CS 584-04: Machine Learning

Spring 2020 Assignment 2

Jason Yeoh (A20457826)

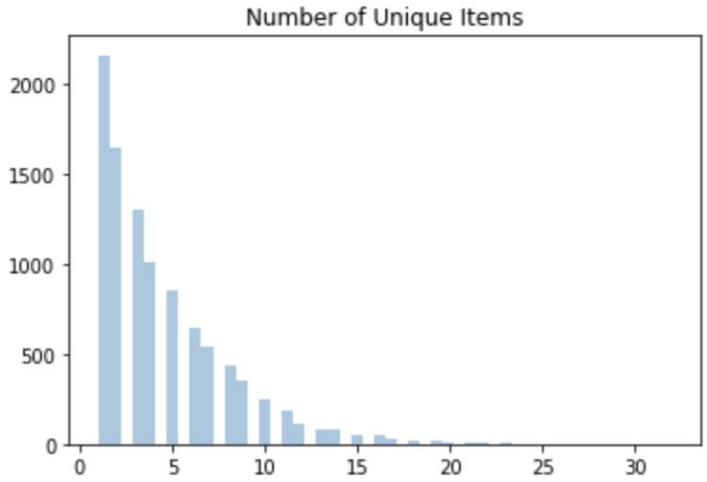
Question 1 (35 points)

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?



Frequency of Number of Items Purchase

1	2159
2	1643
3	1299
4	1005
5	855
6	645
7	545
8	438
9	350
10	246
11	182
12	117
13	78

14	77
15	55
16	46
17	29
18	14
19	14
20	9
21	11
22	4
23	6
24	1
26	1
27	1
28	1
29	3
32	1

25th Percentile: 2.00
50th Percentile: 3.00
75th Percentile: 6.00

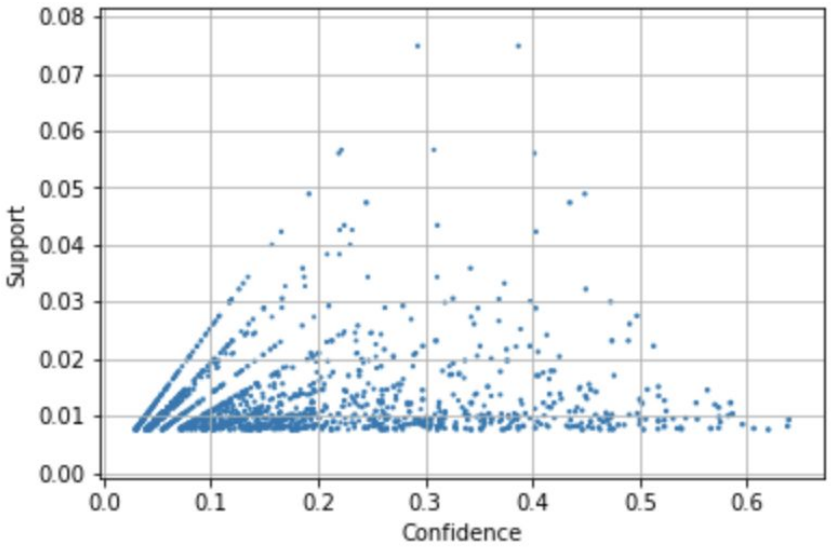
b) (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy five (75) customers. How many itemsets can we find? Also, what is the largest *k* value among our itemsets?

There are 524 itemsets with the largest k value of 4.

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.

There are 1228 association rules.

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.



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

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(root vegetables, butter)	(whole milk)	0.012913	0.255516	0.008236	0.637795	2.496107	0.004936	2.055423
1	(butter, yogurt)	(whole milk)	0.014642	0.255516	0.009354	0.638889	2.500387	0.005613	2.061648
2	(root vegetables, other vegetables, yogurt)	(whole milk)	0.012913	0.255516	0.007829	0.606299	2.372842	0.004530	1.890989
3	(other vegetables, yogurt, tropical fruit)	(whole milk)	0.012303	0.255516	0.007626	0.619835	2.425816	0.004482	1.958317

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?

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

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

FWD	**226**
RWD	**110**
AWD	**92**

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

831.5483173812232 units

For exact computation, check the .py code attached. Thanks!

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

36 units

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?

