

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:

Distinct Items of Customers:

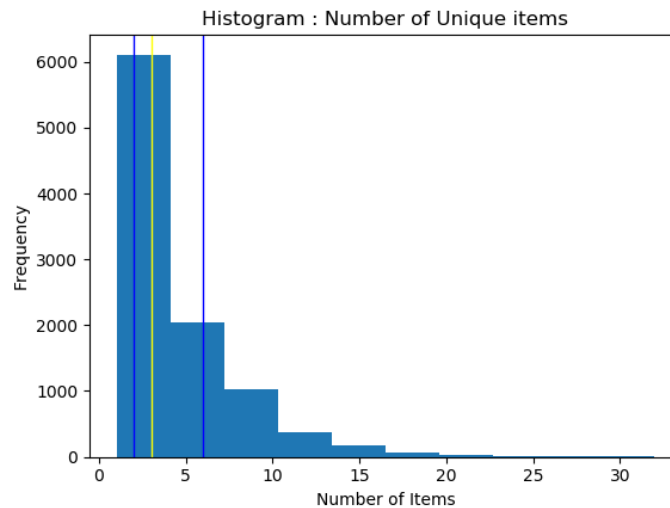
Customer

1	4
2	3
3	1
4	4
5	4
..	
9831	17
9832	1
9833	10
9834	4
9835	5

25th Percentile : 2.0

50th Percentile : 3.0

75th Percentile : 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:

Frequently bought Itemsets:

```

0          (Instant food products)
1          (UHT-milk)
2          (baking powder)
3          (beef)
4          (berries)
...
517        (tropical fruit, whole milk, soda)
518        (yogurt, whole milk, soda)
519 (tropical fruit, whole milk, whipped/sour cream)
520        (tropical fruit, yogurt, whole milk)
521        (yogurt, whole milk, whipped/sour cream)

```

Name: itemsets, Length: 522, dtype: object

Total No. of Itemsets: 522

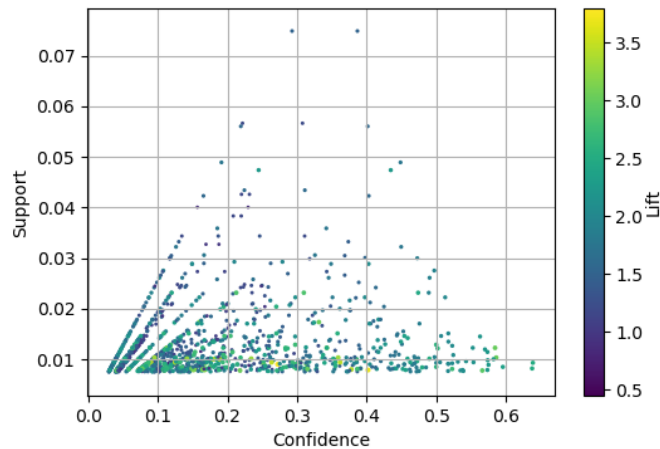
Largest k value: 3

- 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: Number of Association Rules Found: 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:



- 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:

	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, \dots, v_p . Their frequencies (i.e., number of observations) are f_1, \dots, 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: **Frequencies of the categorical feature Type:**
Hybrid : 3
SUV : 60
Sedan : 262
Sports : 49
Truck : 24
Wagon : 30

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

Ans: **Frequencies of the categorical feature DriveTrain:**
AWD : 92
FWD : 226
RWD : 110

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

Ans: **Frequencies of the categorical feature Origin:**
Asia : 158
Europe : 123
USA : 147

Distance between Origin=Asia and Origin=Europe:
0.014459195224863643

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

Ans: Frequencies of the categorical feature Cylinders:
6.0 190
4.0 136
8.0 87
5.0 7
12.0 3
10.0 2
3.0 1
Missing Values: 2
Distance between Cylinders=5 and Cylinders=Missing:
0.6428571428571428

- 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: Number of Observation in in three clusters:
1. 217
2. 116
3. 95

Centroids:
[['Sedan' 'Asia' 'FWD' '4.0']
['Sedan' 'USA' 'FWD' '6.0']
['Sedan' 'Europe' 'RWD' '8.0']]

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

Ans:

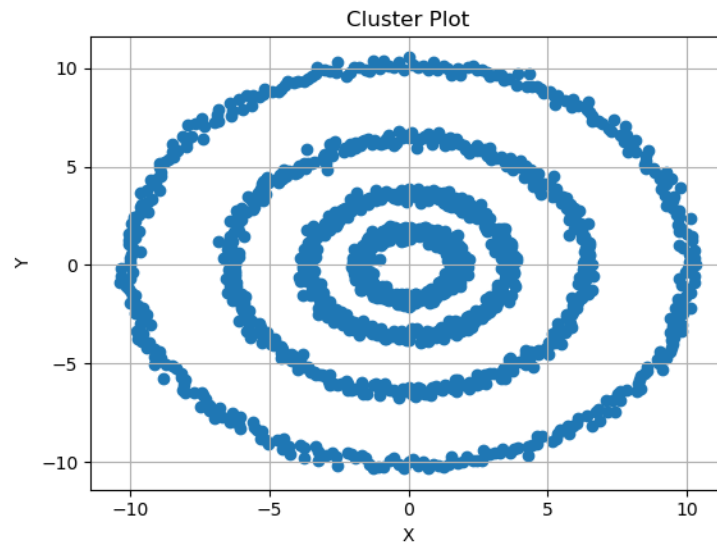
Cluster_number	0	1	2
Origin			
Asia	153	0	5
Europe	27	26	70
USA	37	90	20

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.

