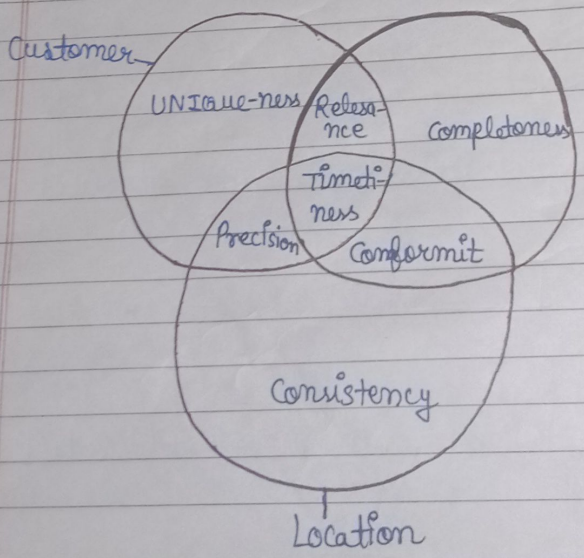
Unit 2

★ Data Quality :-

* Definition :-

- Data quality refers to the overall utility of a data- sells as a function of its ability to be easily processed and analyzed for other uses, usually by a database, data warehouse, or data analysis system.
- Data quality is the measure of how well suited a data set is to serve its specific purpose.
- Measure of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness and timeliness.



Page No. :
Date :

☆ What is Data Quality :-

- Data quality refers to the development and implementation of activities that apply quality management techniques to data in order to ensure the data is fit to serve the specific needs of an organization in a particular context.
- Data that is deemed fit for its intended purpose is considered high quality data.
- Examples of data quality issues include duplicated data, incomplete data, incorrect data, poorely defined data, poorly organised data and poor data security.

☆ Data Quality Dimensions :-

1. Accuracy :-

- The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable sources.

2. Completeness :-

- completeness is a measure of the data's ability to efficitively deliver all the required values that are available.

3. Consistency :-

- Data consistency refers to the uniformity of data as it moves across networks and applications.

- The same data values stored in different locations should not conflict with one another.

4. Validity :-

- Data should be collected according to defined business rules and parameter, and should conform to the right format and fall within the right range.

5. Uniqueness :-

- Uniqueness ensures there are no duplications or overlapping of values across all data sets.
- Data cleaning and deduplication can help remedy a low uniqueness score.

6. Timeliness :-

- Timely data is data that is available when it is required.
- Data may be updated in real time to ensure that it is readily available and accessible.

☆ How to Improve Data Quality :-

- Data quality measures can be accomplished with data quality tools, which typically provide data quality management capabilities such as :

1. Data Standardization :-

Page No. :
Date :

- Disparate data sets are confirmed to a common data format.

2. Geocoding :-

- The description of a location is transformed into coordinates that confirm to U.S and worldwide geographic standards.

3. Matching or linking :-

- Data matching identifies and merges matching pieces of information in big data sites.

4. Data Quality Monitoring :-

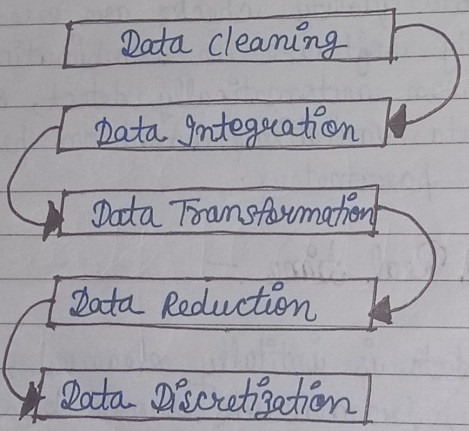
- Frequent data quality checks are essential.
- Data quality software in combination with machine learning can automatically detect, report, and correct data variations based on predefines business rules and parameters.

5. Batch and Real time :-

- once the data is intitally cleansed, an effective data quality framework should be able to deploy the same rules and processes across all applications and data types at scale.

