COMPUTER ENGINEERING DEPARTMENT

ASSIGNMENT NO-02

SUB: Data Warehousing & Mining

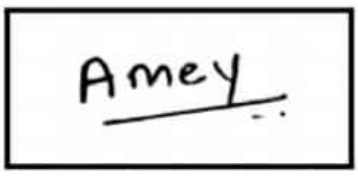
COURSE: T.E. **Year: 2020-2021** **Semester: VI**
DEPT: Computer Engineering
SUBJECT CODE: CSC603 **SUBMISSION DATE: 04/05/2021**

Name: Amey Thakur **Roll No.: 50**

Class: TE Comps B-50 **ID: TU3F1819127**

Assignment No 2

Sr. No.	Question	CO mapping
1	Differentiate between Classification, Prediction and Clustering.	CO1
2	Explain Accuracy and Error measures for Multiple linear regression Model Evaluation & Selection.	CO1
3	Explain the Apriori algorithm with an example.	CO2
4	Explain Mining Multilevel Association Rules, Multidimensional Association Rules in detail.	CO2
5	Differentiate between Spatial Vs. Classical Data mining.	CO3
6	Write a short note on types of web mining.	CO3



AMEY B 50 Amey

Page No.: YOUVA
Date:

Q1. Differentiate between Classification, Prediction and Clustering.

Ans:

Classification	Prediction	Clustering
In classification, a training set containing data that have been previously categorised and based on this training set, the algorithm finds the category that the new data point belong to.	It is used to access the values or value ranges of an attribute that a given sample is likely to have.	The characteristics of similarity of data is not known in advance so using statistical concepts we split datasets into sub-datasets such that they have similar data called as clusters
Classification is supervised learning	Both Classification and Clustering are used to make predictions.	Clustering is unsupervised learning.
You're given an unseen tuple and you are suppose to set a label or a class to that tuple.	Construction and use of a model to access the class of an unlabelled sample.	You are given a set of transaction history that gives us details about which customer bought what item.
Example: - Group patients based on their known medical data and treatment outcome then its a classification	Example: - If a classification model is used to predict the treatment outcome for a new patient, then it would be a prediction	Example: - By clustering techniques you can tell the segmentation of your customer.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Q.2. Explain accuracy and Error measures for multiple linear regression model evaluation and selection.

Ans:

Accuracy and Error Measures

Accuracy of a classifier M, acc(M) is the percentage of set tuples that are correctly classified by the model M.

Basic Concepts

Partition of data randomly into three sides.

Training Set:

- It is a set of instances that have not been used in training process. The models performance is evaluated on unseen data testing just estimate the probability of success of unknown data.

Validation Data:

- It is used for parameters tuning but it can not be the test data. validation data can be training or a subset of training data.

Generalization error: Model error on test data

→ Success: instance class is predicted correctly.

→ Error: instance class is predicted incorrectly.

The confusion Matrix

- It is useful tool for analyzing how well your classifier can recognize types of different classes

AMEY B 50 Amey

Page No.:
Date:
YOUVA

→ TP: Class member which are classified as class member.
→ TN: Class non-member which are classified as non-member.
→ FP: Class non-member which are classified as member.
→ P: Number of positive tuples.
→ N: Number of negative tuples.
→ N': Number of tuples that were labelled negatively.
→ P': Number of tuples that were labelled positively.

Sensitivity.
- True positive recognition rate which is the proportion of positive tuples that are correctly identified sensitivity TP/P.

Specificity
- True negative recognition rate which is the proportion of negative tuples that are correctly classified.

$$\text{Specificity} = \frac{TN}{N}$$

Classifier accuracy on recognition rate.

$$\text{Accuracy} = \frac{(TP + TN)}{D}$$

$$\text{Accuracy} = \text{Sensitivity} \frac{P}{(P+N)} + \text{Specificity} \frac{N}{(P+N)}$$

$$\text{Error rate} = 1 - \text{Accuracy}$$

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

AMEY B 50 Amey

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|}$$

Fmeasure

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_\beta = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

β = non-negative real number.

AMEY B 50 Amey

Page No.: Date: YOUVA

Q3. Explain the apriori algorithm with example.

Ans:

Apriori Algorithm

- The Apriori algorithm uses frequent itemsets to generate association rules, and it is designed to work on the dataset that contain transactions.
- With the help of association rule, it determines how strongly or how weakly two objects are connected.
- This algorithm uses a breadth-first search and Hash tree to calculate the itemset associations efficiently
- It is the iterative process for finding the frequent itemsets from large dataset.

Steps for Apriori Algorithm.

step 1 - Determine the support of itemsets in the transactional database, and select the minimum support and confidence.

step 2 - Take all supports in the transaction with higher support value than the minimum or selected support value.

step 3 - Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

step 4 - Sort the rules as the decreasing order of lift.

