CS 584-04: Machine Learning

Fall 2018 Assignment 2 Answer Key

# Question 1 (20 points)

Suppose a market basket can possibly contain these seven items: A, B, C, D, E, F, and G.

1. (1 point) What is the number of possible itemsets?

Number of possible itemsets is: 27 – 1 = 128 – 1 = 127.

1. (3 points) List all the possible 1-itemsets.

There are seven 1-itemsets. They are: {A}, {B}, {C}, {D}, {E}, {F}, {G}.

1. (3 points) List all the possible 2-itemsets.

There are twenty-one 2-itemsets. They are:  
{A,B}, {A,C}, {A,D}, {A,E}, {A,F}, {A,G},  
{B,C}, {B,D}, {B,E}, {B,F}, {B,G},  
{C,D}, {C,E}, {C,F}, {C,G},  
{D,E}, {D,F}, {D,G},  
{E,F}, {E,G},  
{F,G}.

1. (3 points) List all the possible 3-itemsets.

There are thirty-five 3-itemsets. They are:   
{A,B,C}, {A,B,D}, {A,B,E}, {A,B,F}, {A,B,G}, {A,C,D}, {A,C,E}, {A,C,F}, {A,C,G}, {A,D,E}, {A,D,F}, {A,D,G},  
{A,E,F}, {A,E,G}, {A,F,G},  
{B,C,D}, {B,C,E}, {B,C,F}, {B,C,G}, {B,D,E}, {B,D,F}, {B,D,G}, {B,E,F}, {B,E,G}, {B,F,G},  
{C,D,E}, {C,D,F}, {C,D,G}, {C,E,F}, {C,E,G}, {C,F,G},  
{D,E,F}, {D,E,G}, {D,F,G},  
{E,F,G}.

1. (3 points) List all the possible 4-itemsets.

There are thirty-five 4-itemsets. They are:  
{A,B,C,D}, {A,B,C,E}, {A,B,C,F}, {A,B,C,G}, {A,B,D,E}, {A,B,D,F}, {A,B,D,G}, {A,B,E,F}, {A,B,E,G}, {A,B,F,G}, {A,C,D,E}, {A,C,D,F}, {A,C,D,G}, {A,C,E,F}, {A,C,E,G}, {A,C,F,G}, {A,D,E,F}, {A,D,E,G}, {A,D,F,G}, {A,E,F,G},  
{B,C,D,E}, {B,C,D,F}, {B,C,D,G}, {B,C,E,F}, {B,C,E,G}, {B,C,F,G}, {B,D,E,F}, {B,D,E,G}, {B,D,F,G},  
{B,E,F,G},  
{C,D,E,F}, {C,D,E,G}, {C,D,F,G}, {C,E,F,G},  
{D,E,F,G}.

1. (3 points) List all the possible 5-itemsets.

There are twenty-one 5-itemsets. They are:  
{A,B,C,D,E}, {A,B,C,D,F}, {A,B,C,D,G}, {A,B,C,E,F}, {A,B,C,E,G}, {A,B,C,F,G}, {A,B,D,E,F}, {A,B,D,E,G}, {A,B,D,F,G}, {A,B,E,F,G}, {A,C,D,E,F}, {A,C,D,E,G}, {A,C,D,F,G}, {A,C,E,F,G}, {A,D,E,F,G},  
{B,C,D,E,F}, {B,C,D,E,G}, {B,C,D,F,G}, {B,C,E,F,G}, {B,D,E,F,G},  
{C,D,E,F,G}.

1. (3 points) List all the possible 6-itemsets.

There are seven 6-itemsets. It is {A,B,C,D,E,F}, {A,B,C,D,E,G}, {A,B,C,D,F,G}, {A,B,C,E,F,G}, {A,B,D,E,F,G}, {A,C,D,E,F,G}, {B,C,D,E,F,G}.

1. (1 point) List all the possible 7-itemsets.

The is only one 7-itemset. It is {A, B, C, D, E, F, G}.

# Question 2 (30 points)

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

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

The data is already sorted in ascending order by Customer and then by Item. Also, all the items bought by each customer are all distinct.

After you have imported the CSV file, please discover association rules using this dataset.

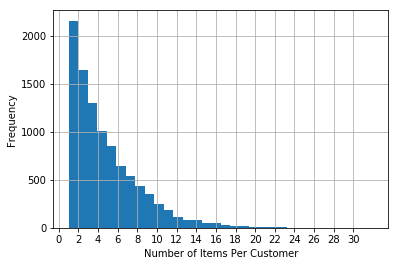
1. (2 points) How many customers in this market basket data?

There are 9,835 unique customers.

1. (2 points) How many unique items in the market basket across all customers?

There are 169 unique items across all customers.

1. (5 points) Create a dataset which contains the number of distinct items in each customer’s market basket. Draw a histogram of the number of unique items. What are the median, the 25th percentile and the 75th percentile in this histogram?



The median number of unique items is 3, the 25th percentile is 2 and the 75th percentile is 6.

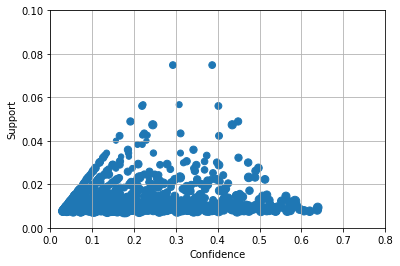
1. (5 points) Find out the *k*-itemsets which appeared in the market baskets of at least seventy five (75) customers. How many itemsets have you found? Also, what is the highest *k* value in your itemsets?

We specify the minimum support as MIN\_SUPPORT = 75/9835. Since the maximum number of items ever purchased by a customer is 32, we specify the maximum length of an itemset as MAX\_LEN = 32 accordingly. Based on the results, the highest *k* value is 4.

1. (5 points) Find out the association rules whose Confidence metrics are at least 1%. How many association rules have you found? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent.

We specify MIN\_THRESHOLD = 0.01. We found 1,228 association rules.

1. (5 points) Graph the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (e). Please use the Lift metrics to indicate the size of the marker.



1. (5 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |

1. (1 point) What similarities do you find among the consequents that appeared in (g)?

* Consequents are “whole milk”
* Antecedents contain one of these items: “butter”, “root vegetables”, “other vegetables”, or “yogurt”.

# Question 3 (20 points)

You are asked to write a Python program to calculate the Elbow value and the Silhouette value. For this question, you will use the CARS.CSV dataset to test your program. Here are the specifications for performing the respective analyses.

**Clustering**

* The input interval variables are Horsepower and Weight
* The distance metric is Euclidean
* The maximum number of clusters is 15
* Consider the silhouette\_score function for calculating the Silhouette value.

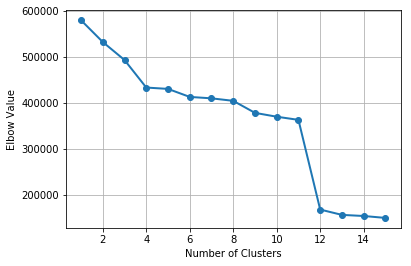
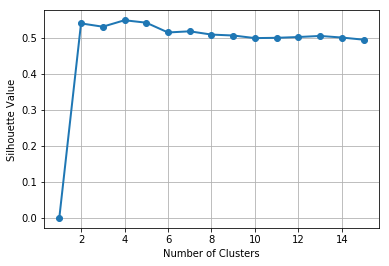
Please answer the following questions.