Data Pre-Processing

- Data preprocessing is a data mining techniques which is used to transform the raw data (real-world data in the form of text, image, videos, etc, is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn't have a regular, uniform design) in a useful and efficient format.
- It is also an important step in data mining as we can't work with raw data.
- The quality of the data should be checked before applying machine learning or data mining algorithms.



Page No. :
Date :

1. Data cleaning :-

- Data is cleaned through processes such as filling in missing values or deleting rows with missing data, smoothing the noise data, or resolving the inconsistencies in the data.
- Smoothing noisy data is particularly important for ML datasets.
- Since machines can't make use of data they can't interpret.
- Data can be cleaned by dividing it into equal size segments that are thus smoothed (binning), by fitting it to a linear or multiple regression function, or by grouping it into clusters of similar data.
- Data inconsistencies can occur due to human errors.
- Duplicated values should be removed through deduplication to avoid giving that data object an advantage.

2. Data Integration :-

- Data with different representations are put together and conflicts within the data are reduced.

3. Data transformation :-

- Data is normalized and generalized.
- Normalization is a process that ensures that no data is redundant, it is all stored in a single place, and all the dependencies are logical.

1. Data Reduction :-

- when the volumn of data is huge, databases can become slower, costly to access, and challenging to properly store.
- Data reduction aims to present a reduced representation of the data in a data warehouse.
- There are various methods to reduce data.
- for ex, once a subset of relevant attributes is chosen for its significance, anything below a given level is discarded.

5. Data Discretization:-

- Data could also be discretized to replace raw values with internal levels.
- This step involves the reduction of a number of values of a continous attribute by dividing the range of attribute internals.
- It convert a huge number of data values into smaller one so that the evaluation and managment of data become easy.

Classification Analysis

Page No.:
Date:

* Definition :-

- Classification analysis is a data analysis task within data mining.
- That identifies and assigns categories to a collection of data to allow for more accurate analysis.
- The classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

☆ What is classification in Data Mining :-

- Classification in Data mining is a common technique that separates data points into different classes.
- It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones.
- It primarily involves using algorithms that you can easily modify to improve the data quality.
- This is a big reason why supervised learning is particularly common with classification in techniques in data mining.
- The primary goal of classification is to connect a variable of interest with the required variables.
- The variable of interest should be of qualitative type.

☆ There are two steps in the construction of a classification model :-

- Learning Step
- Classification Step

1. Learning step:-

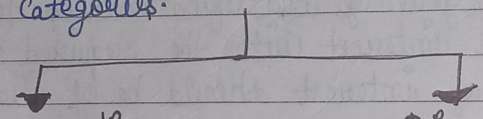
- This is where different algorithms are used to build a classifier by making the model learn using the training set available.
- The model has to be trained for the prediction of accurate results.

2. Classification step:-

- This is where the model used to predict class labels, tests the constructed model on test data.
- which in turn estimates the accuracy of the classification rules.

☆ Types of classification techniques in Data Mining:-

- Before we discuss the various classification algorithms in data mining.
- Primarily, we can divide the classification algorithms in 2 categories.



1. Generative:-

- A generative classification algorithms models the distribution of individual classes.

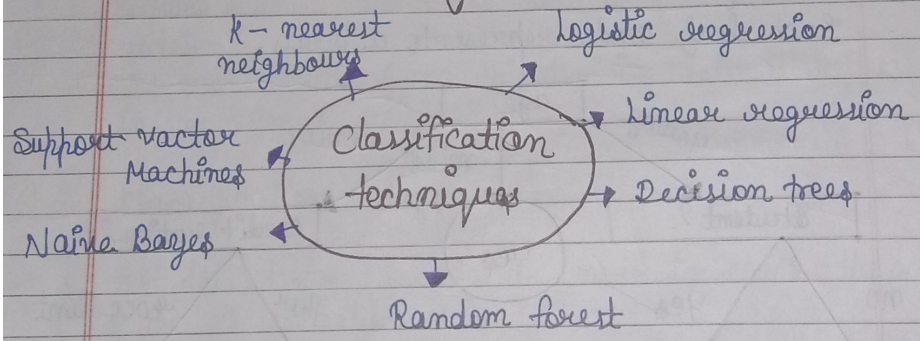
Page No. :
Date :

- It tries to learn the model which create the data through estimation of distribution and assumptions of the model.
- You can use generative algorithms to predict unseen data.
- It is a Naive Bayes classifier.

