Database Management System- CS-2376

A
PROJECT
REPORT
ON

K-MEANS CLUSTERING ALGORITHM

Submitted to
Prof. Anirban Mondal

Submitted by:
Manish Yadav
Aman Antil
Deepak Chauraisya

Department of Computer Science
Ashoka University, Sonipat

Objective:

Implement the k-means clustering algorithm from scratch.

Introduction:

Clustering is one of the techniques of observing the structure of the data. K-means is one of the most used methods of clustering algorithms. What k-means algorithm does is it tries to make partitions of the dataset into n-overlapping subgroups(clusters) in which each data point belongs to a particular group.

Input:

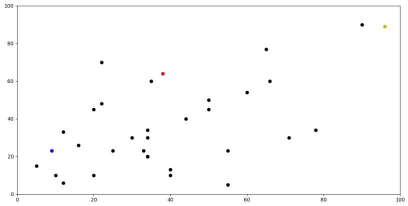
A list of 2 dimensional dataset. It either can be in csv format or we can manually give x,y coordinates in a list format.

Output:

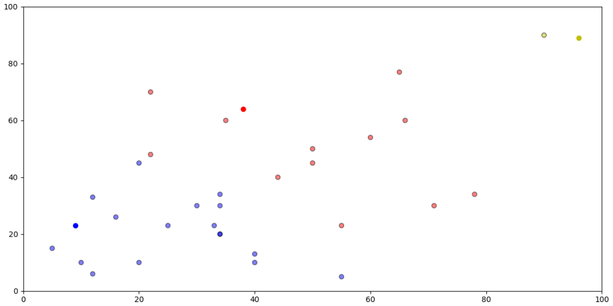
Coloured cluster of points in 2 dimensional plane.

Experiment:

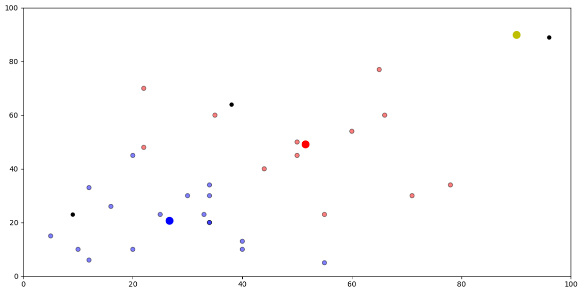
For this assignment, we have created a 2-dimensional dataset in csv format. It can also be generated in excel sheets. We took three random centroids and visualized it on the graph with other points available in the csv file. All cetroids have their specific colors. Here is how the graph looks like:



Three colored points are our centroids and remaining points are values in the csv file. We then calculate the distance of each point from these three centroids and cluster them based on their distances from these centroids. We change the color of black points into the color of the centroid which is closest to it. Here is the result:



We have clustered the given data. Now, we update the centroids into the average of their cluster's points. For instance, we have a centroid and there are three points P1, P2 , P3 into its cluster. So, we find the average of these three points and that average would be our new centroid. The graph given below shows the new positions of our three centroids which are in the same color but we only increased their sizes. Three black colored points represent their older position.



We repeat the algorithm with these new centroid and clustered data until we get a constant cluster which will not change with further iteration.

Conclusion:

Since K-means clustering can't "see" the best clustering, the only way to find it is to keep track of the clusters and their total variance. Then, repeat the process with different starting points.

References:

- https://www.geeksforgeeks.org/k-means-clustering-introduction/
- https://gist.github.com/pmsosa/5454ade527adbee105dd51066ef30c5f