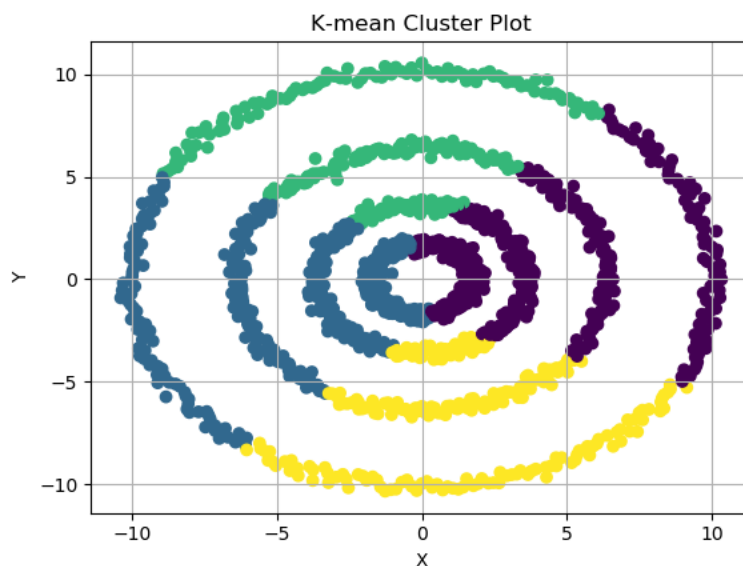
- g) (5 points) Plot y on the vertical axis versus x on the horizontal axis. How many clusters are there based on your visual inspection?

Ans: Based on the visual inspection we can see there are 4 clusters



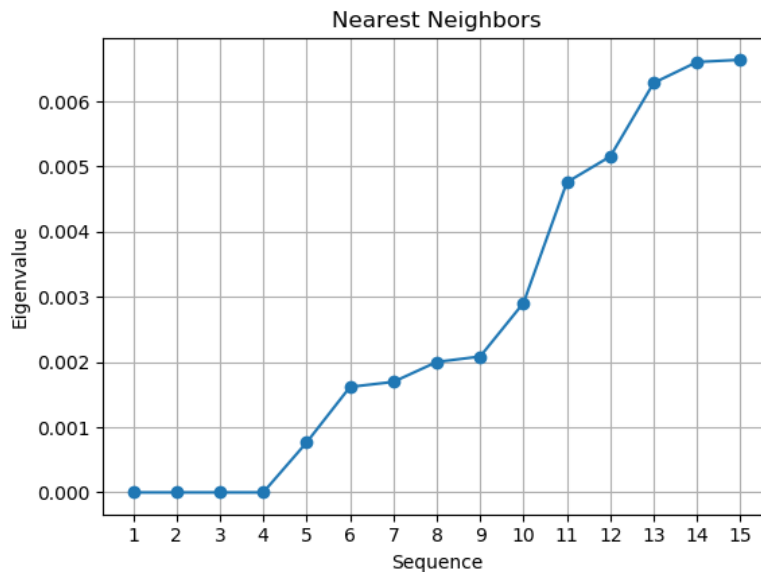
- h) (5 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifiers to control the color scheme. Please comment on this K-mean result.

Ans: As we can see from the image K-mean doesn't give proper clusters. Thus k-means doesn't make proper clusters with this type of spiral representation.



- 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: 6 is the smallest number of neighbors that we should use to discover the clusters correctly. At 6 we can see a steep increase in the graph.



- 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: There are 4 eigen values that are determined practically zero.

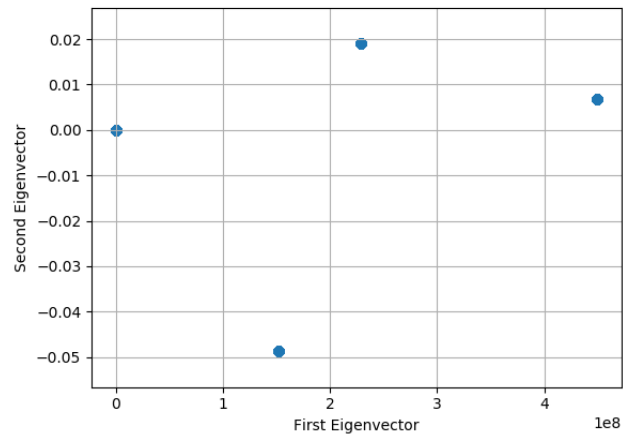
Eigen Values:

-1.688394e-15

-1.103030e-15

-1.022557e-15

-5.517553e-17



- 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:

