CS 584-04: Machine Learning

Spring 2020 Assignment 2

Question 1 (35 points)

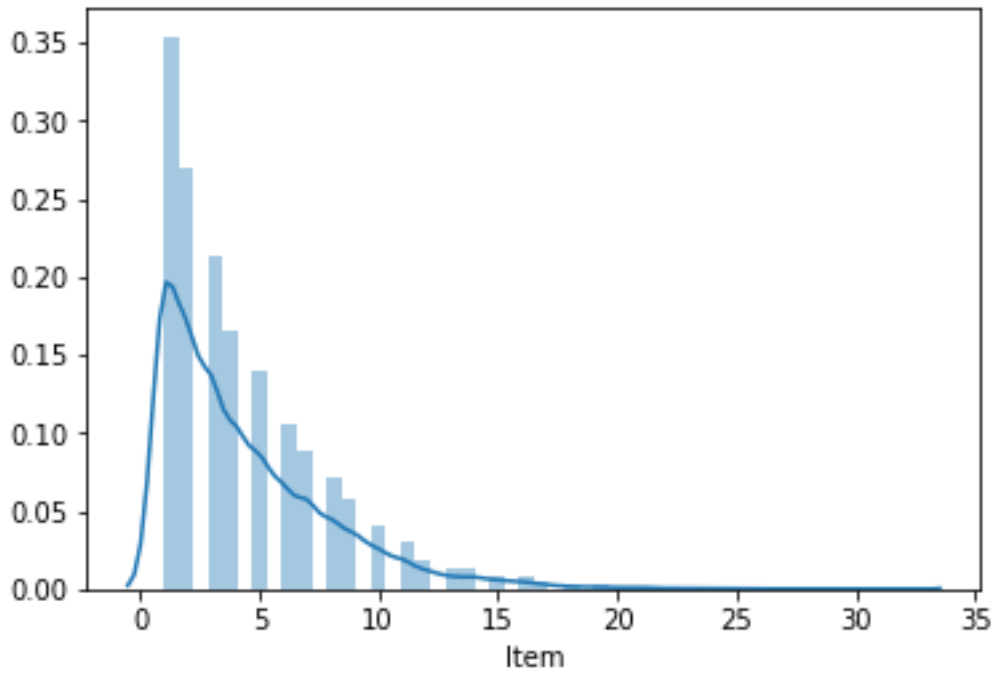
The file Groceries.csv contains market basket data. The variables are:

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**



The 25th percentile of the histogram : 2.0
The median of this histogram is: 3.0
The 75 percentile of this histogram is : 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 itemsets?

ANS:
Frequent itemsets are:

	Support	Itemsets
0	0.008033	(Instant food products)
1	0.033452	(UHT-milk)
2	0.017692	(baking powder)
3	0.052466	(beef)
4	0.033249	(berries)

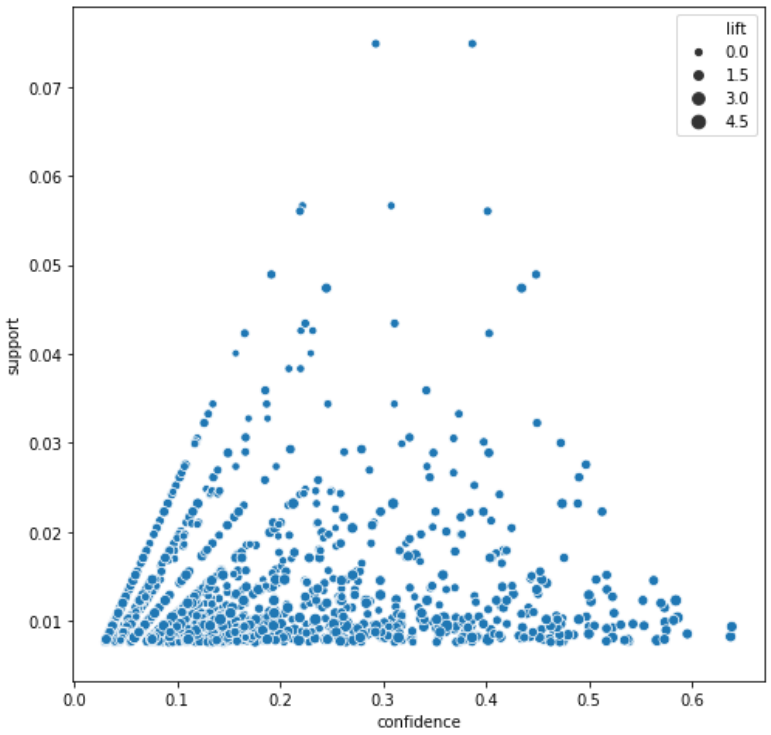
Total number of frequeny itemsets 522
The highest value of k in the itemset: 3

