CS 584-04: Machine Learning

Spring 2020 Assignment 2

Question 1 (35 points)

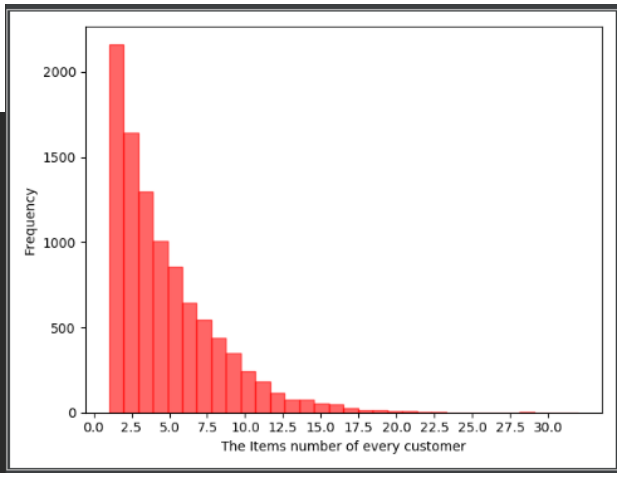
The file Groceries.csv contains market basket data. The variables are:

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?

```
count    9835.000000
mean        4.409456
std         3.589385
min         1.000000
25%         2.000000
50%         3.000000
75%         6.000000
max        32.000000
```



According to the describe function of the data, the median is 3, the 25th percentile is 2 and 75th percentile is 6.

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?

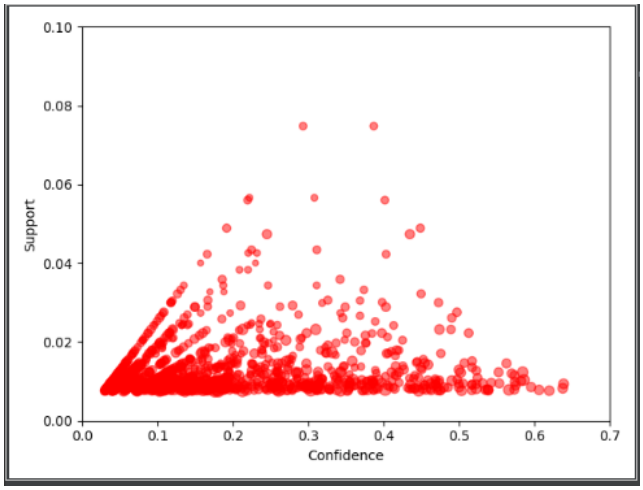
	support	itemsets
0	0.008033	(Instant food products)
1	0.033452	(UHT-milk)
2	0.017692	(baking powder)
3	0.052466	(beef)
4	0.033249	(berries)
..	...	...
519	0.007931	(tropical fruit, whole milk, whipped/sour cream)
520	0.015150	(yogurt, tropical fruit, whole milk)
521	0.010880	(yogurt, whole milk, whipped/sour cream)
522	0.007829	(yogurt, other vegetables, root vegetables, wh...
523	0.007626	(yogurt, other vegetables, tropical fruit, who...

Because the minimum support is 75/9835 and the max are 32 (According to the maximum number of distinct items bought by the customer). So we can get the result and the largest k value is 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.

```
We can find 1228 Association rules
                        antecedents  ... conviction
0                         (UHT-milk)  ...   1.065626
1                (other vegetables)  ...   1.008964
2                         (UHT-milk)  ...   1.069409
3                              (soda)  ...   1.010750
4                    (baking powder)  ...   1.560725
...                              ...  ...        ...
1223  (tropical fruit, whole milk)  ...   1.166976
1224                        (yogurt)  ...   1.039756
1225             (other vegetables)  ...   1.025257
1226               (tropical fruit)  ...   1.054357
1227                    (whole milk)  ...   1.018081
```

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.

	A	B	C	D	E	F	G	H	I	J
		antecedents	consequents	antecedent support	onsequent suppor	support	confidence	lift	leverage	conviction
726		frozenset({'butter', 'root vegetables'})	frozenset({'whole milk'})	0.012913066	0.255516014	0.00823589	0.637795276	2.4961069	0.004936397	2.055423178
732		frozenset({'butter', 'yogurt'})	frozenset({'whole milk'})	0.014641586	0.255516014	0.00935435	0.638888889	2.5003869	0.005613187	2.061647961
1203		frozenset({'yogurt', 'root vegetables', 'other vegetables'})	frozenset({'whole milk'})	0.012913066	0.255516014	0.00782918	0.606299213	2.3728423	0.004529686	1.890989324
1217		frozenset({'yogurt', 'other vegetables', 'tropical fruit'})	frozenset({'whole milk'})	0.012302999	0.255516014	0.00762583	0.619834711	2.4258155	0.004482213	1.958316571

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?

	index	total
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?

	index	total
0	FWD	226
1	RWD	110
2	AWD	92

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

```
The distance metric between 'Asia' and 'Europe' is: 0.0145
```

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

```
The distance metric between 'Asia' and 'Europe' is: 0.0145
```

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?

```
The number of observations in cluster 1: 207
The number of observations in cluster 2: 152
The number of observations in cluster 3: 69

The cluster 1 Data is :  ['Sedan', 'Asia', 'FWD', 4.0]
The cluster 2 Data is :  ['Sedan', 'Europe', 'FWD', 6.0]
The cluster 3 Data is :  ['SUV', 'USA', 'AWD', 8.0]
```

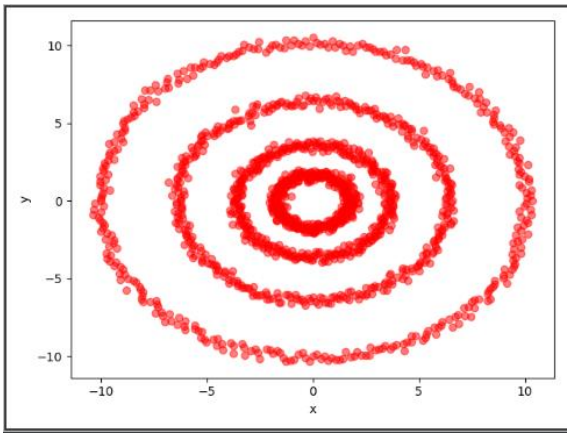
f) (5 points) Display the frequency distribution table of the Origin feature in each cluster.

```
The Cluster 1 is :
Asia      144
USA        38
Europe     25
Name: Origin, dtype: int64

The Cluster 2 is :
Europe    90
USA       62
Asia       0
Name: Origin, dtype: int64

The Cluster 3 is :
USA       47
Asia      14
Europe     8
Name: Origin, dtype: int64
```

Question 3 (35 points)

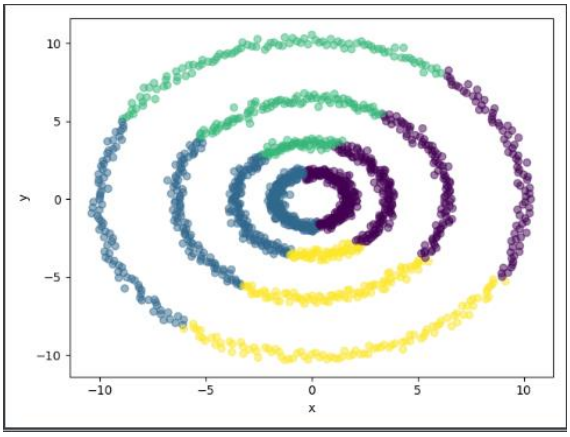
Apply the Spectral Clustering method to the FourCircle.csv. Your input fields are x and y. Wherever needed, specify random_state = 60616 in calling the KMeans function.

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?



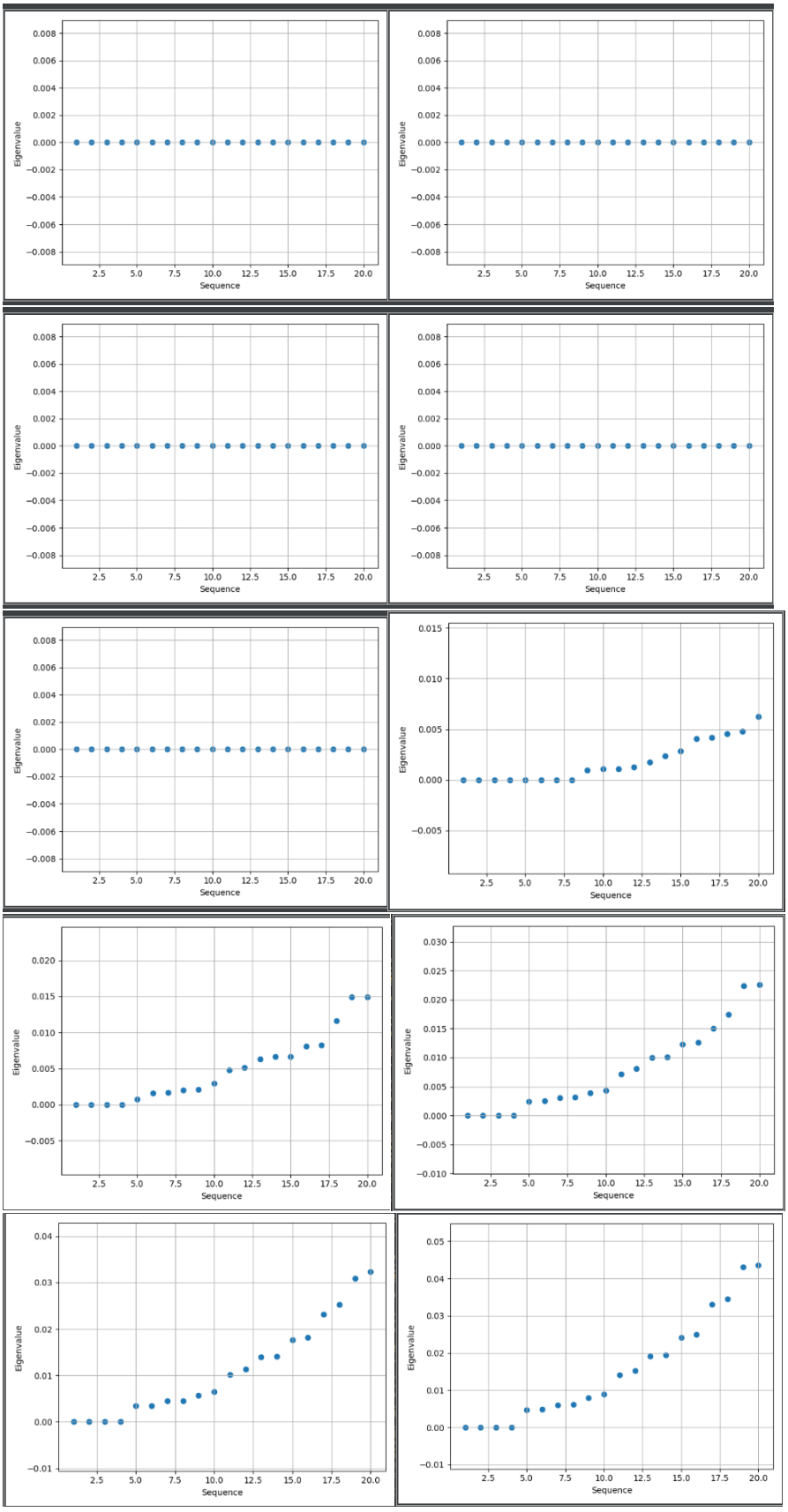
According to the graph of the data, there are four clusters by visual inspection.

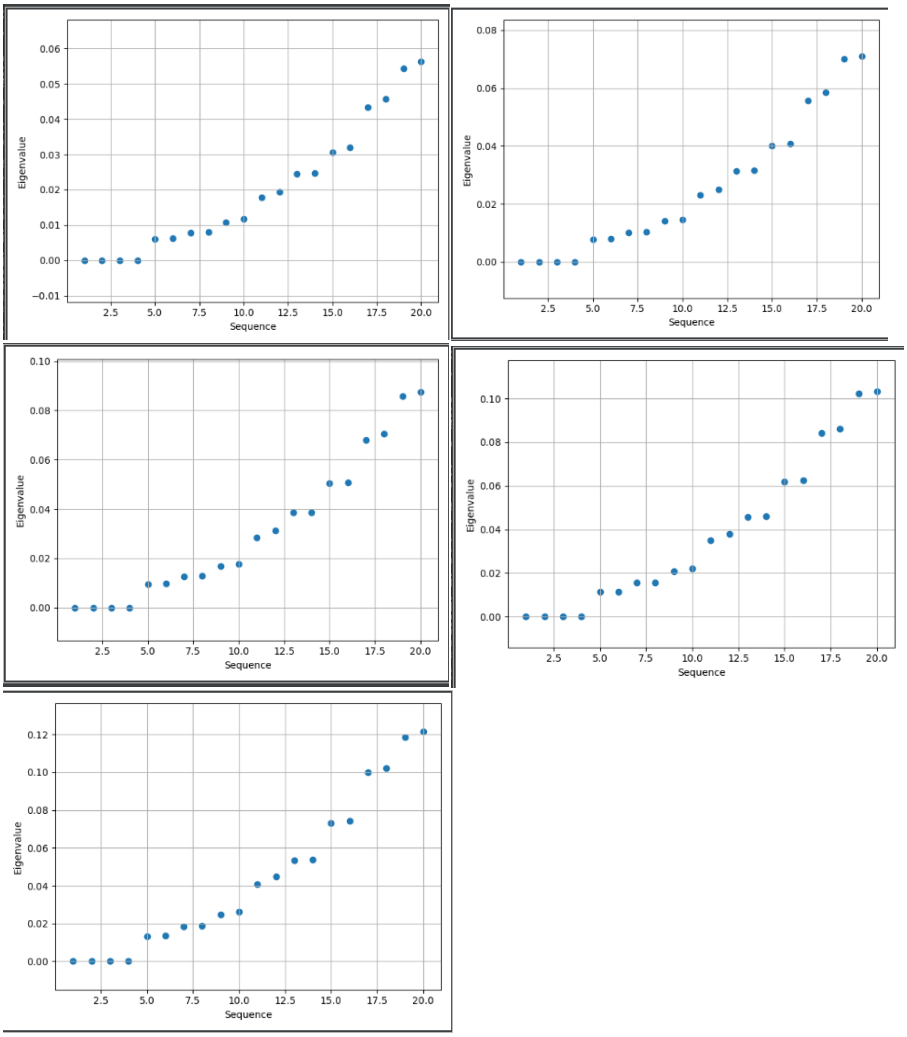
h) (5 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifiers to control the color scheme. Please comment on this K-mean result.



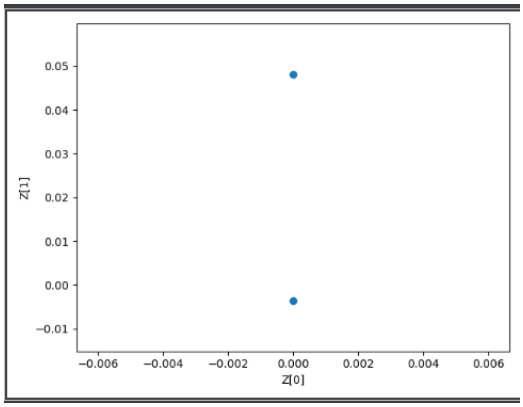
This plot is the result by using four clusters as the number of clusters while applying K-mean algorithm. It is not the result we want obviously.

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.


Eigenvalue
Sequence



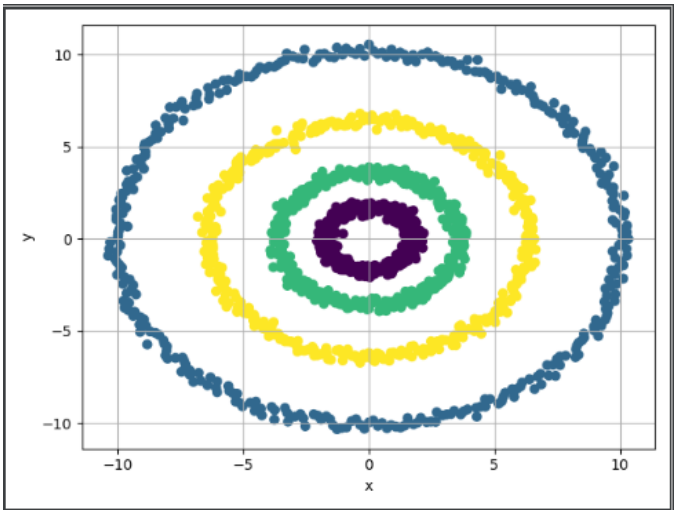
j) (5 points) Using your choice of the number of neighbors in (c), calculate the Adjacency matrix, the Degree matrix, and finally the Laplacian matrix. How many eigenvalues do you determine are practically zero? Please display their calculated values in scientific notation



```
Eigenvalue:  0
              Mean =  0.017287002390780886
Standard Deviation =  0.01988979619768845
  Coeff. Variation =  1.1505636285615155
Eigenvalue:  1
              Mean =  0.01431797080461654
Standard Deviation =  0.022123294431042398
  Coeff. Variation =  1.5451417475938134
Eigenvalue:  2
              Mean =  0.013176156917373177
Standard Deviation =  0.02282177322937909
  Coeff. Variation =  1.7320508075680143
Eigenvalue:  3
              Mean =  0.004121722177984368
Standard Deviation =  0.026027982071838724
  Coeff. Variation =  6.314831749423514
Eigenvalue:  4
              Mean =  2.09770472597231e-15
Standard Deviation =  0.026352313834736504
  Coeff. Variation =  12562451477779.795
Eigenvalue:  5
              Mean =  0.0
Standard Deviation =  0.02635231383473648
  Coeff. Variation =  inf
Eigenvalue:  6
              Mean =  2.220446049250313e-16
Standard Deviation =  0.026352313834736494
  Coeff. Variation =  118680270766469.64
```

```
Eigenvalue:  7
              Mean =  -6.050715484207103e-16
Standard Deviation =  0.026352313834736487
  Coeff. Variation =  -43552392941823.71
Eigenvalue:  8
              Mean =  -1.8257000849391462e-16
Standard Deviation =  0.026352313834736487
  Coeff. Variation =  -144340869851111.7
Eigenvalue:  9
              Mean =  -4.1078251911130794e-16
Standard Deviation =  0.026352313834736518
  Coeff. Variation =  -64151497711605.266
```

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.



By apply K-mean algorithm on these two eigenvectors to find a four clusters solution, the result is absolutely same with our expectation.