CS 6220 Data Mining — Assignment 4

Due: March 1, 2023(100 points)

Yuze Zhang
humorson2021
zhang.yuze1@northeastern.edu

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.

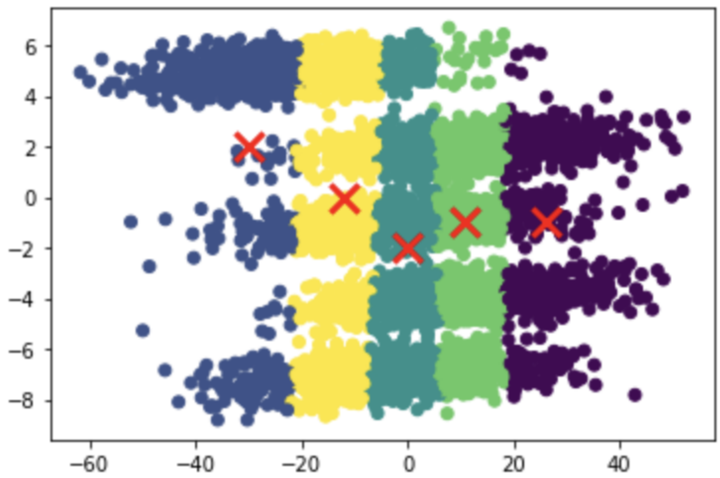
Final 5 centroids: $[26, -1], [-30, 2], [0, -2], [11, -1], [-12, 0]$
[5 pts] Question 1b.) Scatter the results in two dimensions with different clusters as different colors. You can use **matplotlib**'s **pyplot** functionality:
See ipynb file.

```
>> import matplotlib.pyplot as plt
>> plt.scatter(data.values[:, 0], data.values[:, 1], c=labels)
>> plt.scatter(fcentroids[:, 0], fcentroids[:, 1], marker='x', s=200, linewidths
>> plt.show()
```

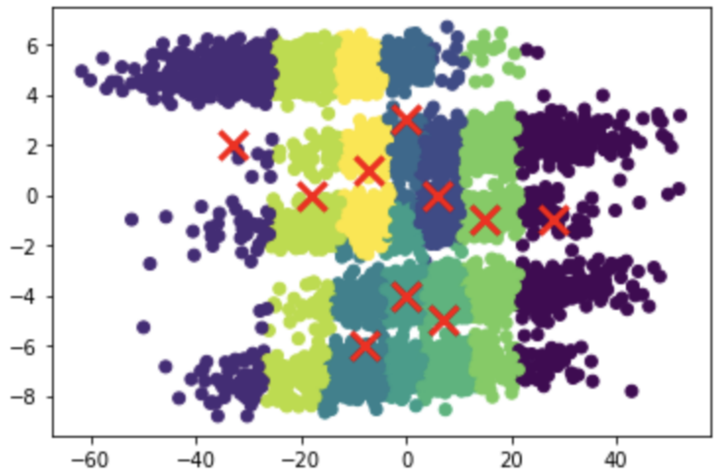
[5 pts] Question 1c.) You will notice that in the above, there are only five initialization clusters. Why is $k = 5$ a logical choice for this dataset? After plotting your resulting clusters, what do you notice? Did it cluster very well? Is there an initialization that would make it cluster well?

It is proper for getting final centroids. If we select k = 6, then the new centroid will occupy part of one of first five centroids. If even more, clusters will be in mess. Now, there are five parallel clusters and it looks well.

k = 5 :



k = 10 :



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:

