Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: March 1, 2023(100 points)

YOUR NAME
YOUR GIT USERNAME
YOUR E-MAIL

K-Means

The normalized automobile distributor timing speed and ignition coil gaps for production F-150 trucks over the years of 1996, 1999, 2006, 2015, and 2022. We have stripped out the labels for the five years of data.

Each sample in the dataset is two-dimensional, i.e. $\mathbf{x}_i \in \mathbb{R}^2$ (one dimension for timing speed and the other for coil gaps), and there are $N = 5000$ instances in the data.

Question 1 [20 pts total]

[10 pts] Question 1a.) Implement a simple k-means algorithm in Python on Colab with the following initialization:

$$\mathbf{x}_1 = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} -10 \\ -10 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} -3 \\ -3 \end{pmatrix},$$

You need only 100 iterations, maximum, and your algorithm should run very quickly to get the results.

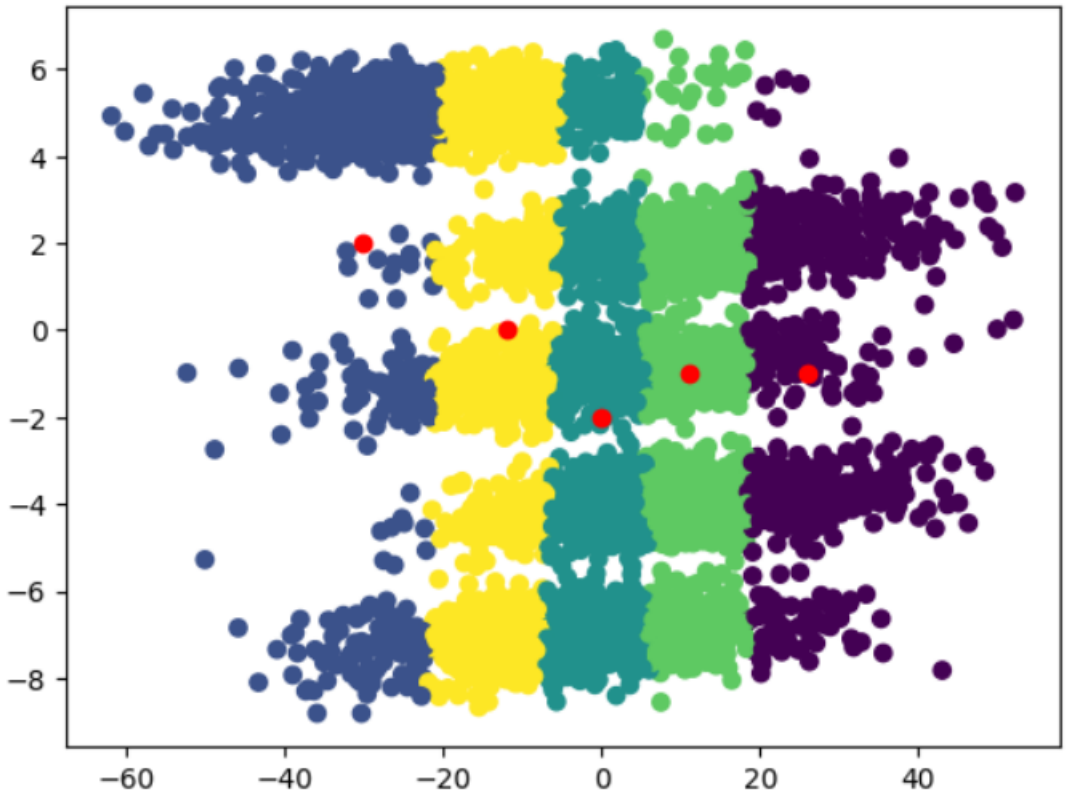
```
def my_kmeans(xs: np.ndarray, init_centers: np.ndarray, n_iter=100):
    N,D = xs.shape
    K = init_centers.shape[0]
    final_centers = init_centers
    for it in range(n_iter):
        dist = cdist(xs,final_centers)
```

```
        cluster_id = np.argmin(dist,axis=1)
        for i in range(K):
            final_centers[i,:] = np.mean(xs[cluster_id==i],axis=0)

    dist = cdist(xs,final_centers)
    cluster_id = np.argmin(dist,axis=1)
    return final_centers,cluster_id
```