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: **No of associaton rules 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: **Scatter plot for support and confidence metrics**



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

ANS: **Rules whose confidence metrics are greater than or equal to 60%**

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(butter, root vegetables)	(whole milk)	0.012913	0.255516	0.008236	0.637795	2.496107	0.004936	2.055423
1	(yogurt, butter)	(whole milk)	0.014642	0.255516	0.009354	0.638889	2.500387	0.005613	2.061648

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 global frequencies (i.e., number of observations) are $f_1, \ldots, f_p$. Please be noted that these global frequencies do not change with the cluster assignment. 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:**Frequencies of the categorical feature Type**

	Type	Counts
0	Sedan	262
1	SUV	60
2	Sports	49
3	Wagon	30
4	Truck	24
5	Hybrid	3

b) (5 points) What are the frequencies of the categorical feature DriveTrain?
ANS: **Frequencies of DriveTrain**

	DriveTrain	Counts
0	FWD	226
1	RWD	110
2	AWD	92

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

ANS:Distance between 'Asia' and 'Europe' **is 0.0144592**

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

ANS:Distance between Cylinders=5 and Cylinder=Missing **is 0.6428571**

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: **Below are the observations and centroids of three clusters**

```
{0: 250, 1: 107, 2: 71}
[['Sedan' 'Asia' 'FWD' '6.0']
 ['Sedan' 'Europe' 'RWD' '8.0']
 ['Sedan' 'USA' 'FWD' '4.0']]
```

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

ANS: Frequency Distribution of Origin feature

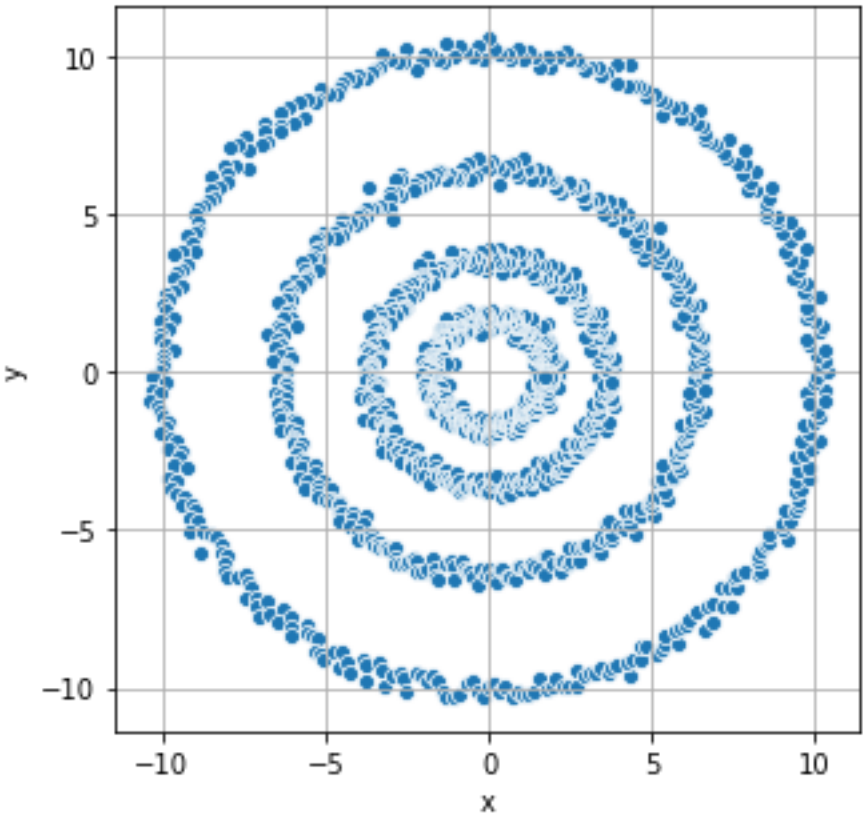
Cluster Origin

0	**Asia 153**
	Europe 30
	USA 67
1	**Asia 5**
	Europe 71
	USA 31
2	**Europe 22**
	USA 49

dtype: int64

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.