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

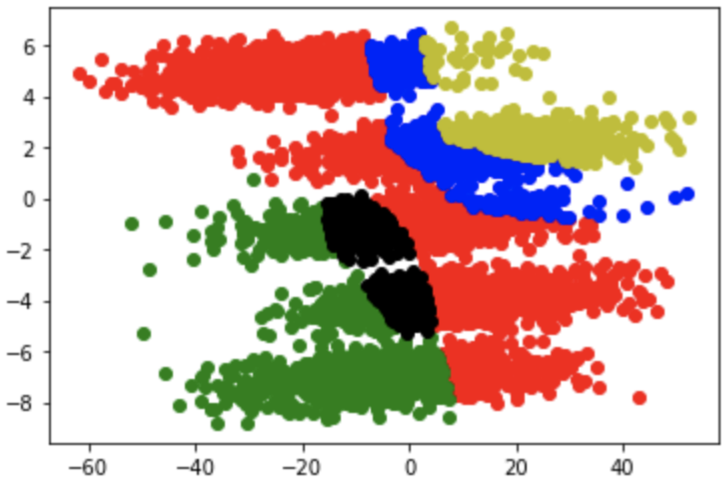
where $\mathbf{x}$ and $\mathbf{y}$ are two instances of dimensionality 2, and $d(\mathbf{x}, \mathbf{y})$ is the distance between them. In the case of the F150 engine components, P is a known relationship through Ford's quality control analysis each year, where it is numerically shown as below:

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

[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.

$$P^{(} - 1) = \begin{pmatrix} 1/30 & -1/15 \\ 4/3 & 4/3 \end{pmatrix}$$



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

$Eigenvalues$: [322.5071327317.38845582]

$Eigenvectors$: [[0.998383170.05684225][−0.056842250.99838317]]

Thus, the first eigenvector corresponds to the principal component with the highest variance and so on.

[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?

They are different from aggregate data and 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?
P' can be used to transform the data such that it is aligned with its principal axes and has unit variance along each of those axes.

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.

For $F_3 \times F_1$:

$\{1, 2, 3\} \times \{4\} = \{1, 2, 3, 4\}$

$\{1,2,3\}\times\{5\}=\{1,2,3,5\}$

$\{1,2,4\}\times\{5\}=\{1,2,4,5\}$

$\{2,3,4\}\times\{5\}=\{2,3,4,5\}$

All candidate 4-itemsets are $\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{2,3,4,5\}$.

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

For $F_3 \times F_3$, merge only if first 2 items are identical:

$\{1,2,3\}\times\{1,2,4\}=\{1,2,3,4\}$

$\{1,2,3\}\times\{1,2,5\}=\{1,2,3,5\}$

$\{1,2,4\}\times\{1,2,5\}=\{1,2,4,5\}$

$\{2,3,4\}\times\{2,3,5\}=\{2,3,4,5\}$

All candidate 4-itemsets are $\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{2,3,4,5\}$.

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

Then,

For $\{1,2,3,4\}$: $\{1,2,3\}, \{1,2,4\}, \{2,3,4\}$ are all frequent

For $\{1,2,3,5\}$: $\{1,2,3\}, \{1,2,5\}, \{2,3,5\}$ are all frequent

For $\{1,2,4,5\}$: $\{1,2,4\}, \{1,2,5\}$ are frequent, but $\{2,4,5\}$ is not frequent

For $\{2,3,4,5\}$: $\{2,3,4\}, \{2,3,5\}\{3,4,5\}$ are all frequent

Therefore, the final list of candidate 4-itemsets are $\{1,2,3,4\}, \{1,2,3,5\}, \{2,3,4,5\}$.

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)?

The set of items: $\{Beer, Diapers, Milk, Bread, Butter, Cookies, Eggs\}$ have 7 items

Then maximum number of association rules will be

$$3^7 - 2^{7+1} + 1 = 1932 \tag{0.1}$$

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

confidence of {Milk, Diapers} ⇒ {Butter} = $\textit{support}\{Milk, Diapers, Butter\}/\textit{support}\{Milk, Diapers\}$ $= 2/4 = 0.5$

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

It is same as support of $\{Milk, Diapers, Butter\}$/total transactions = 2 / 10 = 0.2

[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} is subset of {a,b,c,d} and any subset of an itemset must appear in the dataset at least as frequently as the original itemset. Thus, the frequency of {a,b} exceeds threshold and it is always 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.

True. Because {a,b}, {b,c}, and {a,c} are subsets of {a,b,c} and any transaction that contains a,b,c always contain these three frequent itemsets. Thus, the support of {a,b,c} is at least same as these subsets which means it is also 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. It could be 20 or 30 also. If all transactions that contain {a,b} also contain {b,c}, then it will be 20.

[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. It is choose 2 from 5 which is $5!/2!(5-2)! = 10$.

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