AMEY B 50 Amey

Page No.: YOUVA
Date:

Example:

TID	ITEMS
T1	L_1, L_2, L_5
T2	L_2, L_4
T3	L_2, L_3
T4	L_1, L_2, L_4
T5	L_1, L_3
T6	L_2, L_3
T7	L_1, L_3
T8	L_1, L_2, L_3, L_5
T9	L_1, L_2, L_3

Minimum support Count = 2

Minimum Confidence = 60%

Step 1: $k = 1$

Create a table containing support Count of each item present in dataset called (Candidate Set)

Itemset	Support Count
L_1	6
L_2	7
L_3	6
L_4	2
L_5	2

Compare candidate set items support count with minimum support count.

This gives itemset L1.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

Step 2 : K=2

Generate candidate set C2 using L1.

- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.
- Now find support count of three itemsets by searching in dataset.

itemset	Support Count
L1, L2	4
L1, L3	4
L1, L4	1
L1, L5	2
L2, L3	4
L2, L4	2
L2, L5	2
L3, L4	0
L3, L5	1
L4, L5	0

Compare candidate (C2) support count with minimum support count that gives us itemset L2

itemset	Support Count
L1, L2	4
L1, L3	4
L1, L5	2
L2, L3	4
L2, L4	2
L2, L5	2
L4, L5	2

AMEY B 50 Amey

Page No.: Date: YOUVA

Step 3:

Similarly

itemset	Support count
L_1, L_2, L_3	2
L_1, L_2, L_4	2

Compare C3 this gives L3

itemset	Support count
L_1, L_2, L_3	2
L_1, L_2, L_5	2

We stop here because no frequent itemsets are found further.

Confidence:

$$\text{Confidence}(A \rightarrow B) = \frac{\text{support count}(A \cup B)}{\text{support count}(A)}$$