- ['SUV', 'USA', 'AWD', '6.0'] **145 observations**
- ['Sedan', 'Europe', 'RWD', '6.0'] **118 observations**
- ['Sedan', 'Asia', 'FWD', '4.0'] **165 observations**

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

Cluster 0: ['SUV', 'USA', 'AWD', '6.0']

USA 96

Asia 31

Europe 18

Cluster 1: ['Sedan', 'Europe', 'RWD', '6.0']

Europe 83

Asia 19

USA 16

Cluster 2: ['Sedan', 'Asia', 'FWD', '4.0']

Asia 108

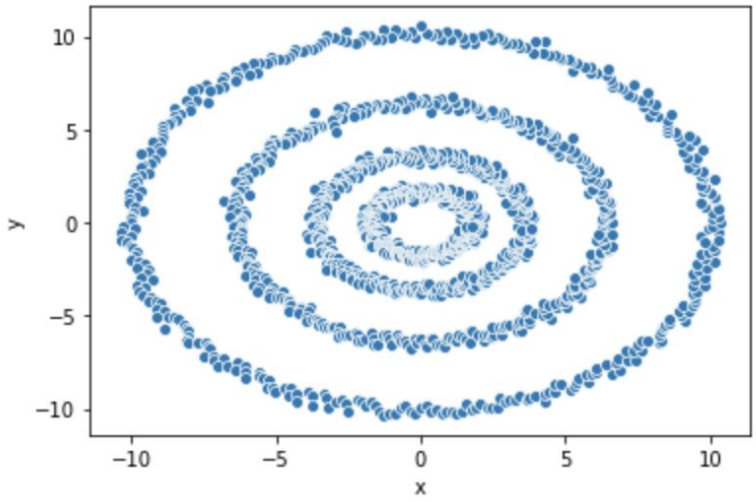
USA 35

Europe 22

Question 3 (35 points)

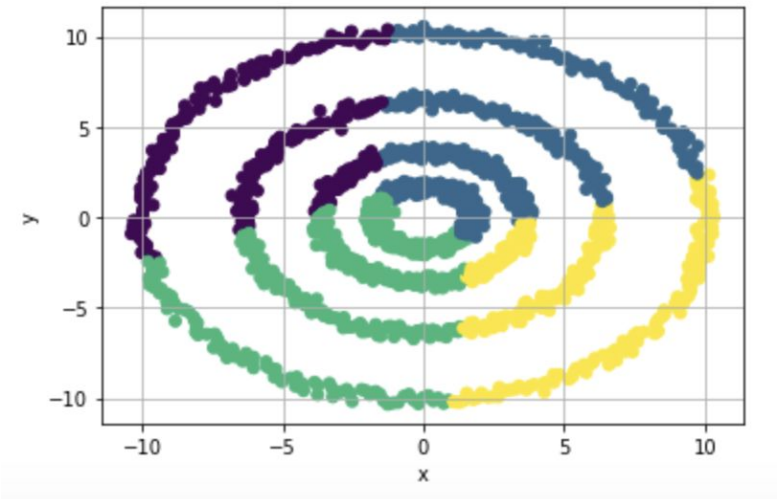
Apply the Spectral Clustering method to the FourCircle.csv. Your input fields are x and y. Wherever needed, specify random_state = 60616 in calling the KMeans function.

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?



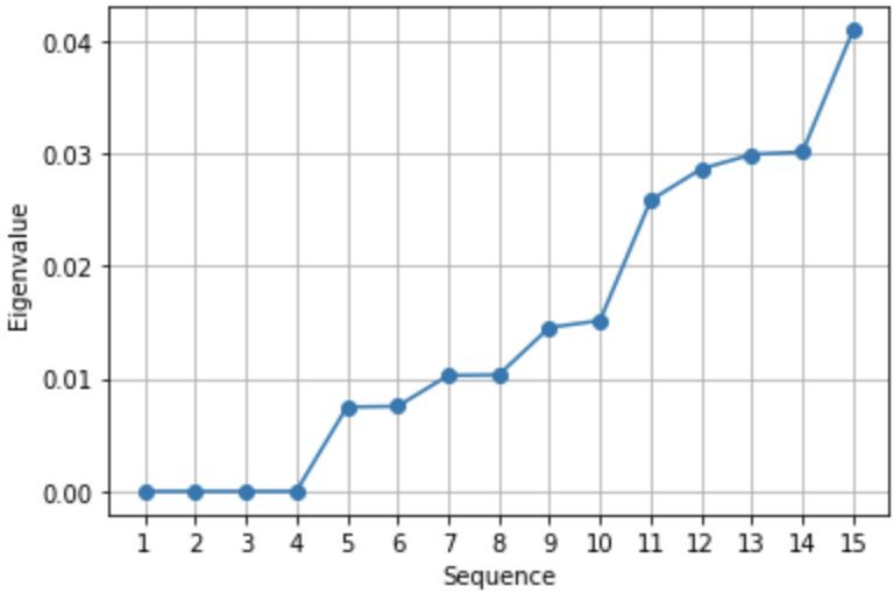
Based on observation, there are 4 spiral clusters on this plot.

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.



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.

10 neighbors



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.

There were 4 eigenvalues with zero value.

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.