2. Discriminative :—

- It's a rudimentary classification algorithm that determines a class for a row of data.
- It models by using the observed data and depends on the data quality instead of its distributions.
- Its excellent type is logistic regression.

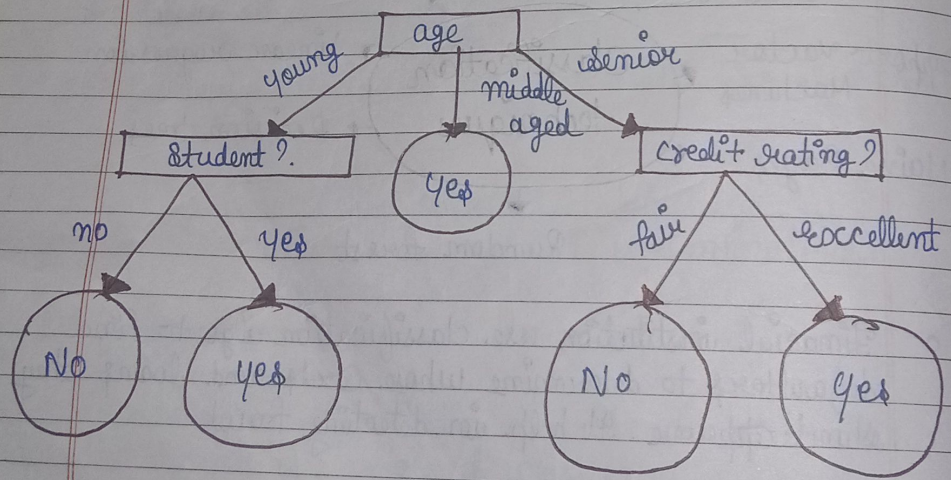
Classification techniques



- Finacial institution use classification algo to find defaulters to determine whose cards and loans they should approve. It help in detecting fraud.

Decision tree

- A decision tree is a structure that includes a root node, branches, and leaf nodes.
- Each internal decision node denotes a test on an attribute, each branch denotes the outcome of a test.
- Each leaf node holds a class label.
- The topmost node in the tree is the root node.
- The following decision tree is for the concept buy-computer that indicates whether a customer at a company is likely to buy a computer or not.
- Each internal node represents a test on an attributes.
- Each leaf node represents a class.



Page No.:
Date:

* Benefits of having a decision tree are as follows:-

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

→ A classifier (tree structured).
→ Also used for regression.

