Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: March 1, 2023(100 points)

Weiran Guo
panunburn
guo.weir@northeastern.edu

K-Means

Question 1 [20 pts total]

[10 pts] Question 1a.)
Please see ipynb file.

[5 pts] Question 1b.)
Please see ipynb file.

[5 pts] Question 1c.)
We choose $k = 5$ for this data as there are five production years, we want to cluster to that specific production years. The scatter plot shows that for clustering jobs, it does not cluster well, as we can see on the plot, the data might be clustered on horizontal ways, but the scattered result is based on vertical clustering. Yes, we might try to initialize with different centroids, so that the clustering will be horizontal. We can try random, or by setting default value based on the data graph.

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:

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

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

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

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

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.) Please see ipynb file.
Now, we can see that data are clustered in horizontal ways, the by changing the distance measuring algorithm, we can get more precise results, even with same initialization instances centroids.
[5 pts] Question 2b.) Please see ipynb file.

[5 pts] Question 2c.) Please see ipynb file. We can see that for cluster they have different pincipal components of aggregated data. As each cluster represent a part of data, and principle components capture the pattern on that subset of data. They are not same as each other.

[5 pts] Question 2d.) Please see ipynb file.
P' is a new matrix that is obtained by transforming the original data with the eigenvectors and eigenvalues of P^T. The relationship between P' and the data is that P' is a new representation of the data in a coordinate system that is aligned with the principal axes of variation in the data. P' can capture the most important features of the data in a more concise and informative way.

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.

The frequent 1-itemsets are:

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

Taking the cross product of frequent 3-itemsets with frequent 1-itemsets, we get the following candidate 4-itemsets:

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

By using $F_{k-1} \times F_{k-1}$, we merge only if first k-2 items are identical.
we have

`{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.
After pruning, we see that

`{1, 3, 5}`

is a subset of

`{1, 2, 3, 5}`

and

`{2, 4, 5}`

is a subset of

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

but are not in 3-itemsets, so final result 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)?
By equation

$$R = 3^n - 2^{n+1} + 1$$

we have,

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

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

$$confidence = \frac{sup\{\text{Milk, Diapers, Butter}\}}{sup\{\text{Milk, Diapers}\}} = \frac{2}{4} = 0.5$$

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

$$confidence = \frac{sup\{\text{Milk, Diapers, Butter}\}}{total} = \frac{2}{10} = 0.2$$

[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, as {a,b} is a subset of a,b,c,d}, a subset of a frequent itemset is always 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.

False, we can generate {a,b,c} candidate itemset, but we should look at data to determine whether it is frequent.

[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.

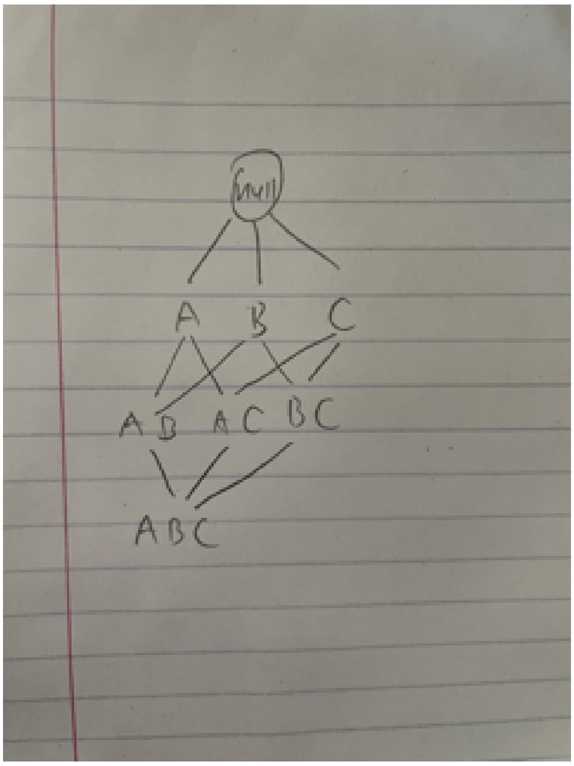
False. the support can be larger than 30, simply let itemsets be union of 20 {a,b}s and 30 {b,c}s, then there are total 50 bs.

[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. By combination, we have 5 choose 2, which is

$$\frac{5 * 4}{1 * 2} = 10$$

[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