[5 pts] Question 1b.) Scatter the results in two dimensions with different clusters as different colors. You can use **matplotlib**'s **pyplot** functionality:

```
import matplotlib.pyplot as plt
data = np.array([[10,10],[-10,-10],[2,2],[3,3],[-3,-3]])
centroids,cluster_id = my_kmeans(x,data)
plt.scatter(x[:,0],x[:,1],c=cluster_id)
plt.scatter(centroids[:,0],centroids[:,1],c='r')
plt.show()
```



[5 pts] Question 1c.) You will notice that in the above, there are only five initialization clusters. Why is $k = 5$ a logical choice for this dataset? After plotting your resulting clusters, what do you notice? Did it cluster very well? Is there an initialization that would make it cluster well?

From the data scatter plot, we can see that the data's distribution are five short strips and each strip is a cluster. Therefore, we choose k=5 as the clustering number.
When we draw the resulting clusters, we can see that the result breaks the original distribution of data and it cuts each strip into five parts. Therefore, it doesn't cluster well.

There doesn't exist an initialization that would make it cluster well. Kmeans can't do well in such shape of data.

Question 2)[30 pts total]

In the data from Question 1, let $\mathbf{x}$ and $\mathbf{y}$ be two instances, i.e., they are each truck with separate measurements. A common distance metric is the *Mahalanobis Distance* with a specialized matrix $P \in \mathbb{R}^{2\times 2}$ that is written as follows:

$$d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T P^{-1} (\mathbf{x} - \mathbf{y})$$

In scalar format (non-matrix format), the Mahalanobis Distance can be expressed as:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{2} \sum_{j=1}^{2} (x_i - y_i) \cdot P_{i,j}^{-1} \cdot (x_j - y_j)$$

where $\mathbf{x}$ and $\mathbf{y}$ are two instances of dimensionality 2, and $d(\mathbf{x}, \mathbf{y})$ is the distance between them. In the case of the F150 engine components, P is a known relationship through Ford's quality control analysis each year, where it is numerically shown as below:

$$P = \begin{pmatrix} 10 & 0.5 \\ -10 & 0.25 \end{pmatrix}$$

[15 pts] Question 2a.) Using the same data as **Question 1** and the same initialization instances $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ implement a specialized k-means with the above Mahalanobis Distance. Scatter the results with the different clusters as different colors.

What do you notice? You may want to pre-compute P^{-1} so that you aren't calculating an inverse every single loop of the the k-Means algorithm.