Step 4: Association rules

$[L_1 \wedge L_2] \Rightarrow [L_3]$ // $2/4 * 100 = 50\%$

$[L_1 \wedge L_3] \Rightarrow [L_2]$ // $2/4 * 100 = 50\%$

$[L_2 \wedge L_3] \Rightarrow [L_1]$ // $2/4 * 100 = 50\%$

$[L_1] \rightarrow [L_2 \wedge L_3]$ $2/6 * 100 = 33\%$

$[L_2] \rightarrow [L_1 \wedge L_3]$ $2/7 * 100 = 28\%$

$[L_3] \rightarrow [L_1 \wedge L_2]$ $2/6 * 100 = 33\%$

So if minimum confidence = 50%

Then first 3 rules can be considered as strong association rules

AMEY B 50 Amey

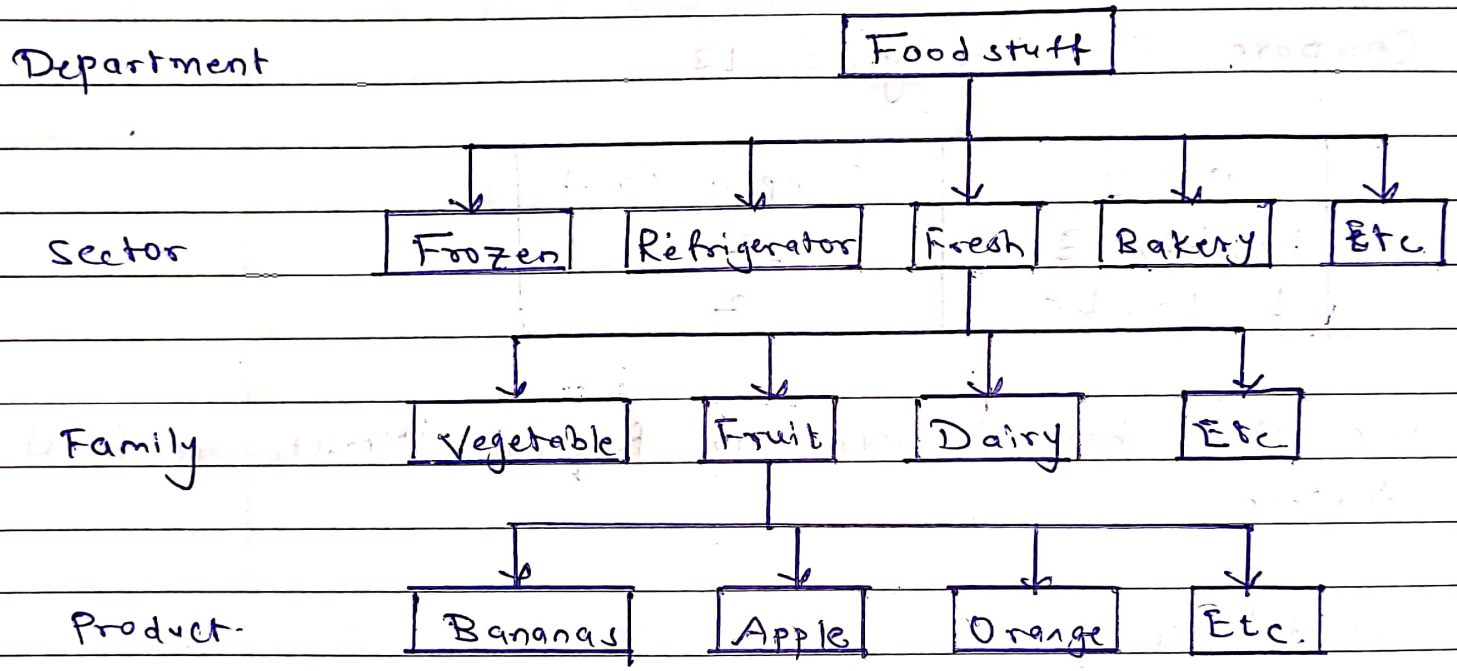
Page No.:
Date:
YOUVA

Q4. Explain mining of multilevel association rule, multidimensional Association Rule.

Ans:

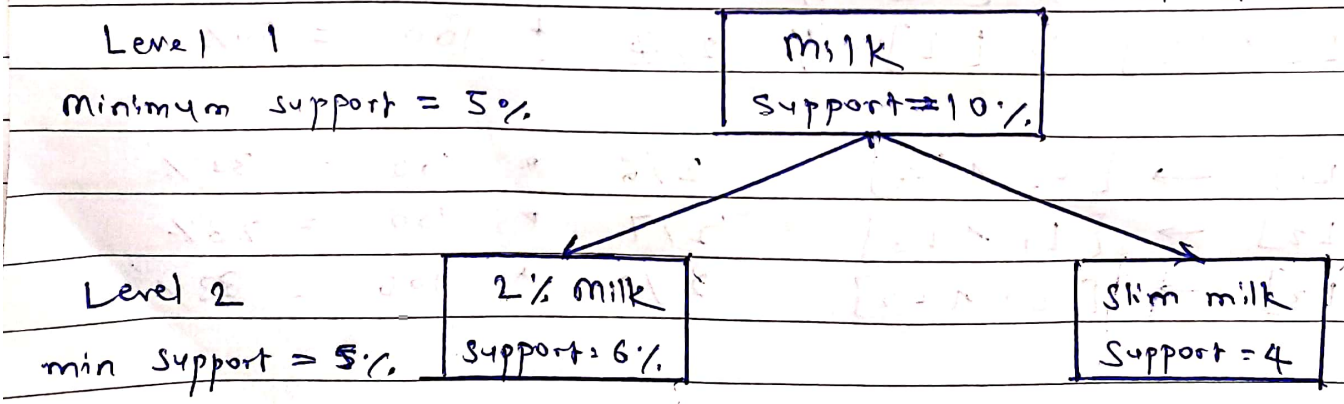
Multilevel Mining Association rules

- Items often form hierarchy
- Items of the lower has lower support.
- Rules which contain associations with hierarchy of concepts are called multilevel association rules.



(1) Using uniform support level for all levels

- The same minimum support for all levels
- There is only one minimum support threshold no need to examine itemsets
- If support threshold is too low → generate too many high level association.

