Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: March 1, 2023(100 points)

Wanqing Wang
wqwang-cerealk
wang.wa@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.

[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
>> plt.scatter(<YOUR CODE HERE>)
```

[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?

I think k = 5 is logical because as we can see from the above graph, the points are assigned to clusters ranges from left to right. The leftmost cluster is the purple one, and the rightmost is the light blue one. The x-axis ranges from -60 to 40, which can be divided into 5 portions.

I noticed that the data is not clustered well since there are obviously two larger clusters while the middle three clusters are smaller.

I think a more logical way to initialize centroids are to assign centroids more evenly across the quandrant. Currently, the distance between x3, x4, and x5 are too close, which causes the middle three clusters' areas are smaller. Also, we can move x1 and x2 to (20, 20) and (-20 -20) which matches the range of the data points more than current initialization.

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:

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

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.

I notice that the clusters are totally different than using Euclidean distance. And the clusters look more decent and reasonable since in Euclidean distance, the x-axis unit is larger than unit in y-axis, which causes inaccurate clusters and Mahalanobis Distance makes more sense in this situation.

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

[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 not the same as aggregate data. Although they are not exactly the same as each other, they look pretty close in this case, especially for first three clusters.

[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' is closed to cluster principal component.

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, 4}`

In itemset {1, 2, 3, 5}, the subset {1, 3, 5} does not exist in the original given F3.

In itemset {1, 2, 4, 5}, the subset {1, 4, 5} does not exist in the original given F3.

In itemset {2, 3, 4, 5}, the subset {2, 4, 5} does not exist in the original given F3.

So all of them are pruned. The left one is itemset {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)?

Based on formula, there are total $3^7 - 2^8 + 1 = 1932$ can be generated.

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

Confidence = frequency of {Milk, Diapers, Butter} / frequency of {Milk, Diapers} = 2 / 4 = 0.5

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

Support = frequency of {Milk, Diapers, Butter} / $|T|$ = 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. Because if an itemset is frequent, then all of its subsets must also be frequent (anti-monotone property of support). {a,b} is {a, b, c, d}'s subset, so if {a, b, c, d} is a frequent itemset then {a, b} must be frequent.

[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, in this case, if we do not have a min threshold then we assume {a, b}, {b, c} and {a, c} are frequent itemsets in F2 itemset. Then by combining pair {a, b} and {a, c} or combining pair {a, b} and {b, c}, we can get frequent itemset {a, b, c} in F3 itemset.

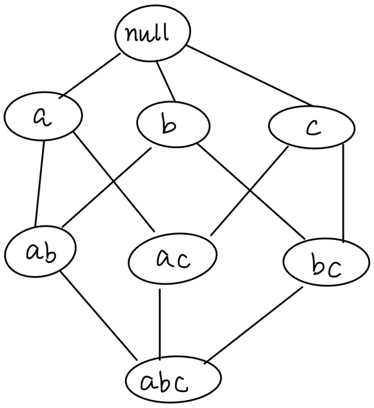
[3 pts] Question 4f.) `True` or `False` with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

False. Given that the support of {a, b} is 20, this means that itemset {a, b} appears 20 and thus itemset {b} appears at least 20. Meanwhile, itemset {b, c} is 30 means that {b} appears at least 30. Therefore, the support of {b} is larger or equal to 30.

[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 of size-2 frequent itemsets is ${}_5C_2 = 10$.

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


null
a
b
c
ab
ac
bc
abc