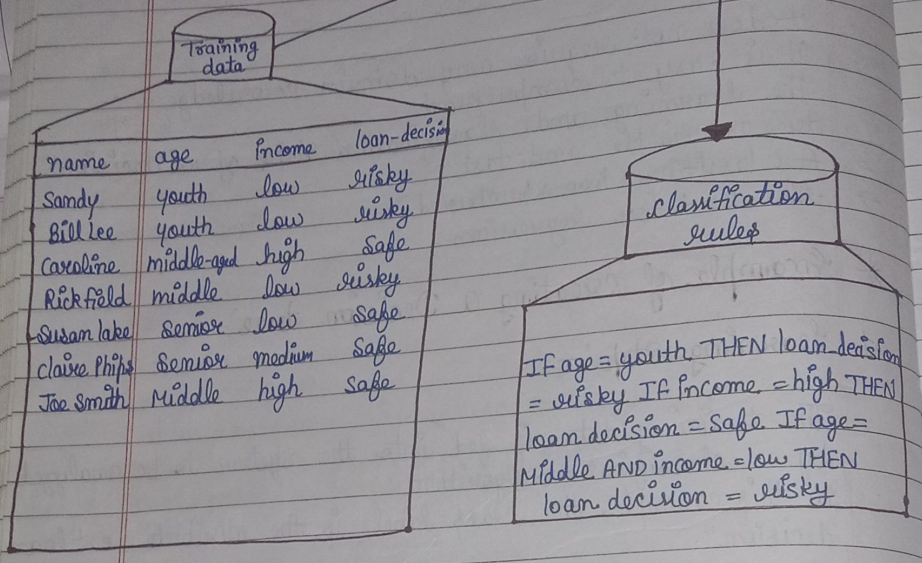
* Example of creating a Decision tree:-

1. Learning step:-

- The training data is fed into the system to be analyzed by a classification algorithm.
- In this example, the class label is the attribute i.e. "loan decision".
- The model built from this training data is represented in the form of decision rules.

2. classification:-

- Test dataset are fed to the model to check the accuracy of the classification rule.
- If the model gives acceptable results then it is applied to a new dataset with unknown class variables.



name	age	income	loan-decision
Sandy	youth	low	risky
Bill lee	youth	low	risky
Caroline	middle-aged	high	Safe
Rick field	middle	low	risky
Susan lake	Senior	low	Safe
Claire Phips	Senior	medium	Safe
Joe Smith	Middle	high	Safe

* Attributes :-

1. Binary Attributes :-

- A binary attribute is a nominal attribute with only two elements or states including 0 or 1, where 0 frequently represents that the attributes is absent, and 1 represents that it is present.
- Binary attributes are defined as Boolean if the two states are equivalent to true and false.
- A binary attribute is symmetric if both of its states are equal valuable and make an equal weight.

- There is no preference on which results must be coded as 0 or 1.
- An example can be the attribute gender having the states male and female.
- A binary attribute is asymmetric if the outcomes of the states are not equally essential, such as the positive and negative outcomes of a medical check for HIV.
- By convention, it can code the most essential result, which is generally the nearest one, by 1 and the different by 0.

2. Nominal attributes :—

- Nominal defines associating with names.
- The values of a nominal attributes are symbols or names of things.
- Each value defines some type of category, code, or state etc.
- Nominal attributes are defined as categorical.
- The values do not have any significant order.
- In computer science, the values are also called enumerations.

3. Ordinal attributes :—

- An ordinal attribute is an attribute with applicable values that have an essential series or ranking among them, but the magnitude b/w successive values is unknown.
- It is an attribute whose possible values have a meaningful order or ranking among them, but the magnitude b/w successive value is not known.

Unit → 4th

* Frequent Item Set Generation :-

Itemset :- An itemset consists of two or more items.
An itemset that occurs frequently is called a frequent.
. (Statistics) A group of related items. forex → Bread & Butter

Frequency :-

. Frequency is the no. of occurences of a repeating event per unit of time. It is also called Frequency.

☆ Frequent itemset :-

. Frequent itemset is an itemset whose support is greater than same user specified minimum support.
. A candidate itemset is a potentially frequent itemset.

Apriori Algorithm

. Apriori algorithm uses frequent itemset to generate association rules.
. It is based on the concept that a subset of a frequent itemset must also be a frequent itemset.

Page No. :
Date :

- Frequent itemset is an itemset whose support value is greater than a threshold value (support).
- It uses a bottom up approach, designed for finding association rule in a database that contains transactions.

* Advantages :-

1. Easy to implement among association rule learning algo.
2. use large itemset-properties.

* Disadvantages :-

1. Very low, and low minimum support in the data set.
2. Requires many database scans.

Apriori Principle

- Apriori Algorithm is a classification algorithm in data mining.
It is used for mining frequent itemset and relevant association rules.
It is devised to operate on a database containing a lot of transactions.
for instance items brought by customer in a store.
When making sure that all of the patterns in a set of data meet the minimum support requirements, we want to find all of the patterns that are supported, and not waste time looking at patterns that aren't.
This seems simple, but in much larger data sets, it can

become difficult to keep track of which patterns are support and which aren't.