Data Management and Data Warehousing- CS-2376

A
PROJECT
REPORT
ON
Visualization and Decision Support

Submitted to:
Prof. Anirban Mondal

Submitted by:
Manish Yadav
Aman Antil
Deepak Chauraisya

Department of Computer Science
Ashoka University, Sonipat

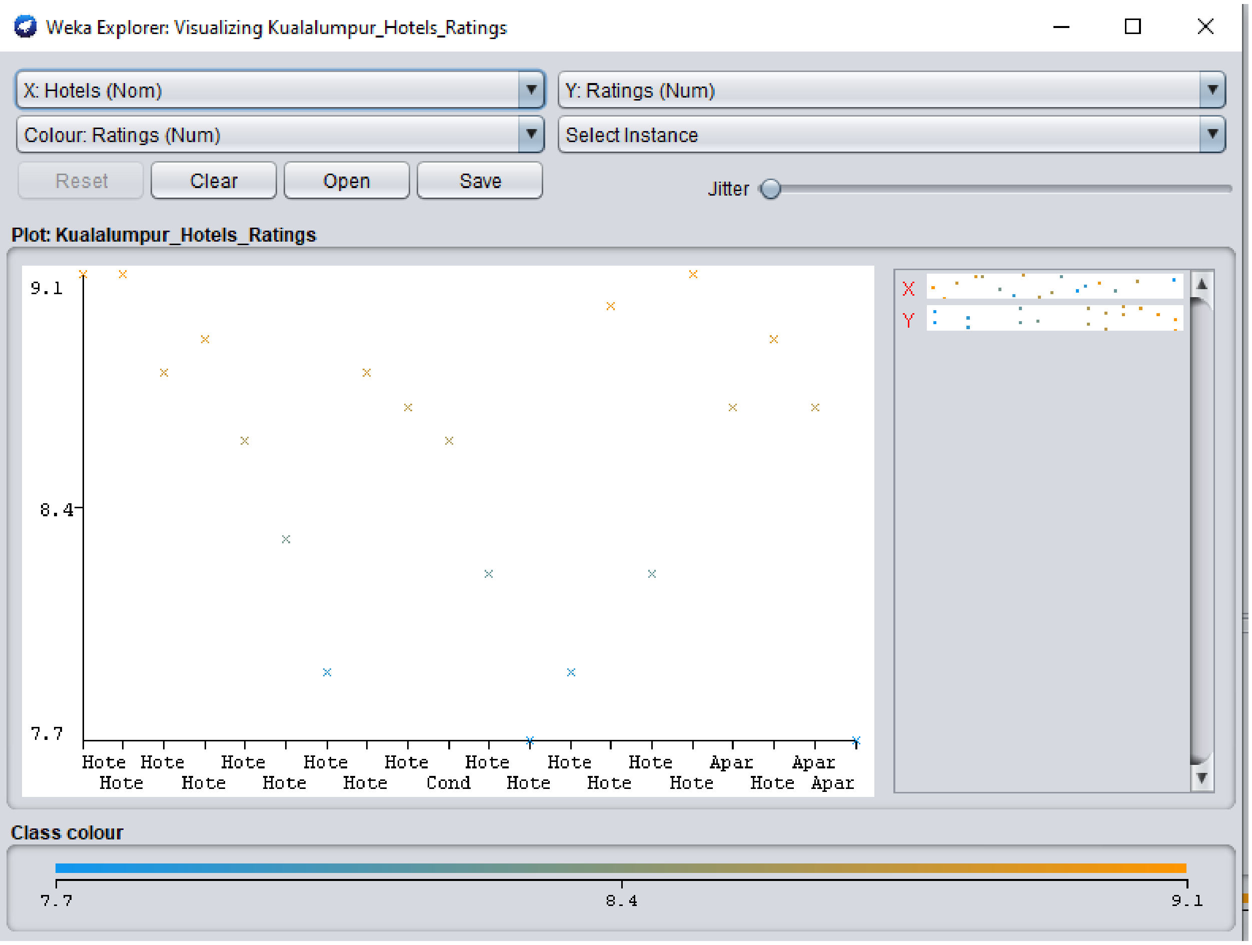
Objective:

Mined the data such as hotel name, rating, reviews, user rating from an e-travel and visualized it

Data Visualization:

We use selenium to mine the data from the given website. We stored relevant details in csv format. Here is how the visualization of hotels with their ratings looks like using weka. I have also clustered hotels based on their ratings.

1. Bangkok



Weka Explorer: Visualizing Kualalumpur_Hotels_Ratings

X: Hotels (Nom)

Y: Ratings (Num)

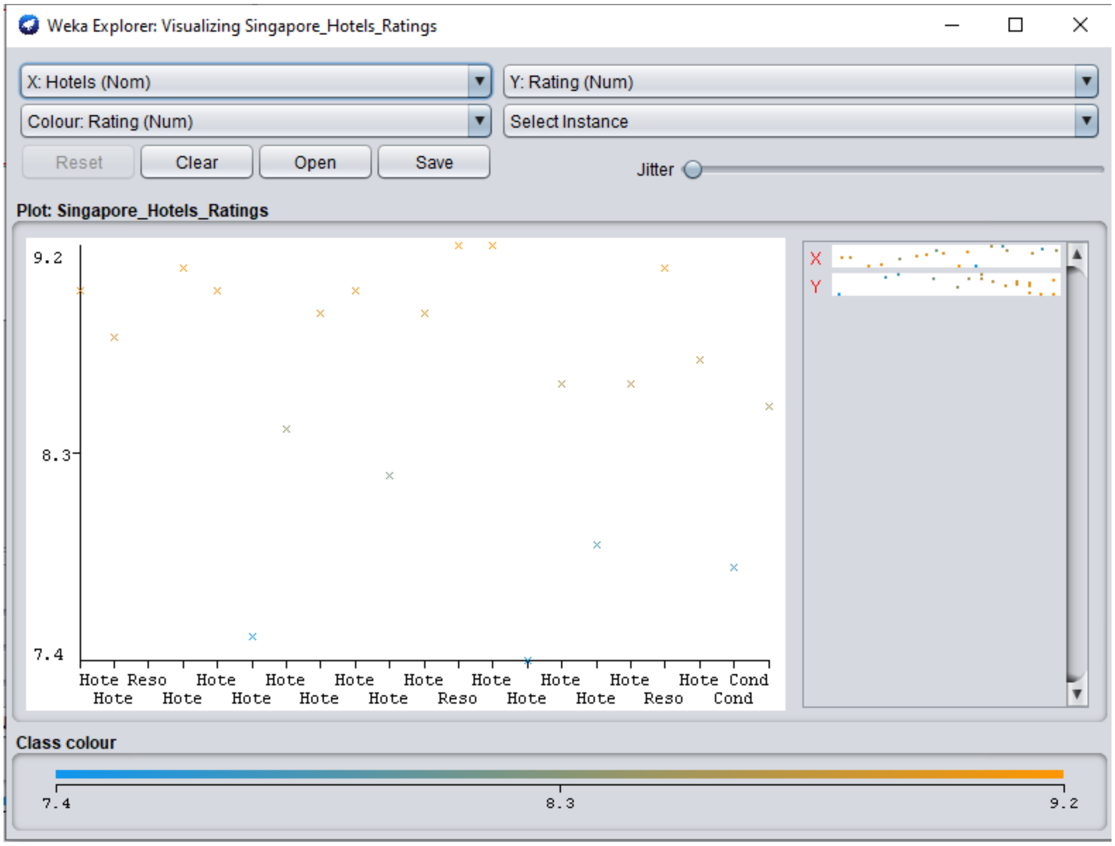
Colour: Ratings (Num)