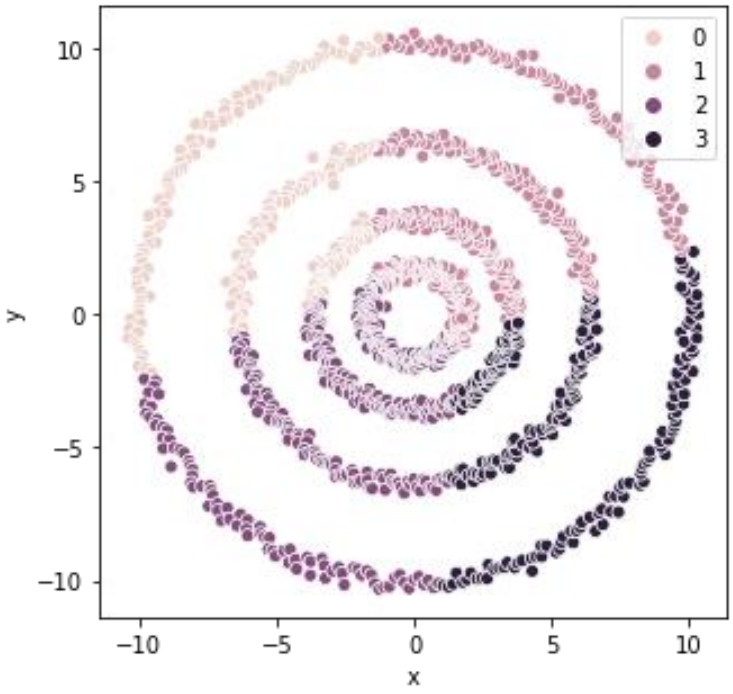
a) (5 points) Plot y on the vertical axis versus x on the horizontal axis. How many clusters are there based on your visual inspection?
ANS:



By visual inspection we can **observe 4 clusters**

b) (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

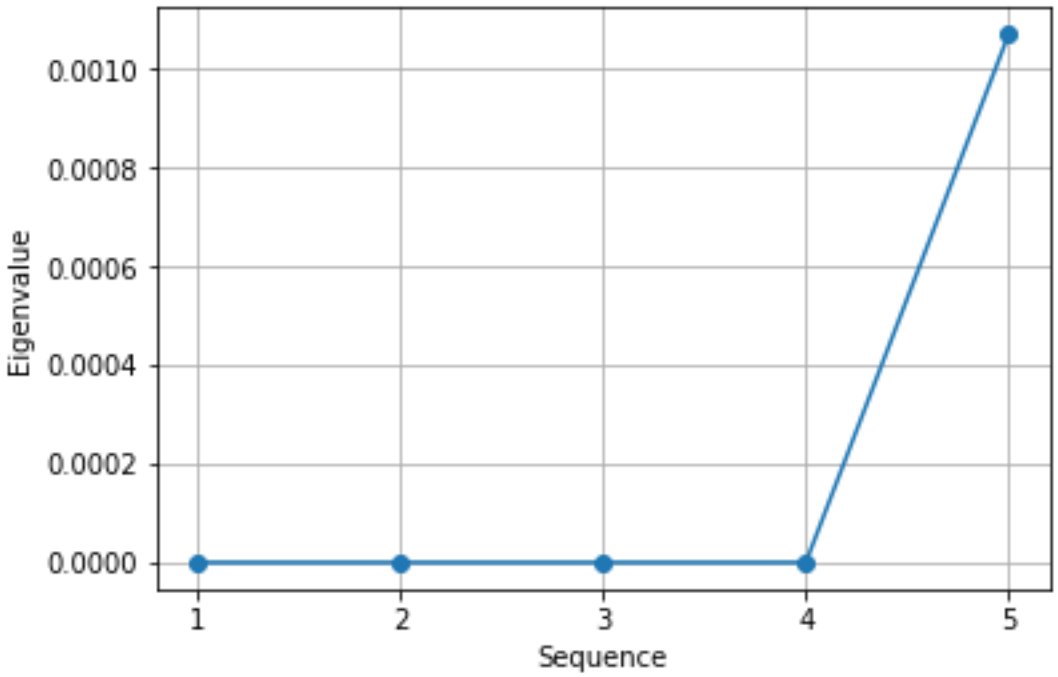
ANS:



The clusters are present in quadrants.Each cluster is there in one quadrant.
So **Cluster 0:Quadrant 2**
Cluster 1:Quadrant 1
Cluster 2:Quadrant 3
Cluster 3:Quadrant 4
Each point in the cluster is closer to its own cluster than to other cluster centers.

c) (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 that we should use to discover the clusters correctly **are 6.**



d) (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 values of the "zero" eigenvalues 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: [[4.80117773 0. 0. ... 0. 0. 0.]
[0. 4.29598338 0. ... 0. 0. 0.]
[0. 0. 5.55116784 ... 0. 0. 0.]
...
[0. 0. 0. ... 5.29371731 0. 0.]
[0. 0. 0. ... 0. 4.88916173 0.]
[0. 0. 0. ... 0. 0. 4.94116662]]
Laplacian 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]]

Eigen values :

Eigenvalue: -2.7188019913617817e-15
Eigenvalue: -7.549863207144599e-16
Eigenvalue: 5.394941757143433e-16
Eigenvalue: 1.0131799766542524e-15

4 eigen values are displayed which are practically 'zero'. So, we got 4 clusters.

e) (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: