CS 584-04: Machine Learning

Spring 2020 Assignment 2

Question 1 (35 points)

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

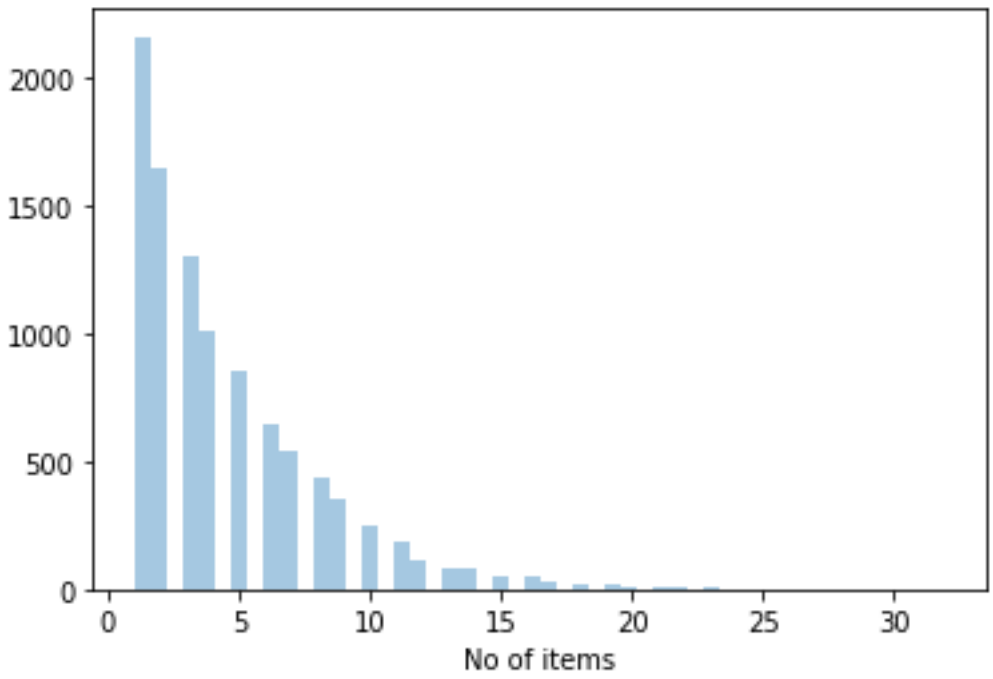
1. Customer: Customer Identifier
2. Item: Name of Product Purchased

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) (5 points) Create a data frame that contains the number of unique items in each customer's market basket. Draw a histogram of the number of unique items. What are the 25th, 50th, and the 75th percentiles of the histogram?

Ans.

Histogram of the number of unique items



Percentiles of Histogram:

25th Percentile of Histogram: 2.0

50th Percentile of Histogram: 3.0

75th Percentile of Histogram: 6.0

b) (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 can we find? Also, what is the largest *k* value among our item-sets?

Ans**. Total number of item-sets with at least 75 customers are 522**

The largest k-value among the item-sets is 3

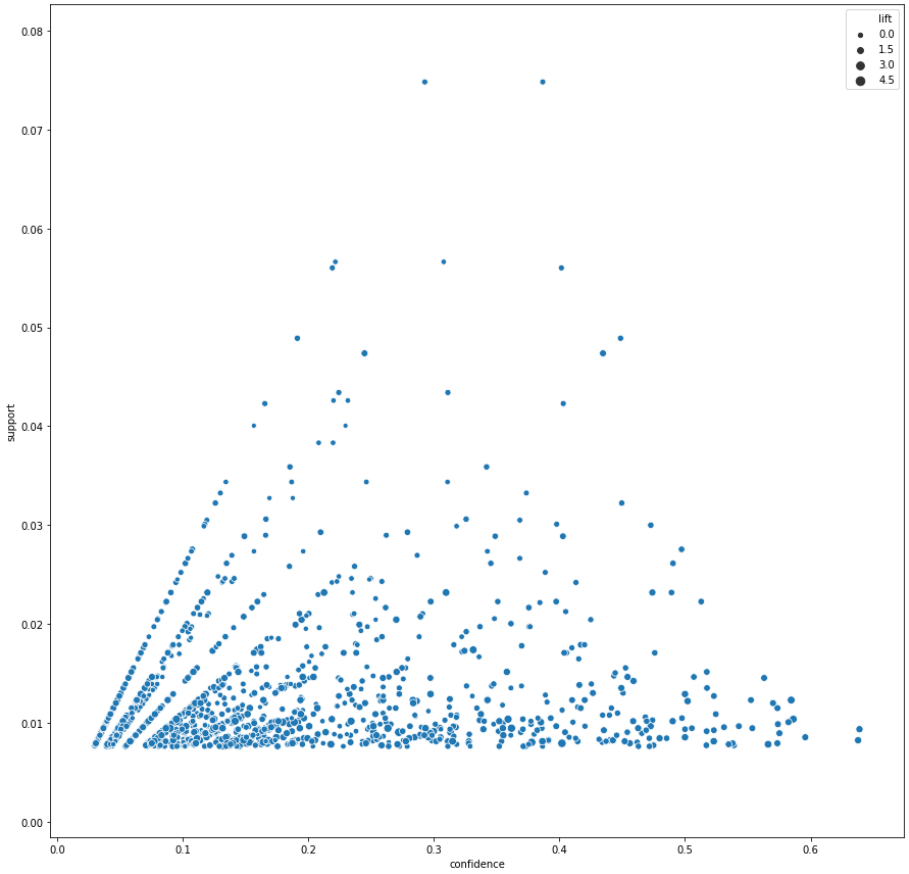
c) (10 points) Find out the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent. Please **do not** display those rules in your answer.