1. (15 points) List the Elbow values and the Silhouette values for your 1-cluster to 15-cluster solutions.

|  |  |  |
| --- | --- | --- |
| Number of Clusters | Elbow Value | Silhouette |
| 1 | 579857.9543 | 0.0000 |
| 2 | 532455.2722 | 0.5391 |
| 3 | 493218.0813 | 0.5300 |
| 4 | 433215.8150 | 0.5479 |
| 5 | 430290.4574 | 0.5411 |
| 6 | 412804.9312 | 0.5140 |
| 7 | 409729.7423 | 0.5172 |
| 8 | 404285.7518 | 0.5081 |
| 9 | 378087.1355 | 0.5056 |
| 10 | 369686.6227 | 0.4984 |
| 11 | 362825.8531 | 0.4992 |
| 12 | 167426.0602 | 0.5014 |
| 13 | 155828.0892 | 0.5044 |
| 14 | 153411.7102 | 0.4999 |
| 15 | 149219.2400 | 0.4940 |

1. (5 points) Based on the Elbow values and the Silhouette values, what do you suggest for the number of clusters?

Based on the two charts below, I would determine the number of clusters as 4. It is because the first elbow occurs at 4 and the Silhouette value attains a local maximum at 4 too.

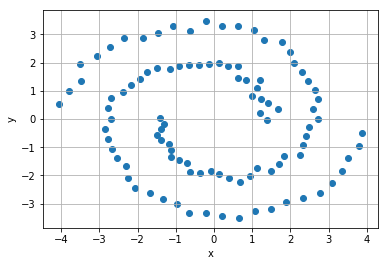
 

# Question 4 (30 points)

Apply the Spectral Clustering method to the Spiral.csv. Your input fields are x and y.

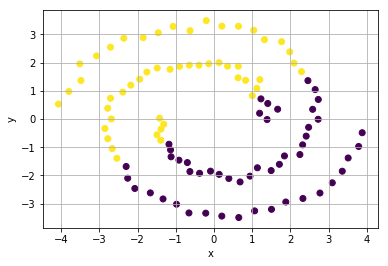
1. (5 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?

Since there are two spirals visually, I would say intuitively that the number of clusters is 2.



1. (5 points) Apply the K-mean algorithm directly using your number of clusters (in a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?

The chart below is the results of applying the K-mean algorithm to find a 2-cluster solution. Obviously, this cluster solution is not what we have expected.

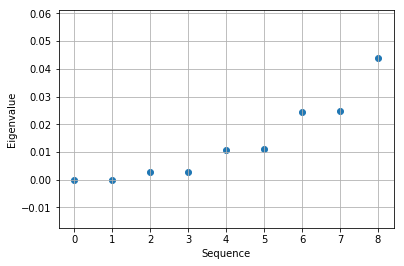


1. (5 points) Apply the nearest neighbor algorithm using the Euclidean distance. How many nearest neighbors will you use?

Actually, I choose three neighbors for the nearest neighbor algorithm after trying other values. Eventually, the three neighbors’ solution leads to the desirable clustering results.

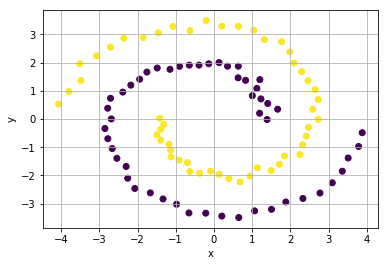
1. (5 points) Generate the sequence plot of the first nine eigenvalues, starting from the smallest eigenvalues. Based on this graph, do you think your number of nearest neighbors (in a) is appropriate?

Visually, the eigenvalues jump between 1 and 2, 3 and 4, 5 and 6, and 7 and 8. In other words, the chart suggests us to try 3, 5, 7 and 9 for the number of neighbors. It turns out 3 is the most appropriate choice.



1. (5 points) Apply the K-mean algorithm on your first two eigenvectors that correspond to the first two smallest eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?

The chart below is the results of applying the K-mean algorithm to find a 2-cluster solution using the first two eigenvectors. Obviously, this cluster solution meets our expectation.



1. (5 points) Comment on your spectral clustering results?

The short comment is that the spectral clustering is the appropriate technique to find the clusters. It definitely produces our expected results.