Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: March 1, 2023(100 points)

Wu Chenjie
@suiboli314
wu.chenj@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?

k=5 is a logical choice for the dataset since the data is actually gathered in 5 different places, or say clusters, as we may visually observe in the plot.
But the centroids of dataset are not in the center of cluster after calculatedb by k-means algorithm in this case. If the data is normalized or standardized, or the initialization is spread out at -8, -4, -2, 2, 6 in y axis, and small covariance in x axis.

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:

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

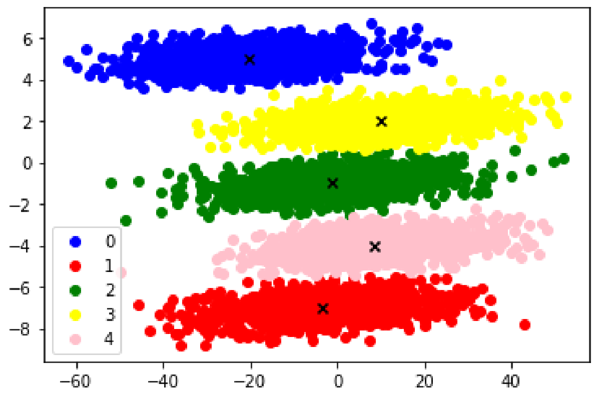
In scalar format (non-matrix format), the Mahalanobis Distance can be expressed as:

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

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 $(P^TP)^{-1}$ so that you aren't calculating an inverse every single loop of the the k-Means algorithm.

K-means is able to find the centroid of cluster successfully.

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

$eigval[322.50713273, 17.38845582]$
$eigvec[[0.99838317, 0.05684225][-0.05684225, 0.99838317]]$

[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?

$eigval[195.0259165, 0.2722047]$
$eigvec[[0.99993527, -0.01137789][0.01137789, 0.99993527]]$
$eigval[204.376807, 0.28339306]$
$eigvec[[0.99992533, -0.01222027][0.01222027, 0.99992533]]$
$eigval[217.34942629, 0.27670722]$
$eigvec[[0.99990986, -0.01342629][0.01342629, 0.99990986]]$
$eigval[204.42164874, 0.26977027]$
$eigvec[[0.99993306, -0.01157047][0.01157047, 0.99993306]]$
$eigval[191.5356272, 0.26355151]$
$eigvec[[0.99989374, -0.01457781][0.01457781, 0.99989374]]$
Eigen vectors are surprisingly similar to each other. But they are not the same as aggregate data.

[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? The relationship between P' and the data the eigenvectors of np.linalg.eig(P.T @ P) are a square matrix similar to the cluster PCA components in Q2c, meaning that p' is a good projection that seperates out our data

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.

To obtain the candidate 4-itemsets using the $F_{k-1} \times F_1$ merging strategy, we need to join the frequent 3-itemsets F_3 with the frequent 1-itemsets F_1

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

that contain an item not in the frequent 3-itemsets.

The generated 4-itemsets are

```
{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 $F_{k-1} \times F_{k-1}$.

To merge the 3-itemsets together with whose have identical 2-tiemsets.
The generated 4-itemsets are

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

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

are not in 3-itemsets, so 4-itemsets are

```
{1, 2, 3, 4}, {1, 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)?

There are 7 distinct items in the data set.
{Beer, Diapers, Milk, Bread, Butter, Cookies, Eggs}
For each item, it can either be included or excluded from the antecedent (left-hand side) of a rule, or from the consequent (right-hand side) of a rule. There are 2 choices for each of the 7 items, giving us a total of 2^7 possible combinations of items. However, we need to exclude the empty set and the full set, since they are not interesting association rules. So the total number of possible association rules is: $2^7 - 2 = 126$

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

Freq{Milk, Diapers} = 4
Freq{Milk, Diapers, Butter} = 2
confidence{Milk, Diapers, Butter} = 2/4 = 0.5

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

Support{Milk, Diapers, Butter} = 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`
If a, b, c, d is a frequent itemset, then by definition it occurs in at least the minimum number of transactions required to be considered frequent. Any subset of a, b, c, d, including a, b, must occur at least as many times as the superset, since it represents a subset of those transactions. Therefore, 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`
Since {a,b,c} might not be frequent due to the minimal support count does not reach.

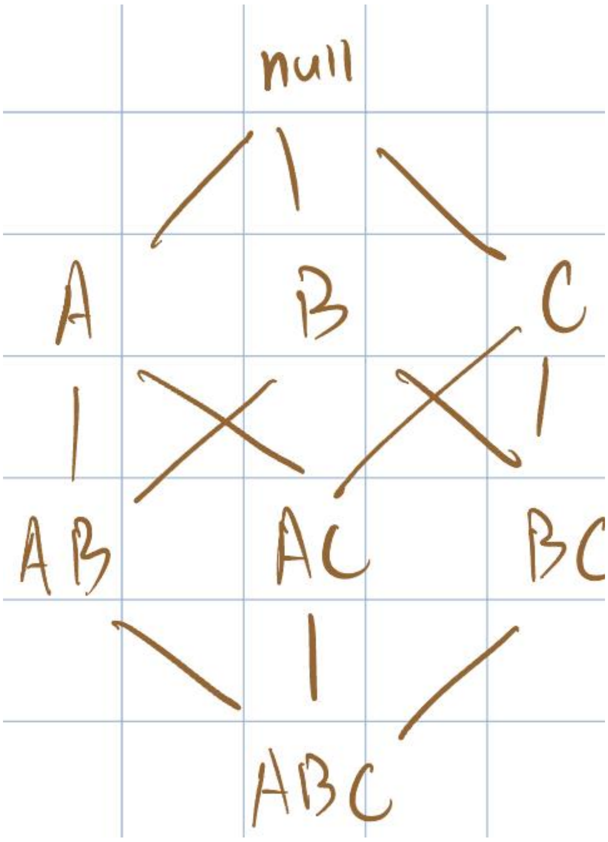
[3 pts] Question 4f.) `True` or `False` with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

`False`
As anti-monotone property of support, If itemset is frequent, then all of its subsets must also be as least frequent as it. So, the support of {b} is larger than 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`
5 choose 2, C(5, 2), is 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