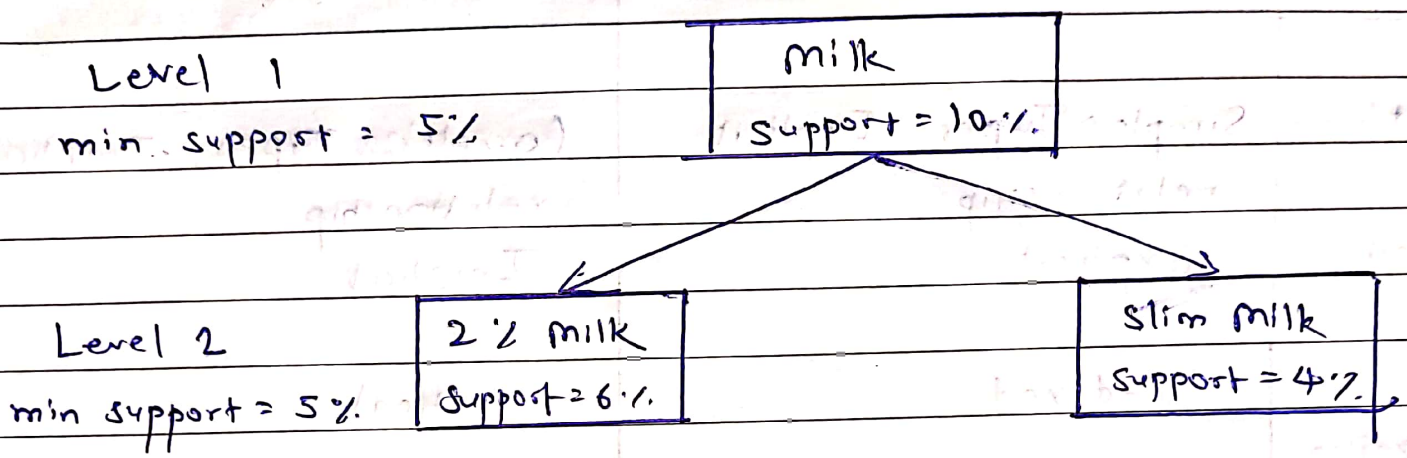


AMEY B. 50 Amey

Page No.:
Date:
YOUVA

(2) Using reduced minimum support level for all levels

- At every level of abstraction, there is its own minimum support threshold
- So minimum support at lower level reduced.



Multidimensional Association Rule.

(1) In multidimensional association.
- Attribute can be categorical or quantitative
- Quantitative attributes are numeric and incorporates hierarchy
- Numeric attributes must be discretized.
- It consists of more than one dimension.

(2) Three approaches in mining multi dimensional association rule.

(i) Using static criticization of quantitative attributes
(ii) Using dynamic discriticization of quantitative attributes
(iii) Using distance based discriticization with clustering

AMEY B 50 Amey

Page No.:
Date:
YOUVA

① Using static discritization of quantitative attributes

- Discritization is static and occurs prior to mining
- Discritized attributes are treated as categorical
- Use apriori algorithm to find all k-frequent predicate sets. (this requires 'k or k+1 table scans.
- Every subset of frequent predicate set must be frequent.
- Eg If in a data cube the 3D cuboid (age, income, buys) is frequent implies (age, income), (age, buys), (income, buys) are also frequent.
- Data cubes are well suited for mining since they make mining faster.
- The cells of an n dimensional data cuboid correspond to the predicate cells.

② Using dynamic discritization of quantitative attribute

- Known as mining quantitative association rules
- Numeric attributes are dynamically discreticized
- Eg: age(X, "20..25") ∧ income(X, "30k..41.k") buys (X, "Laptop Computer")

	Age=20	Age=21	Age=22	Age=23
income, 38 to 41				
income, 34 to 37				
income, 30 to 33				

Grid for Tuples.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

(3) Using distance based discritization with clustering

- It involves a two step mining process
  - → Perform clustering to find the interval of attributes involved.
  - → Object association rules by searching for groups of clusters that occur together.

- The resultant rules may satisfy
  - → Clusters in the rule antecedent are strongly associated with clusters of rules in the consequent.
  - → Clusters in the antecedent occur together.
  - → Clusters in the consequent occur together.

AMEY B 50 Amey

Page No.: YOUVA
Date:

Q5. Differentiate between Spatial Vs. Classical Data mining

Ans:

Parameters	Classical Data mining	Spatial Data mining
Data Definition	Simple	Complex
Relationships	Explicit	Implicit
Data organization	Indexed	Vertical
Statistical Foundation	Independencies of samples	Spatial auto correlation
Output	Set based	Spatial based
Example	Classification Accuracy	Spatial Accuracy
Computation Process	Combinational optimization numerical algorithm	Computational efficiency opportunity. Plane sweeping Spatial auto correlation

AMEY B 50 Amey

Page No.: YOUVA
Date:

Q.6 Write a short note on types of web mining.

Ans:

Web mining

- Web mining is the process of data mining technique to automatically discover and extract information from web documents and services.
- Web mining can be broadly divided into 3 types.

(1) Web content mining

- It is the application of extracting useful information from the content of web documents.
- Web content consists of several types of data: text, images, audio, video, etc.
- Content data is the group of facts that a web page is designed.
- It can provide effective and interesting patterns about user needs.
- Text documents are related to text mining, machine learning and Natural Language Processing

(2) Web structure mining

- It is the application of discovering structure information from the web.
- The structure of the web graph consists of web pages as nodes and hyperlinks as edges connecting related pages.
- Structure mining basically shows the structured summary of a particular website.
- It identifies relationship between web pages linked by information or direct link connection.
- To determine the connection between two commercial websites, web structure mining can be very useful.

AMEY B 50 Amey

Page No.:
Date:
YOUVA

③ Web usage Mining

- It is the application of identifying on discovering interesting usage patterns from large datasets and these pattern enables us to understand the user behaviour or something like that.
- In usage mining, user access data is form so usage mining is also called mining.