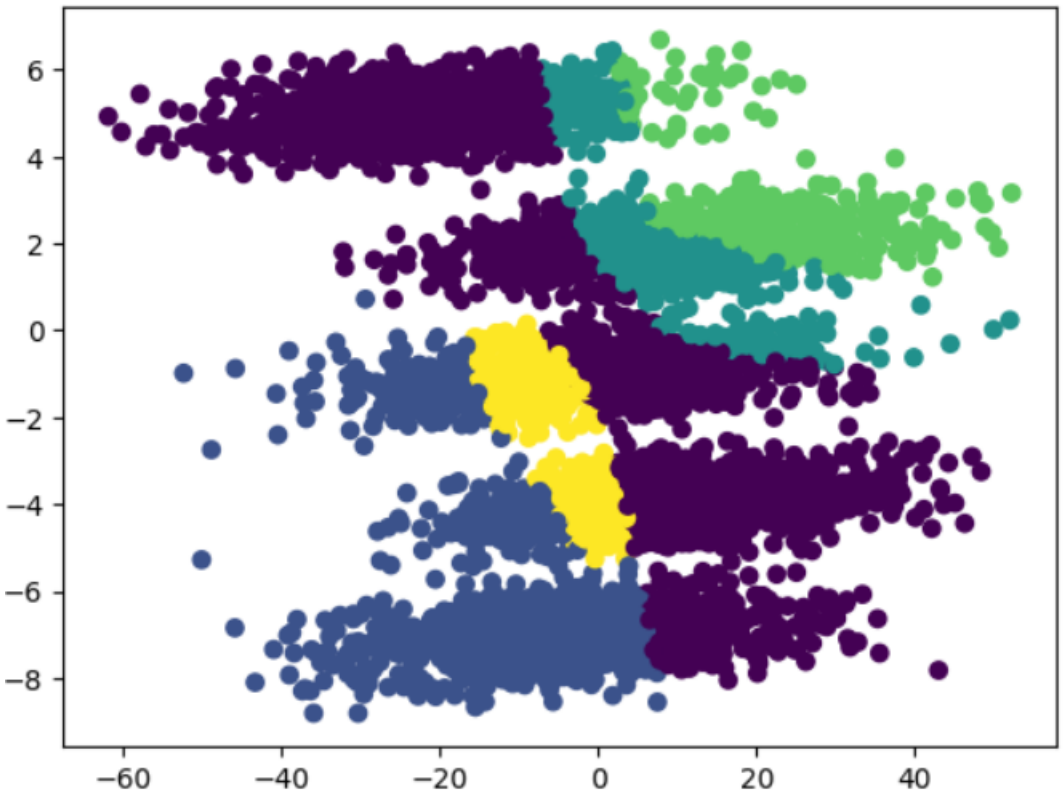
```
def new_kmeans(xs, init_centers, VI, n_iter=100):
    N,D = xs.shape
    K = init_centers.shape[0]
    final_centers = init_centers

    for it in range(n_iter):
        dist = cdist(xs,final_centers,metric='mahalanobis', VI=VI)
        cluster_id = np.argmin(dist,axis=1)

        for i in range(K):
            final_centers[i,:] = np.mean(xs[cluster_id==i],axis=0)

    dist = cdist(xs,final_centers,metric='mahalanobis', VI=VI)
    cluster_id = np.argmin(dist,axis=1)
    return final_centers,cluster_id
p = np.array([[10,0.5],[-10,0.25]])
```

```
inv_p = np.linalg.inv(p)
data = np.array([[10,10],[-10,-10],[2,2],[3,3],[-3,-3]])
new_centroids,new_cluster_id = new_kmeans(x,data,VI=inv_p)
```



[5 pts] Question 2b.) Calculate and print out the principle components of the aggregate data.

1st pc is [-0.99838317, 0.05684225],
2nd pc is [-0.05684225, -0.99838317]

[5 pts] Question 2c.) Calculate and print out the principle components of *each cluster*. Are they the same as the aggregate data? Are they the same as each other?

For cluster 0,the compenents are:[[-0.98439765, 0.17595815] [-0.17595815, -0.98439765]]
For cluster 1,the compenents are:[[-0.9966809, 0.08140747] [0.08140747, 0.9966809]]
For cluster 2,the compenents are:[[0.99089241, -0.13465595] [0.13465595, 0.99089241]]
For cluster 3,the compenents are:[[0.99902772, -0.04408651] [0.04408651, 0.99902772]]
For cluster 4,the compenents are:[[-0.97099153, 0.23911388] [0.23911388, 0.97099153]]

They are not same as the aggregate data. And they are also different from each other.

[5 pts] Question 2d.) Take the eigenvector / eigenvalue decomposition of P^T and subsequently, take their product. That is to say,

$$\{\Lambda, \Phi\} = \texttt{eig}\left(P^T\right)$$

where $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ and Φ is a 2×2 matrix with $\phi_i \in \mathbb{R}^2$, a column in Φ. Calculate a new P'

such that

$$P' = \Lambda\Phi$$

What is the relationship between P' and the data?
Throug coding, we obtain $\Lambda = \begin{pmatrix} 9.45693086 & 0 \\ 0 & 0.79306914 \end{pmatrix}$. The new $P' = \begin{pmatrix} 9.44301625 & 6.95724558 \\ 0.04300577 & 0.53717161 \end{pmatrix}$

Market Basket Analysis and Algorithms

Consider F_3 as the following set of frequent 3-itemsets:

```
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4},
{2, 3, 4}, {2, 3, 5}, {3, 4, 5}.
```

Assume that there are only five items in the data set.

Question 3 [25 pts total]

[10 pts] Question 3a.) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

```
{1,2,3,4},
{1,2,3,5},
{1,2,4,5},
{1,3,4,5},
{2,3,4,5}
```

[10 pts] Question 3b.) List all candidate 4-itemsets obtained by the candidate generation procedure in A Priori, using $F_{k-1} \times F_{k-1}$.

```
{1,2,3,4},
{1,2,3,5},
{1,2,4,5},
{2,3,4,5}
```

[5 pts] Question 3c.) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

{1,2,3,5}'s subset {1,3,5} is not in F_3, therefore, it will be pruned.
{1,2,4,5}'s subset {1,4,5},{2,4,5} are not in F_3, therefore, it will be pruned.
{2,3,4,5}'s subset {2,4,5} is not in F_3, therefore, it will be pruned.
Therefore, the survived candidate is {1,2,3,4}.

Question 4 [25 pts total]

Consider the following table for questions 4a) to 4c):

Transaction ID	Items
1	**{Beer, Diapers}**
2	**{Milk, Diapers, Bread, Butter}**
3	**{Milk, Diapers, Cookies}**
4	**{Bread, Butter, Cookies}**
5	**{Milk, Beer, Diapers, Eggs}**
6	**{Beer, Cookies, Diapers}**
7	**{Milk, Diapers, Bread, Butter}**
8	**{Bread, Butter, Diapers}**
9	**{Bread, Butter, Milk}**
10	**{Beer, Butter, Cookies}**

[3 pts] Question 4a.) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?
There are 7 unique items. Therefore, the maximum number of association rules is $3^7 - 2^8 + 1 = 1932$

[3 pts] Question 4b.) What is the confidence of the rule {Milk, Diapers} ⇒ {Butter}?

$$confidence = \frac{\sigma(\{Milk, Diaper, Butter\})}{\sigma(\{Milk, Diaper\})} = \frac{2}{4} = 0.5$$

[3 pts] Question 4c.) What is the support for the rule {Milk, Diapers} ⇒ {Butter}?

$$support = \frac{\sigma(\{Milk, Diapers, Butter\})}{|T|} = \frac{2}{10} = 0.5$$

[3 pts] Question 4d.) `True` or `False` with an explanation: Given that {a,b,c,d} is a frequent itemset, {a,b} is always a frequent itemset.
True.
$\{a, b\} \in \{a, b, c, d\}$, that means $support(\{a, b\}) \geq support(\{a, b, c, d\})$. Therefore, when $\{a, b, c, d\}$ is a frequent itemset, $\{a, b\}$ is also a frequent itemset.

[3 pts] Question 4e.) `True` or `False` with an explanation: Given that {a,b}, {b,c} and {a,c} are frequent itemsets, {a,b,c} is always frequent.
False.
For example, when {a,b}, {b,c} and {a,c} are frequent itemsets, the itemsets {a,b,c} may not occur in the transactions. At this time, its support is 0. Certainly, it's not a frequent itemset.

[3 pts] Question 4f.) `True` or `False` with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.
False.
The support of {b} is at least larger than $20 + 30 = 50$.

[3 pts] Question 4g.) `True` or `False` with an explanation: In a dataset that has 5 items, the maximum number of size-2 frequent itemsets that can be extracted (assuming minsup > 0) is 20.
False.

The maximum number is $\binom{5}{2} = 10$

[4 pts] Question 4h.) Draw the itemset lattice for the set of unique items $\mathcal{I} = \{a, b, c\}$.