Ans. **Total number of Association Rules with > 1% Confidence are 1200**

d) (5 points) Plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you have found in (c). Please use the Lift metrics to indicate the size of the marker.

Ans.

Support Vs Confidence



e) (5 points) List the rules whose Confidence metrics are greater than or equal to 60%. Please include their Support and Lift metrics.

Ans. Association rules with confidence >= 60%

antecedents	consequents	confidence	support	lift
(root vegetables, butter)	(whole milk)	0.637795	0.008236	2.496107
(yogurt, butter)	(whole milk)	0.638889	0.009354	2.500387

Question 2 (30 points)

The K-means algorithm works only with interval features. One way to apply the k-means algorithm to categorical features is to transform them into a new interval feature space. However, this approach can be very inefficient, and it does not produce good results.

For clustering categorical features, we should consider the K-modes clustering algorithm which extends the K-means algorithm by using different dissimilarity measures and a different method for computing cluster centers. See this article for more details. Huang, Z. (1997). "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." In Proceedings of the *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1–8. New York: ACM Press.

Please implement the K-modes clustering method in Python and then apply the method to the cars.csv. Your input fields are these four categorical features: Type, Origin, DriveTrain, and Cylinders. **Please do not remove the missing or blank values in these four features**. Instead, consider these values as a separate category.

The cluster centroids are the modes of the input fields. In the case of tied modes, choose the lexically or numerically lowest one.

Suppose a categorical feature has observed values $v_1, \ldots, v_p$. Their frequencies (i.e., number of observations) are $f_1, \ldots, f_p$. The distance metric between two values is $d(v_i, v_j) = 0$ if $v_i = v_j$. Otherwise, $d(v_i, v_j) = \frac{1}{f_i} + \frac{1}{f_j}$. The distance between any two observations is the sum of the distance metric of the four categorical features.

a) (5 points) What are the frequencies of the categorical feature Type?

Ans.

Type	Count
Sedan	262
SUV	60
Sports	49
Wagon	30
Truck	24
Hybrid	3

b) (5 points) What are the frequencies of the categorical feature DriveTrain?

Ans.

DriveTrain	Count
FWD	226
RWD	110
AWD	92

c) (5 points) What is the distance between Origin = 'Asia' and Origin = 'Europe'?

Ans. **The distance between ASIA and EUROPE is <u>0.014459195224863643</u>**

d) (5 points) What is the distance between Cylinders = 5 and Cylinders = Missing?

Ans. **The distance between 5 and MISSING is <u>0.6428571428571428</u>**

e) (5 points) Apply the K-modes method with **three clusters**. How many observations in each of these three clusters? What are the centroids of these three clusters?

Ans. **<u>Cluster 0:</u> 250 observations**

<u>Cluster 1</u>: 107 observations

<u>Cluster 2</u>: 71 observations

<u>Centroids</u>

Centroid for Cluster 0: 'Sedan' 'Asia' ' FWD' ' 6.0'

Centroid for Cluster 1: 'Sedan' 'Europe' 'RWD' '8.0'

Centroid for Cluster 2: 'Sedan' 'USA' 'FWD' '4.0'

f) (5 points) Display the frequency distribution table of the Origin feature in each cluster.

Ans.

