CS 484: Introduction to Machine Learning

Spring 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, 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}.

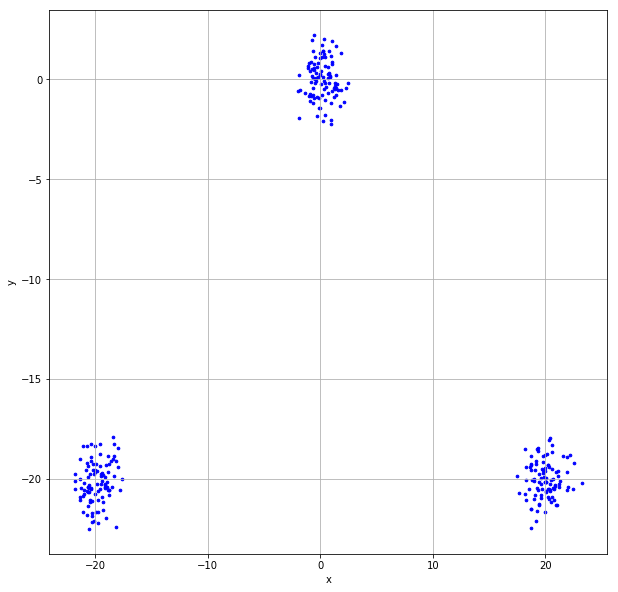
| Friend | Items | {Cheese, Wings} | {Soda} | {Cheese, Wings, Soda} |
| --- | --- | --- | --- | --- |
| Andrew | Cheese, Cracker, Soda, Wings | Yes | Yes | Yes |
| Betty | Cheese, Soda, Tortilla | No | Yes | No |
| Carl | Cheese, Ice Cream, Soda, Wings | Yes | Yes | Yes |
| Danny | Cheese, Ice Cream,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

Support of {Cheese, Wings} = 3 / 6. Support of {Soda} = 4 / 6. Support of {Cheese, Wings, Soda} = 3 / 6.

Confidence of {Cheese, Wings} 🡺 {Soda} = Support of {Cheese, Wings, Soda} / Support of {Cheese, Wings} = (3/6) / (3/6) = 1.0. Expected Confidence of the rule is Support of {Soda} = 4/6. Therefore, the Lift of the rule is 1.0 / (4/6) = 1.5.

# 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



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 is 0.9368.

# Question 4 (15 points)

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

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

| Cluster | Value | Distance from -1 |  | Silhouette Width |
| --- | --- | --- | --- | --- |
| 0 | -2 | 1 |  |  |
| -1 | N/A |
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 1 | 4 | 5 |  |
| 5 | 6 |
| 7 | 8 |
| 8 | 9 |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | Value | Size | Centroid | Intra-Distance |  |  |
| 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, . Similarly, .

Davies-Bouldin value of Cluster 0, and of Cluster 1, .

1. (5 points) What is the Davies-Bouldin Index of this two-cluster solution?

The Davies-Bouldin Index of this two-cluster solution is .