Select Instance

Plot: Kualalumpur_Hotels_Ratings

Reset
Clear
Open
Save
Jitter
9.1
8.4
7.7
Hote Hote Hote Hote Hote Hote Hote Hote Apar Apar
Hote Hote Hote Hote Cond Hote Hote Hote Hote Apar
X
Y
Class colour
7.7
8.4
9.1

2. Singapore



Weka Explorer: Visualizing Singapore_Hotels_Ratings

X: Hotels (Nom)

Y: Rating (Num)

Colour: Rating (Num)

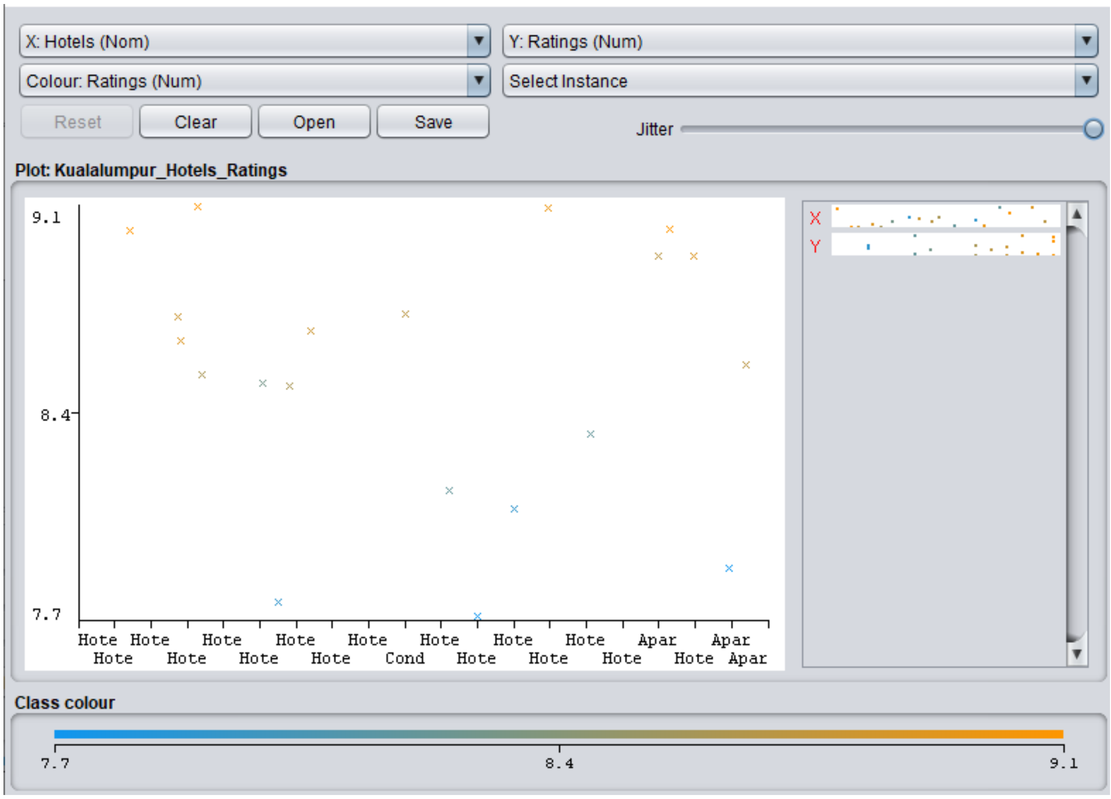
Select Instance

Plot: Singapore_Hotels_Ratings

Class colour

Reset
Clear
Open
Save
Jitter
9.2
8.3
7.4
Hote Reso Hote Hote Hote Hote Hote Hote Hote Hote Cond
Hote Hote Hote Hote Hote Reso Hote Hote Reso Cond
X
Y
7.4
8.3
9.2

3. Kuala-Lumpur



X: Hotels (Nom)

Y: Ratings (Num)

Colour: Ratings (Num)