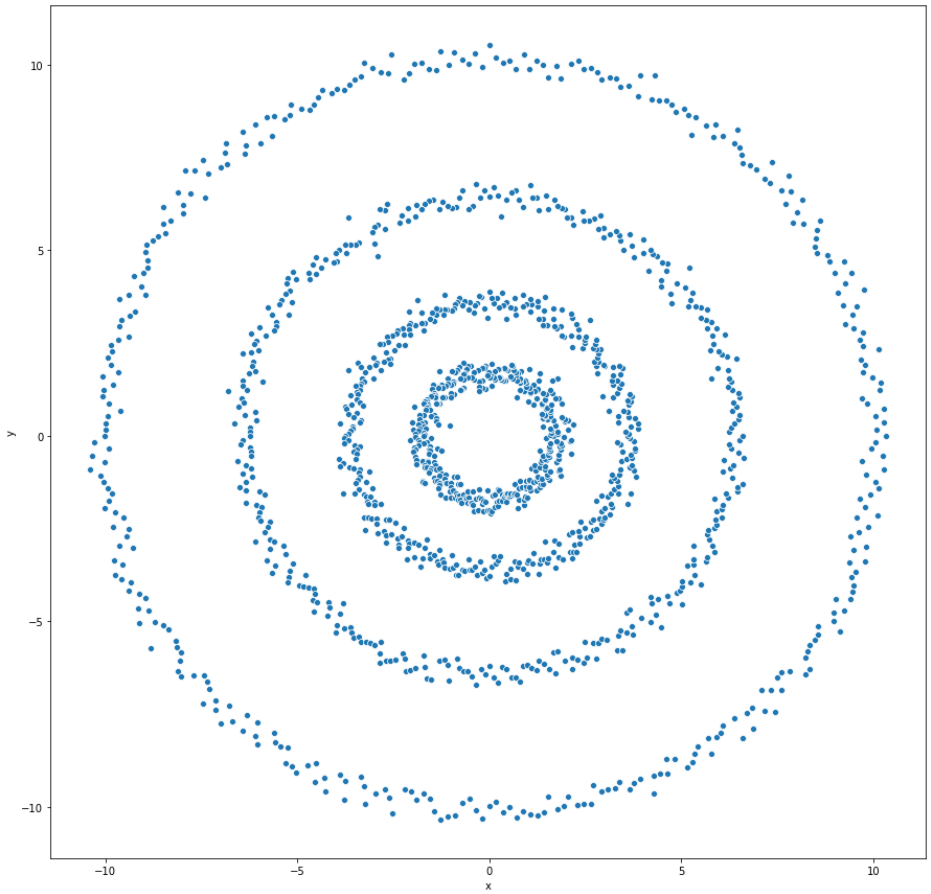
Cluster	Origin	Count
0	Asia	153
	Europe	30
	USA	67
1	Asia	5
	Europe	71
	USA	31
2	Europe	22
	USA	49

Question 3 (35 points)

Apply the Spectral Clustering method to the FourCircle.csv. Your input fields are x and y. Wherever needed, specify random_state = 60616 in calling the KMeans function.

g) (5 points) Plot y on the vertical axis versus x on the horizontal axis. How many clusters are there based on your visual inspection?

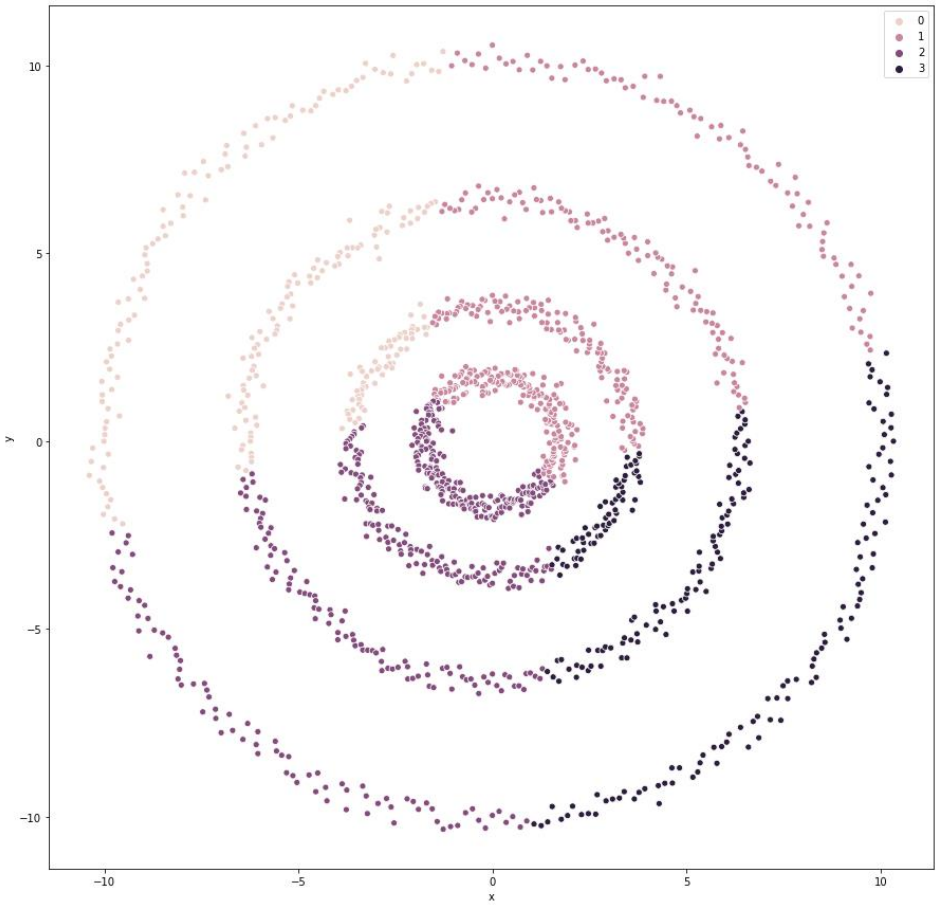
Ans.