# 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.

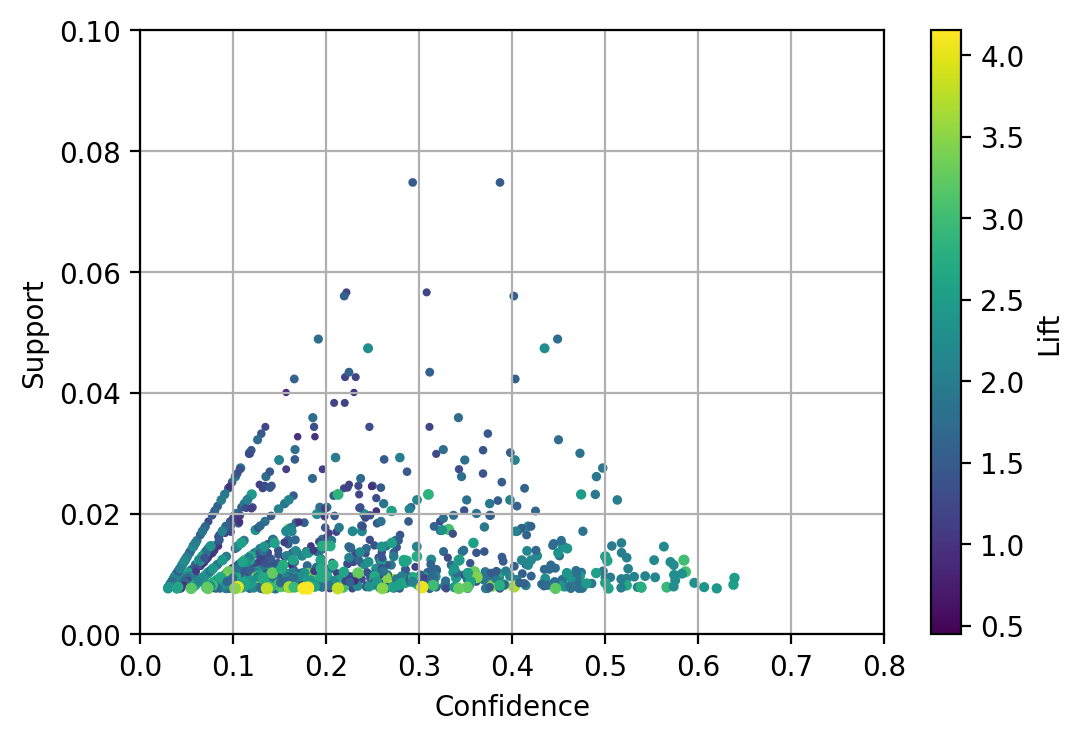
1. (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy five (75) customers. How many itemsets in total can we find? Also, what is the largest *k* value among our itemsets?

Since there are 9,835 customers in the data, we should specify the minimum support as MIN\_SUPPORT = 75/9835. 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.**

1. (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.

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.



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, including the Support, the Confidence, and the Lift metrics, in a table.

Expected Confidence is shown in the consequent support column.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **antecedents** | **consequents** | **antecedent support** | **consequent support** | **support** | **confidence** | **lift** |
| {'butter',  'root vegetables'} | {'whole milk'} | 0.01291 | 0.25552 | 0.00824 | 0.63780 | 2.49611 |
| {'yogurt', 'butter'} | {'whole milk'} | 0.01464 | 0.25552 | 0.00935 | 0.63889 | 2.50039 |
| {'yogurt',  'root vegetables',  'other vegetables'} | {'whole milk'} | 0.01291 | 0.25552 | 0.00783 | 0.60630 | 2.37284 |
| {'yogurt',  'tropical fruit',  'other vegetables'} | {'whole milk'} | 0.01230 | 0.25552 | 0.00763 | 0.61983 | 2.42582 |

# 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.

1. (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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| N Clusters | Total WCSS | Elbow Value | Silhouette Value | Calinkski-Harabasz Score | Davies Bouldin Index |
| 1 | 2813.1682 | 6.5728 |  |  |  |
| 2 | 1361.2005 | 6.7253 | 0.4367 | 454.4064 | 0.8410 |
| 3 | 894.3790 | 7.5695 | 0.3996 | 455.8948 | 0.8279 |
| 4 | 698.3085 | 9.5324 | 0.3608 | 428.0346 | 0.9046 |
| 5 | 629.4303 | 8.8788 | 0.3209 | 366.8878 | 0.9766 |
| 6 | 509.3603 | 9.8247 | 0.3273 | 381.7365 | 0.9490 |
| 7 | 477.0522 | 10.6278 | 0.3073 | 343.6049 | 1.1907 |
| 8 | 396.3455 | 10.2723 | 0.3218 | 365.8661 | 1.0390 |
| 9 | 363.2817 | 10.4267 | 0.3103 | 353.2047 | 1.0485 |
| 10 | 334.3628 | 10.5184 | 0.3075 | 344.3168 | 1.0518 |

1. (10 points) Based on the values in (a), what is your suggested number of clusters?

The Elbow values decelerate when the number of clusters is 4. The Silhouette values attain the highest value when the number of clusters is 2. When the number of clusters is 3, the Calinkski-Harabasz scores attain the highest value and the Davies-Bouldin indices attain the lowest value. Since there are no agreement among the four metrics, we will go with the majority. Our suggested number of clusters is three.

1. (10 points) What are the cluster centroids of your suggested cluster solution? Please show the centroids in their original scales.

The centroids, in their original scales, of the three-cluster solution are:

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Weight | Wheelbase | Length |
| 0 | 4606.55 | 121.36 | 206.93 |
| 1 | 2820.71 | 99.98 | 171.45 |
| 2 | 3670.04 | 108.46 | 188.07 |