Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 4

Due: March 1, 2023 (100 points)

Suisui Xia
SuisuiXia
xia.su@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. $x_i \in \mathbb{R}^2$(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] ***(Please see .ipynb file for this problem)***

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

$$x_1 = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, x_2 = \begin{pmatrix} -10 \\ -10 \end{pmatrix}, x_3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, x_4 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, x_5 = \begin{pmatrix} -3 \\ -3 \end{pmatrix},$$

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?

Question 2 [30 pts total] ***(Please see .ipynb file for this problem)***

In the data from Question 1, let x and 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:

$$R = \left(P^T P\right)^{-1}$$

$$d(x, y) = (x - y)^T R(x - y)$$

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

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

where x and y are two instances of dimensionality 2, and d(x, 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 $\{x_1, x_2, x_3, x_4, x_5\}$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 k-Means algorithm.

[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?

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

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

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

$$P' = \Lambda\Phi$$

What is the relationship between P' and the 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.

$F1 = \{1\},\{2\},\{3\},\{4\},\{5\}$, it is all frequent itemset of size 1

$F3$ = $\{1,2,3\}$, $\{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}$, it is all frequent itemset of size 3 given by problem

$F4$ candidates can be generated by $F3 \times F1$ so we can have all the 3-itemsets in F3 merge with a frequent item in F1:

By using this $F_{k-1} \times F_1$ mering strategy, we can get result as below:

$$F4\ candidates = \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{1,3,4,5,\}, \{2,3,4,5\}$$

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

In order to get all candidate 4-itemsets by using $F_{k-1} \times F_{k-1}$, we will need to merge every two frequent 3-itemsets only if their first k-1 items are same.

$$F3 = \{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}$$

For example, we will merge {1,2,3} with {1,2,4} to get {1,2,3,4}, since the first k-1 items are same, {1,2} is same as {1,2}

$$Merge(\{1,2,3\}\, With\, \{1,2,4\}) \rightarrow \{1,2,3,4\}$$

$$Merge(\{1,2,3\}\, With\, \{1,2,5\}) \rightarrow \{1,2,3,5\}$$

$$Merge(\{1,2,4\}\, With\, \{1,2,5\}) \rightarrow \{1,2,4,5\}$$

$$Merge(\{2,3,4\}\, With\, \{2,3,5\}) \rightarrow \{2,3,4,5\}$$

With this method, we can get result as below:

$$F4\ candidates = \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{2,3,4,5\}$$

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

By using Apriori algorithm we get the F4 candidates as below:

$$F4\ candidates = \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{2,3,4,5\}$$

And F3 is given:

$F3 = \{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{2,3,4\},\{2,3,5\},\{3,4,5\}$

In order to survive from candidate pruning, we will need to make sure its subset of size k-1 is frequent, otherwise we will prune it. For example, given candidate 4-itemsets, we will need to make sure its subset with size 3 are also frequent(means appearing in F3)

The subsets of size 3 for $\{1,2,3,4\}$ are $\{1,2,3\}$, $\{1,2,4\}$, $\{2,3,4\}$, $\{1,3,4\}$, and these subsets are all frequent since appear in F3.

since $\{1,2,3,5\}$ contains $\{1,3,5\}$ which is infrequent (not appear in F3), we will need to prune it.

Same as above example, $\{1,2,4,5\}\ and\ \{2,3,4,5\}$ contains $\{2,4,5\}$, which is infrequent too. So it will not be survive.

After pruning step, we will get the result as below:

$F4 = \{1,2,3,4\}$ survived

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 maximum numbers of association rules can be extracted from this data by using formula(show below), which i indicates the unique items from data set, and in this table we have 7 unique items, they are beer, diapers, milk, bread, butter, cookies, eggs

$= 3^i - 2^{i+1} + 1$

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

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

To find out the confidence of {Milk, Diapers} ⇒ {Butter}, we need to use this formula:

$$confidence = \frac{\sigma(\{Milk, Diapers, Butter\})}{\sigma(\{Milk,\ Diapers\})}$$

{Milk, Diapers} appear in transactions 4 times (transactions 2,3,5,7)

The transactions contain {Milk, Diapers} appear 4 times in transaction 2,3,5,7

Within these transactions 2,3,5,7, and also contains {Butter} appear 2 times in transaction 2,7.

So we can calculate using formula above:

$Confidence = 2 / 4 = 0.5$

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

To find out the support of {Milk, Diapers} ⇒ {Butter}, we need to use this formula:

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

Numbers of transactions that contains {Milk, Diapers, Butter} appear 2 times with transaction 2,7

T is Total of transactions, in this data table it will be 10

So we can calculate using formula above:

$Support = 2 / 10 = 0.2$

[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 its subsets will always be 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. If $\{a, b\}$,$\{b, c\}$ and $\{a, c\}$ are frequent, $\{a, b, c\}$ is not necessarily be frequent. $\{a, b, c\}$ is frequent or not depends its support count in dataset

[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 $\{b\}$ means how often b appears in transactions. From question we will know b appear in {a,b} 20 transactions, in {b,c} 30 transactions, so we will know b appearing at least in 30 transactions, so statement is false

[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. If the dataset that has 5 items, and we want to get all the size-2 frequent itemsets, which means 5 choose 2, equals 10

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