By visual inspection, number of clusters are 4

h) (5 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifiers to control the color scheme. Please comment on this K-mean result.

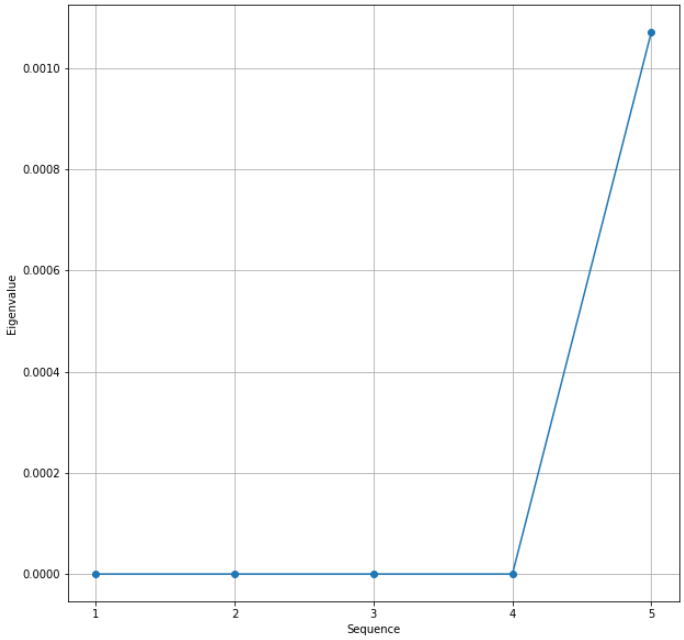
Ans.



Each cluster drops into each quadrant of the graph.
Cluster 0 is in quadrant 2.
Cluster 1 is in quadrant 1.
Cluster 2 is in quadrant 3.
Cluster 3 is in quadrant 4.

i) (10 points) Apply the nearest neighbor algorithm using the Euclidean distance. We will consider the number of neighbors from 1 to 15. What is the smallest number of neighbors that we should use to discover the clusters correctly? Remember that we may need to try a couple of values first and use the eigenvalue plot to validate our choice.

Ans. **The smallest number of neighbors are <u>6 (Six)</u> to discover the clusters correctly.**



Number of Clusters are 4(Four)

j) (5 points) Using your choice of the number of neighbors in (c), calculate the Adjacency matrix, the Degree matrix, and finally the Laplacian matrix. How many eigenvalues do you determine are practically zero? Please display their calculated values in scientific notation.

Ans. **Adjacency Matrix:**

```
[[1.       0.       0.       ... 0.       0.       0.       ]
[0.       1.       0.       ... 0.       0.       0.       ]
[0.       0.       1.       ... 0.       0.96602229 0.       ]
...
[0.       0.       0.       ... 1.       0.       0.       ]
[0.       0.       0.96602229 ... 0.       1.       0.       ]
[0.       0.       0.       ... 0.       0.       1.       ]]
```

Degree Matrix:

```
[[0. 0. 0. ... 0. 0. 0.]
[0. 0. 0. ... 0. 0. 0.]
[0. 0. 0. ... 0. 0. 0.]
 ...
[0. 0. 0. ... 0. 0. 0.]
[0. 0. 0. ... 0. 0. 0.]
[0. 0. 0. ... 0. 0. 0.]]
```

Laplace Matrix:

```
[[ 3.80117773  0.         0.         ...  0.         0.
0.        ]
[ 0.         3.29598338  0.         ...  0.         0.
0.        ]
[ 0.         0.         4.55116784 ...  0.        -0.96602229
0.        ]
...
[ 0.         0.         0.         ...  4.29371731  0.
0.        ]
[ 0.         0.        -0.96602229 ...  0.         3.88916173
0.        ]
[ 0.         0.         0.         ...  0.         0.
3.94116662]]
```

Practically Zero Eigen Values are 4.

Eigenvalue 1: -5.840164970722635e-15
Eigenvalue 2: 1.6481834523013946e-16
Eigenvalue 3: 2.911241034756213e-16
Eigenvalue 4: 1.0391558562349583e-15

So there are Four Clusters

k) (10 points) Apply the K-mean algorithm on the eigenvectors that correspond to your "practically" zero eigenvalues. The number of clusters is the number of your "practically" zero eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme.

Ans.