388 | 9.72723 |
| **6** | 359.861 | 10.4177 |
| **7** | 312.959 | 10.8884 |

1. (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?

**According to the chart, 4 would be optimal number of clusters**

**Chart, line chart

Description automatically generated**

1. (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 recommend 4 clusters because the Elbow Value from the scaled datasets correlate better with both graphs that look like there are 4 clusters from the dataset.**