Components of Apriori Algorithm

1. Support
2. Confidence
3. Lift

1. Support :-

- Support refers to the default popularity of any product.
- you find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions.
- Hence we get

$$\text{Support (Biscuits)} = \frac{\text{(Transaction relating Biscuits)}}{\text{(Total transactions)}}$$

$$= 400/4000 = 10 \text{ percent.}$$

2. Confidence :-

- Confidence refers to the possibility that the customers bought both biscuits and chocolates together.

Page No. :
Date :

- So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.
Hence,

$$\text{confidence} = \frac{(\text{Transactions relating both biscuits and chocolates})}{(\text{Total transactions including Biscuits})}$$

$$= 200/400 = 50 \text{ percent.}$$

3. Lift :-

- consider the above example.
- Lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits.
- The mathematical equations of lift are given below.

$$\text{Lift} = (\text{confidence (biscuit - chocolates)}) / (\text{support (Biscuits)})$$

$$= 50/10 = 5$$

* Example :-

Transaction ID	Rice	Pulse	oil	Milk	Apple
T_1	1	1	1	0	0
T_2	0	1	1	1	0
T_3	0	0	0	1	1
T_4	1	1	0	1	0
T_5	1	1	1	0	1
T_6	1	1	1	1	1

- If you implement the threshold assumption, you can figure out that the customer's set of three product is RPO.
- We have considered an easy example to discuss the apriori algorithm in data mining.
- In reality, you find thousands of such combination.

FP-Growth Algorithm

- FP growth algorithm (frequent pattern growth).
- FP growth algorithm is an improvement of apriori algorithm.
- FP growth algorithm used for finding frequent item-set in a transaction database without candidate generation.
- FP growth represents frequent items in frequent pattern trees or FP-tree.

* Advantages of FP growth algorithms :-

1. Faster than apriori algorithm.
2. No candidate generation.
3. only two passes over dataset.

* Disadvantages of FP growth algorithms :-

1. FP tree may not fit in memory.
2. FP tree is expensive to build.

* FP growth Example :-

- Consider the following database (D)
- Set minimum support = 3%

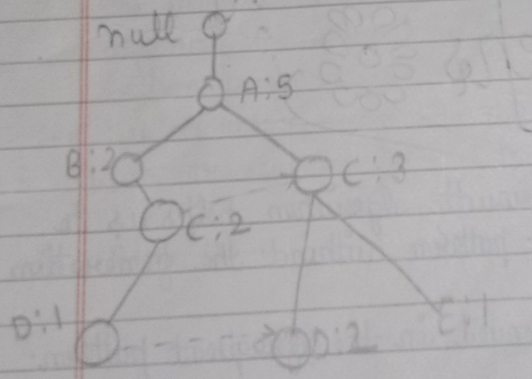
D

Item_ID	Item Set
1	F, A, C, D, G, M, P
2	A, B, C, F, L, M, O
3	B, F, N, O
4	B, K, C, D
5	A, F, C, L, P, M, N

Count Support for each item →

C_1

Item	Support
A	3
B	3
C	4
D	1
F	4
G	1
K	1
L	2
M	3
N	1
O	2
P	3



compare candidate Support Count with minimum Support count and arrange item's with descending Support.

L_1

Item	Support
F	4
C	4
A	3
B	3
M	3
P	3

Compare L_1 table item with D

Ex → In L_1 first item is 'F', Now check if item 'F' is present in table D, then put item 'F' in table L_1. Now 2nd item in L_1 table is 'C'. Now check if item 'C' is present in table D, then put item 'C' in table C_2.

C_2

Item ID	Order Set item
1	F, C, A, M, P
2	F, C, A, B, M
3	F, B
4	C, B
5	F, C, A, M, P

Frequent Pattern Algorithm Steps

- The frequent pattern growth algorithm helps us to discover the frequent pattern without the generation of candidates.
- Let us see the steps followed in the frequent pattern.

Page No. :
Date :

* Step 1 :-

- The 1st step is to scan the entire database to find the possible occurences of the item sets in the database.
- This step is similar to the 1st step of a apriori algorithm.
- No. of 1- itemsets in the database is called support count or frequency of 1- itemset.

* Step 2 :-

- It is to construct the FP tree.
- Create the root of the tree where the root is represented by null.

* Step 3 :-

- The next step is to scan the database once again and observe the transactions.
- Examine the 1st transaction and find the itemset in the database.
- The itemset with the maximum count is taken at the top and the itemset with lower count is taken at the bottom and so on.
- Which means that the branch of the tree is constructed with transaction item sets in descending order of count.

* Step 4 :-

- The next step is to examine the transaction in the database.

- The itemset are sorted in descending order of count.
- If any itemset of this transaction is already present in any other branch, then this transaction branch may share a common prefix to the root of the FP growth algo.
- This means that the common itemset is connected to the new node of another itemset in this transaction.

* Step 5:-

- Next step is that the count of the itemset is increased as it occurs in the transactions.
- Both the common node and new node count is incremented by 1 as they are created and linked according to transactions.

* Step 6:-

- This step is done to mine the FP tree which is created.
- In this step, the lowest node is examined first along with the connections of the lowest nodes.
- The lowest node represents the frequency pattern of length 1.
- Reverse the path in the FP tree.
- These path are called as a conditional pattern base.
- This base is a sub-database consists of prefix paths in the FP tree with the lowest node as suffix.

* Step 7:-

- Next step is to construct a conditional FP tree, which is formed by a count of itemsets in the path.
- The itemset which satisfies the threshold support are considered in the conditional FP tree.

* Step 8:-

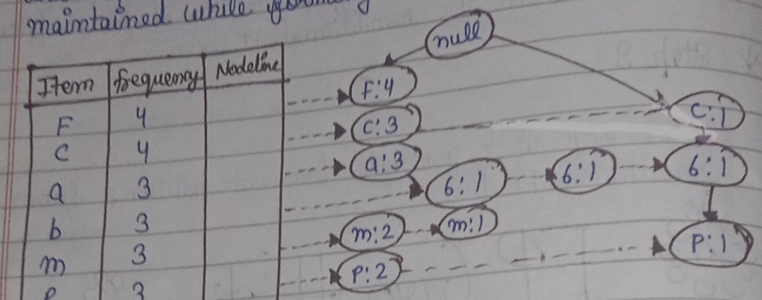
- Final step is to generate frequent patterns from the conditional FP tree.

FP Growth tree

- FP growth represents frequent items in frequent pattern trees which can also be called as FP- tree.
- FP-tree is a tree-like structure which is made with the initial item set of the database.
- The purpose of the FP tree is to find out the most frequent pattern where each node of the FP tree represents an item of the itemset.
- This structure will maintain the association or relation between the itemset.
- The database is separated by using one frequent item.
- This fragmented or separated part is called as pattern fragment.
- The item sets of these fragmented pattern are studied.
- Hence, the search for frequent item set is decreased comparatively.

In FP tree the root node represents null while the lower nodes represents the itemset in database.
The relation of the nodes with the lower nodes that is the item sets with the other item sets are maintained while forming the tree.

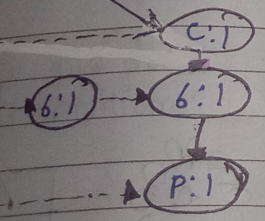
Item	Frequency	Nodelink
F	4	
C	4	
a	3	
b	3	
m	3	
p	3	



* Difference b/w FP growth algo. and Apriori algorithm!-

FP growth algorithm	Apriori algorithm
1. This algorithm is faster than apriori algorithm.	1. This algorithm is slower than FP growth.
2. FP growth is a tree based algo.	2. Apriori is an array based algo.
3. Requires only two data base scan.	3. Requires multiple database scan to generate a Candidate set.

sents null while the
set in database.
the lower nodes
ther item sets are
.



4.	Requires less space.	4. Requires more space.
5.	Consume less time	5. consume more time.
6.	use less memory	6. use more memory.
7.	More accurate	7. Less accurate.
8.	It takes less run time	8. It take more run time.
9.	No Candidate generation	9. use Candidate generation.
10.	It uses a depth search.	10. It uses breadth first-Search.