Select Instance

Plot: Kualalumpur_Hotels_Ratings

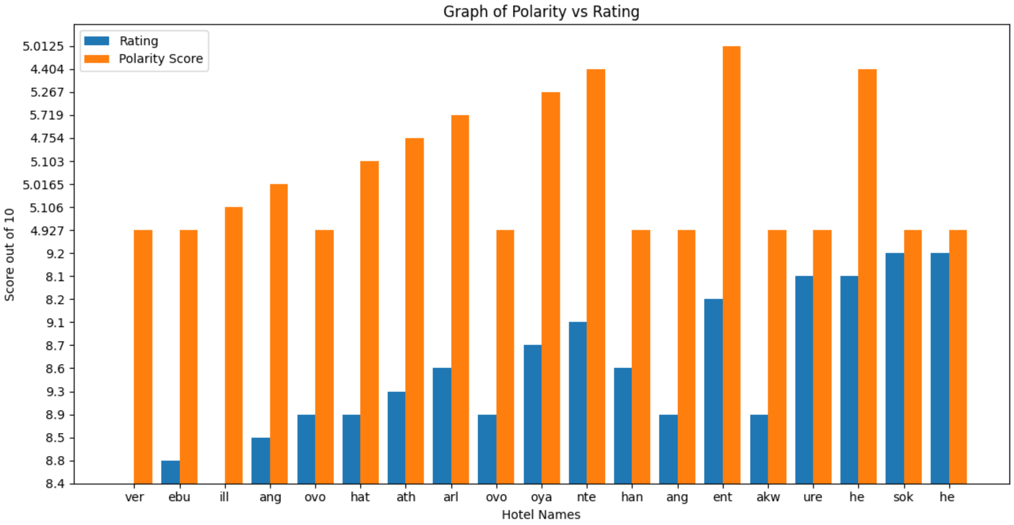
Class colour

Reset
Clear
Open
Save
Jitter
9.1
8.4
7.7
Hote Hote Hote Hote Hote Hote Hote Hote Apar Apar
Hote Hote Hote Hote Cond Hote Hote Hote Hote Apar
X
Y
7.7
8.4
9.1

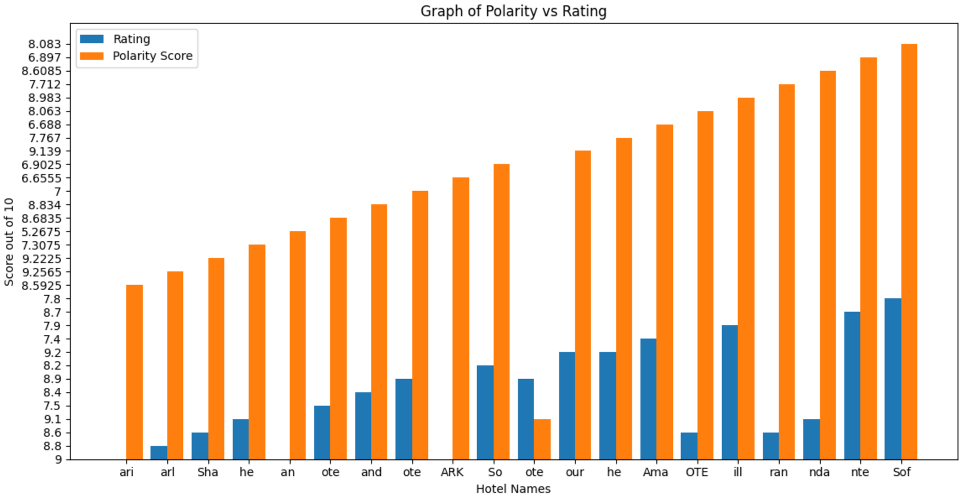
Sentiment Analysis:

We extract reviews of hotels from the csv file and calculate polarity score. We visualized polarity scores and ratings of the hotels as follows:

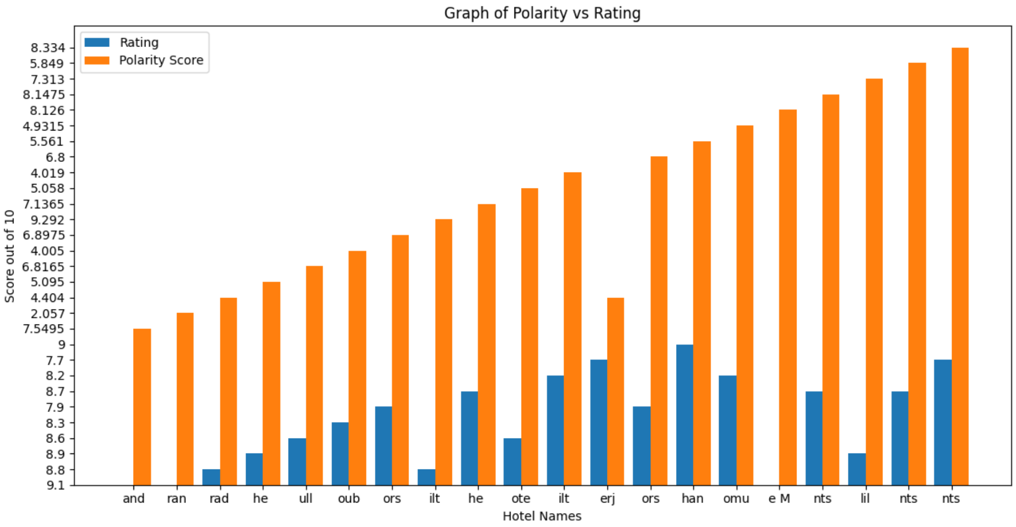
1. Bangkok



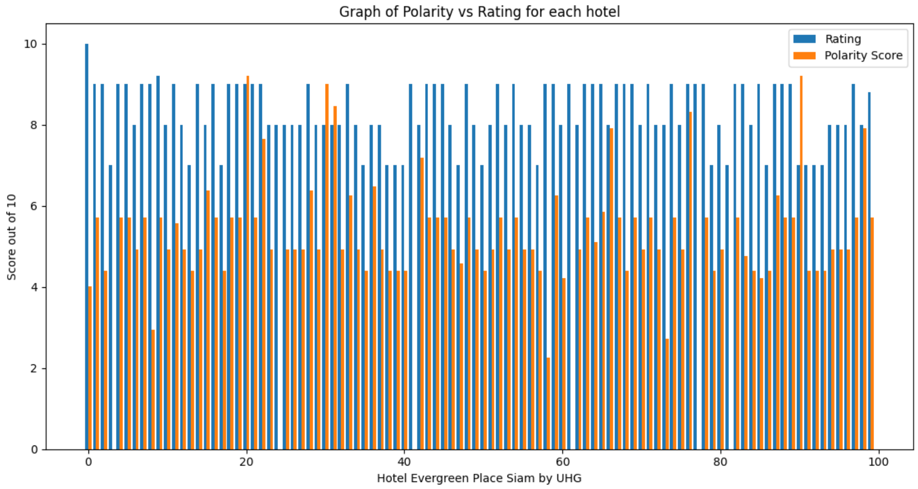
2. Singapore



3. Kuala-Lumpur



Now, We have written code to get the polarity score and rating of a particular hotel in one csv file. There are 60 hotels in total so it would take lots of space to paste all images. We visualized polarity scores and ratings of a particular (Hotel Evergreen Place Siam by UHG) hotel here. We can change the index in code and visualize other hotels too.



Data Management and Data Warehousing- CS-2376

A
PROJECT
REPORT
ON

ASSOCIATION RULE MINING ALGORITHM

Submitted to:
Prof. Anirban Mondal

Submitted by:
Manish Yadav
Aman Antil
Deepak Chauraisya

Department of Computer Science
Ashoka University, Sonipat

Objective:

To implement the Association Rule Mining algorithm.

Introduction:

The association rule mining is one of the ways to figure out such a pattern which puts the similar things together for example; people who buy coca cola are likely to buy namkeen or chips too. It is an implication expression of the form X→Y, where X and Y are disjoint itemsets.

Input:

The inputs are a transaction dataset, two threshold values (minimum support value) and the minimum confidence value. We can also provide the transaction dataset manually.

Output:

The output of the Association Rule Mining is a set of association rules on the dataset by using the minimum threshold value and minimum confidence value. The threshold support is the number of transaction which consists the set of A∪B where A and B are items. The confidence of the rule is the number of transactions that holds A∪B/A.

Experiment:

Step 1:- In the very first step, we have a **csv** file which is our required data. Now, we read the given dataset and create a list of items.

Step 2:- Then we convert the above list to dataframe with boolean values. In these steps we check that the particular items exist in datasets or not.

Step 3:- Will create an array from the dataset list with true/false values(depending if an item shows up/doesn't show up in the list together). Now, we are at the point to convert this array to a dataframe (df) utilizing things as section names.

Step 4:- Now from our python library we will import the Apriori function. Then we apply the algorithm to our data to extract the itemsets that have a minimum support value and also find confidence values that measure how often each item appears in the B .

Minimum support = sigma(A+B)/total
conf(A=>) = sup(AUB)/sup(A)

Step 5:- Mining the association rules, we will discover the association rules for the continuous itemsets which we determined in the above step. Then we import the necessary function from the page to decide the association rule for a given dataset utilizing some arrangement of boundaries. Then we will apply it to the frequent item dataset which we created in the above step.
There are three association ways to measure association. Here, A stands for antecedent and C stands for consequent.

1. Support: It says how popular an itemset is. It’s measured by the proportion of transactions in which an itemset appears.

$$\text{support}(A \rightarrow C) = \text{support}(A \cup C), \text{ range: } [0,1]$$

2. Confidence: **It** says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}.

$$\text{confidence}(A \rightarrow C) = \text{support}(A \rightarrow C)/ \text{ support}(A), \text{ range: } [0,1]$$

3. Lift: It says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

$$\text{lift}(A \rightarrow C) = \text{confidence}(A \rightarrow C)/\text{support}(C), \text{ range: } [0,\infty]$$

If the Lift value is near 1 shows that A and B almost often appear together as expected. If the Lift value is greater than 1 which means that both appear together more than expected and less than 1 means both appear less than expected. If value is more greater that means there is a strong association.

4. Leverage:

$$\text{levarage}(A \rightarrow C) = \text{support}(A \rightarrow C) - \text{support}(A) \times \text{support}(C), \text{ range: } [-1,1]$$

5. Conviction:

$$\text{conviction}(A \rightarrow C) = 1 - \text{support}(C)/1 - \text{confidence}(A \rightarrow C), \text{ range: } [0,\infty]$$

We didn’t need to define all the functions because it’s pre-defined in the python library.

Conclusion:

We have implemented the Association Rule Mining for market basket analysis. We focused on theory and application of the most common algorithms.

References:

- http://www.philippe-fournier-viger.com/spmf/AssociationRules.php#:~:text=The%20output%20of%20an%20association%20rule%20mining%20algorithm,works,%20it%20is%20necessary%20to%20review%20some%20definitions.
- http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/
- https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Mid-Term Exam

Name : Manish Yadav
Course Code : 2376
Prof. Anirban Mondal
Date : 16 March, 2020
Roll No : 1020191493

Question 1:

Show the execution of the k-means clustering algorithm on the following data, where k =2. Show the relevant steps clearly and in as much detail as you can. [8]

Solution 1:

We use k-means clusters to divide the data of similar interest in k non-overlapping subgroups. K is the value that we get as an input which tells how many clusters we want from the dataset. We than follow these steps:

Step 1: From the given data we choose k points and consider it as our centroid.

Step 2: We take a particular point and find the Euclidean Distance from the above k centroids. We assign the point to the closest cluster (based on closest distance)

Step 3: We then find the average or mean of all points in a particular cluster and consider it as our new mean. We do the same thing for other clusters also.

We repeat step one and two until we get a stabilized stage where centroid will not change further.

Input : k = 2 so we need to find 2 clusters. First we'll take two random centroids.

Centroid:

$K_1 = (1,1) = P_1$

$K_2 = (1,2) = P_2$

Points: $P_1 = (1,1)$

$P_2 = (1,2)$

$P_3 = (4,3.5)$

$P_4 = (5,4)$

Now will calculate distance of points from centroid.

for P_3

$d(P_1 P_3) = \sqrt{(4-1)^2+(3.5-1)^2} = \sqrt{9+6.25} = \sqrt{15.25}$

$d(P_2 P_3) = \sqrt{(4-1)^2+(3.5-2)^2} = \sqrt{9+2.25} = \sqrt{11.25}$

So P_3 in cluster K_2

for P_4

$d(P_1 P_4) = \sqrt{(5-1)^2+(4-1)^2} = \sqrt{16+9} = \sqrt{25}$

$d(P_2 P_4) = \sqrt{(5-1)^2+(4-2)^2} = \sqrt{16+4} = \sqrt{20}$

So P_4 falls in cluster K_2

K_1 cluster has — P_1

K_2 has — P_2, P_3, P_4

New centroid in case K_2 = $\left(\frac{4+1+5}{3}\right), \left(\frac{2+4+3.5}{3}\right)$

$= (3.3, 3.16)$

Again will calculate the distance using centroid $K_1 \neq (1,1)$ $K_2 = (3.3, 3.16)$

$d(K_1 P_2) = \sqrt{(1-1)^2+(1-2)^2} = \sqrt{1} = 1$

$d(K_2 P_2) = \sqrt{(3.3-1)^2+(3.16-2)^2} = \sqrt{5.29+1.35} =$

For P_3 :

$d(K_1 P_3) = \sqrt{(1-4)^2 + (1-3.5)^2} = \sqrt{9 + 6.25} = \sqrt{15.25}$

$d(K_2 P_3) = \sqrt{(3.3-4)^2 + (3.16-3.5)^2} = \sqrt{0.49 + 0.34} =$

For P_4 :

$d(K_1 P_4) = \sqrt{(1-5)^2 + (1-4)^2} = \sqrt{16+9} = \sqrt{25}$

$d(K_2 P_4) = \sqrt{(3.3-5)^2 + (3.16-4)^2} = \sqrt{2.89 + 0.71} =$

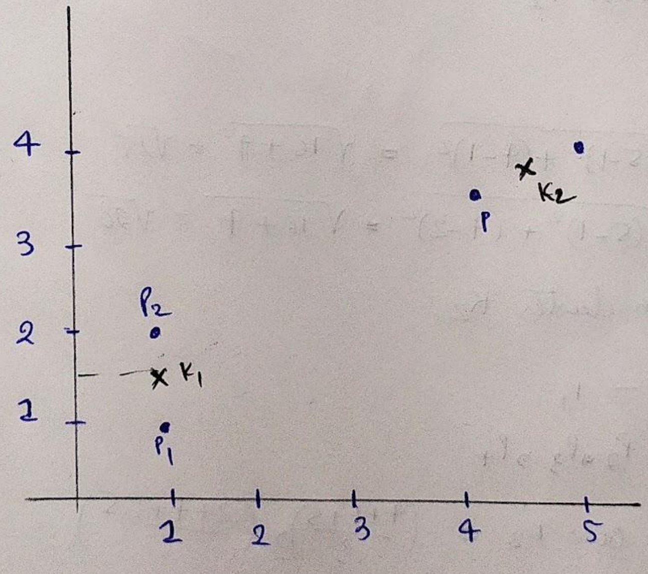
Now, K_1 cluster has – P_1 & P_2

K_2 cluster has – P_3 & P_4

Centroid will

$K_1 = (\cancel{1,1})\ \frac{1+1}{2}, \frac{1+2}{2} = 1, 1.5$

$K_2 = (4.5, 3.75)$



Question 2;

X is a store selling a mix of cheap as well as expensive products. You are given the data concerning a total of 10 million transactions that occurred in X during the past one year. (Each transaction can involve multiple items.) To reduce the number of transactions that the association rule mining algorithm needs to consider, the CIO of X has recently proposed that only transactions above $10 should be considered by the association rule mining algorithm.

Under what circumstances would the CIO's heuristic fail, assuming that the objective of the top management of X is to maximize total sales revenue? (For simplicity, ignore tax effects and other externalities.) Can you propose some heuristics for the circumstances in which the CIO's heuristic would fail? [4]

Solution 2:

Three scenarios where association rule mining would be ineffective:

1. We define support threshold (S) and based on that value we decide which product would be considered frequent or in the given sub-dataset. Let assume we have decided S (say a particular rupees) but most of the product let say 90% is below than S. Since products above S are only considered in the association rule so most of the products will be ignored.
2. In order to increase the profit we need to increase our sales. Since we aren't considering the huge amount (90%) of data in our association rule so we would not think about it. We need to find patterns and other relations between these 90% products because our profit totally depends on these 90% products.
3. Most transactions can be below S so the large number of transactions that occurred would be below S. So, if we had to make a cluster or subsets of the given data, it would not be useful because we are not considering 90% products. Frequently sold products might be in that 90% which helps to decide patterns.

   So, this association algorithm would fail.

Question 3:

How would you use data mining and analytics to improve customer satisfaction at a restaurant? First, define the context of the application scenario. You can make any reasonable assumptions. Discuss how you would proceed for each phase of data mining i.e., starting from data collection all the way to the interpretation of results. Clearly mention which data mining algorithms/techniques (e.g., association rule mining, clustering etc.) you would deploy and justify why your chosen algorithm is a good fit to the given application scenario. [8]

Note: This is an open-ended question. Please feel free to make any reasonable assumptions.

Solution 3:

Data mining and analytics can also be used to improve customer satisfaction in brick and mortar retail stores. It can be used in following way:

1. To get a pattern of what is frequently bought together which can help the store to recommend products while a customer makes a particular order. For instance, people order bricks with cement because it's probably bought for a house building. So, we can recommend different brands of cements such as ACC, JSW etc. People also buy tap, basin, and other home products that we can recommend to them while they order something. Like that, we can get customers choices based on their past purchasing.
2. Stores can also keep track of what people buy to decorate their houses and increase that particular stock.

We follow the data mining phases as follows:

1. Data Collection/ Consolidation
    a. Data can be collected on the basis of price, mode of transaction such as card, cash, or google pay. We also need to give preference to our regular customers. We collect the data of the variety of items ordered. Will also have sets of products ordered together with particular products. We also look for the brand of cement, brick, iron road that is frequently bought in a huge amount. Raw data is complicated because it's unstructured or and heterogeneous data. We need to process this data in a well structured format.
2. Data Preparation
    a. We get data in a raw or unstructured form that needs to be structured. So, we Information integration needs to be performed. That's a process of merging raw data or unstructured data. We merge the provided data in a conceptual, contextual way.
    b. Data can not be cleaned 100% so there might be some negligible error in the data.
    So, while performing information integration, try to remove the inconsistencies from the data as much as you can. We remove wrongly recorded data. It can happen because while writing details. Mistakenly, we have changed the decimal place or added zero. Customers may also unintentionally omit the details by cutting it.
    We use different algorithms and principles to find the cluster or dataset. We use the k-means algorithm, association rule, apriori principle and so on. Further, we use data visualization techniques to represent the data in a visual way.
3. Decision Making
    We visualize the data to find the pattern or products which are in more demand. Based on the cluster or dataset, managers can look for the trend to make a decision. So, he can decide which product should be recommended with what. He can also find which product

is frequently sold on a particular day or month. Based on the geographic data the manager can decide to vary the price of the commodity from place to place.

Final-Exam

Name: Manish Yadav
Course Code: 2376
Prof. Anirban Mondal
Date: 16 March 2020
ID: 1020191493

Question 1:

Which of the following queries/activities are related to data warehousing and OLAP? [2]

(A) A user querying for his/her account balance from an ATM

(B) A book publisher deciding whether to increase the price of a set of books by $5 each

(C) A user Jack initiates a transaction to send $100 to another user John

(D) A large retail store trying to determine whether to give a storewide 10% discount during any week of a holiday season.

Your answer to this question should just be letters e.g., {A, B}, {A, B, C) etc. No justification is required. You have to get all the letters right to earn the 2 marks for this question. No partial marks if you get one of the letters right.

Solution 1:

{B, D}

Question 2:

Would agglomerative hierarchical clustering be effective for handling data streams? Justify. [2]

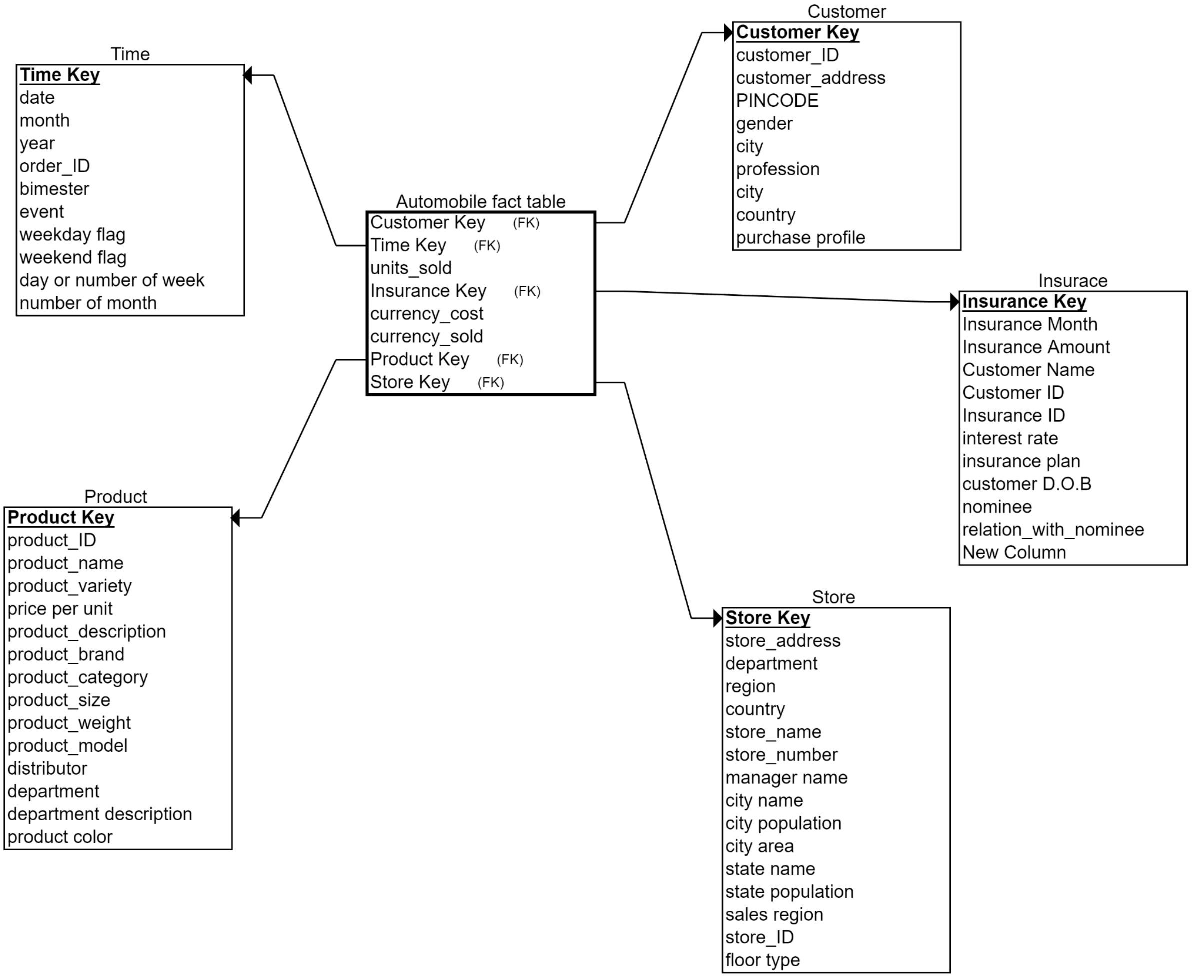
Solution 2:

No! agglomerative clustering wouldn't be effective for handling data streams. It is because agglomerative hierarchical clustering has large storage requirements and they can be computationally complicated to handle. Datastream is a sequence of digitally encoded coherent signals. The amount of data that would be collected will be in large amounts which will make agglomerative clustering more complicated and acquire more space. In comparison to the k-means algorithm, it is four times bigger. We can't reverse the merged data which creates another problem if we have noisy and high dimensional data.

Question 3:

Create a star schema design for organizing sales data for a large multinational automobile company (e.g., Toyota, GM, Ford) [6] • State two reasonable queries for the above scenario. Note: This is an open-ended question. Please feel free to make any reasonable assumptions.

Solution 3:



Question 4:

How would you use data mining and analytics to improve customer satisfaction for a logistics firm such as FedEx? First, define the context of the application scenario. You can make any reasonable assumptions. Discuss how you would proceed for each phase of data mining i.e., starting from data collection all the way to the interpretation of results. Clearly mention which data mining algorithms/techniques (e.g., association rule mining, clustering etc.) you would deploy and justify why your chosen algorithm is a good fit for the given application scenario. [10] Note: This is an open-ended question. Please feel free to make any reasonable assumptions.

Solution 4:

Data mining and analytics can also be used to improve customer satisfaction in brick and mortar retail stores. It can be used in the following way:

1. Getting a pattern of what is frequently bought together can help the store to recommend products while a customer makes a particular order. For instance, people order bricks with cement because it's probably bought for a house building. So, we can recommend different brands of cement such as ACC, JSW etc. People also buy taps, basins, and other home products that we can recommend to them while they order something. Like that, we can get customers' choices based on their past purchasing.
2. Stores can also keep track of what people buy to decorate their houses and increase that particular stock.

We follow the data mining phases as follows:

1. Data Collection/ Consolidation
    a. Data can be collected on the basis of price, and mode of transaction such as card, cash, or google play. We also need to give preference to our regular customers. We collect the data on the variety of items ordered. Will also have sets of products ordered together with particular products. We also look for the brand of cement, brick, and the iron road that is frequently bought in a huge amount. Raw data is complicated because it's unstructured or heterogeneous data. We need to process this data in a well-structured format.
2. Data Preparation
    a. We get data in a raw or unstructured form that needs to be structured. So, Information integration needs to be performed. That's a process of merging raw data or unstructured data. We merge the provided data in a conceptual, contextual way.
    b. Data can not be cleaned 100% so there might be some negligible error in the data.
    So, while performing information integration, try to remove the inconsistencies from the data as much as you can. We remove wrongly recorded data. It can happen because while writing details. Mistakenly, we have changed the decimal place or added zero. Customers may also unintentionally omit the details by cutting them.
    We use different algorithms and principles to find the cluster or dataset. We use the k-means algorithm, association rule, apriori principle and so on. Further, we use data visualization techniques to represent the data in a visual way.
3. Decision Making
    We visualize the data to find the pattern or products which are in more demand. Based on the cluster or dataset, managers can look for the trend to make a decision. So, he can decide which product should be recommended with what. He can also find which product

is frequently sold on a particular day or month. Based on the geographic data the manager can decide to vary the price of the commodity from place to place.