KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore-641 021
(For the candidates admitted from 2017 onwards)
DEPARTMENT OF COMPUTER SCIENCE, CA & IT

SUBJECT CODE : 17CSU602A SUBJECT : DATA MINING
SEMESTER : VI CLASS : III B.Sc. CS L T P C = 4 0 0 4

Course Objectives

This course introduce students to the basic concepts and techniques of Data Mining, develop skills of using recent data mining software for solving practical problems, gain experience of doing independent study and research

Course Outcomes(COs)

- To introduce students to the basic concepts and techniques of Data Mining.
- To develop skills of using recent data mining software for solving practical problems.
- To gain experience of doing independent study and research.
- Possess some knowledge of the concepts and terminology associated with database systems, statistics, and machine learning

UNIT-I

Overview: Predictive and descriptive data mining techniques

UNIT-II

Supervised and unsupervised learning techniques

UNIT-III

Process of knowledge discovery in databases, pre-processing methods

UNIT-IV

Data Mining Techniques: Association Rule Mining, classification and regression techniques, clustering

UNIT-V

Scalability and data management issues in data mining algorithms, measures of interestingness.

SUGGESTED READINGS

1. Pang-Ning Tan., Michael Steinbach.,& Vipin Kumar. (2005). Introduction to Data Mining. New Delhi: Pearson Education.
2. Richard Roiger., & Michael Geatz. (2003). Data Mining: A Tutorial Based Primer. New Delhi: Pearson Education.
3. Gupta, G.K. (2006). Introduction to Data Mining with Case Studies. New Delhi: PHI.
4. Soman, K. P., Diwakar Shyam., & Ajay, V. (2006). Insight Into Data Mining: Theory And Practice. New Delhi: PHI.

WEB SITES

1. Thedacs.Com
2. Dwreview.Com
3. Pcai.Com
4. Eruditionhome.Com

ESE MARKS ALLOCATION

1.	**Section A** 20 x 1 = 20	20
2.	**Section B** 5 x 2 = 10	10
3.	**Section C** **5 x 6= 30** Either 'A' or 'B' choice	30
	Total	60

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Pollachi Main Road, Eachanari Post, Coimbatore - 641021
(For the candidates admitted from 2017 onwards)
DEPARTMENT OF CS, CA & IT

SUBJECT : DATA MINING **SEMESTER : VI**
SUBJECT CODE: 17CSU602A **CLASS : III B.SC(CS)**

LECTURE PLAN

S.NO	LECTURE DURATION (Hour)	TOPICS TO BE COVERED	SUPPORT MATERIALS
		UNIT I	
1	1	**Introduction:** Data Mining	S1: 13-18
2	1	Data Mining Functionalities, Data Mining model tasks	W2
3	1	Predictive Data Model Techniques	W2
4	1	Descriptive Data Model Techniques	W2
5	1	Predictive analytics Process	W2
6	1	Selecting Modeling Paradigm	W2
7	1	Application of analytics	W3
8	1	**Recapitulation and Discussion of Important Questions**	
		Total no. of Hours Planned for Unit – I	**8 Hrs**
		UNIT – II	
1	1	**Learning,** Machine Learning, Definition for Supervised and unsupervised	W4
2	1	Supervised and unsupervised learning with a Real life Example	W4
3	1	Supervised Machine learning	W5
4	1	Unsupervised Machine learning	W5
5	1	Semisupervised Machine learning	W4
6	1	Analysis of Supervised and Unsupervised Data Mining	W4
7	1	Algorithm choice based on supervised and unsupervised learning	W5
8	1	**Recapitulation and Discussion of Important Questions**	
		Total no. of Hours Planned for Unit – II	**8 Hrs**
		UNIT III	

1	1	**Introduction:** What is Data mining, Types of data, Data mining Life cycle	S2:9-15
2	1	Data mining Techniques	S2:47-48
3	1	Classification of Data mining	S2:61-67
4	1	Process of knowledge discovery in databases (KDD)	S2:67-70
5	1	Pre-processing methods	W6
6	1	Data Reduction, Data Discretization	S2:72-88
7	1	**Recapitulation and Discussion of Important Questions**	
		Total no. of Hours Planned for Unit - III	**7 Hrs**
UNIT IV			
1	1	Overview of Data mining techniques	S3:160-166
2	1	Data mining algorithms and techniques	S3:166-170
3	1	Association Rule mining	S3:69-83,W3
4	1	AIS,SETM, Apriori Algorithm	W7,J1
5	1	Classification and Regression: Classification Techniques	S3:119-125
6	1	Contd…Classification Techniques	S3:125-138 S3:170-171
7	1	Regression Techniques	W8
8	1	Contd…Regression Techniques	W8
9	1	Clustering	W9
10	1	Contd… Clustering	W9
11	1	**Recapitulation and Discussion of Important Questions**	
		Total no. of Hours Planned for Unit - IV	**11 Hrs**
UNIT V			
1	1	Scalability and data management issues in data mining algorithms	W3
2	1	measures of interestingness	W3
3	1	**Recapitulation and discussion of important Questions**	
4	1	**Previous Year ESE Questions Discussion**	
5	1	**Previous Year ESE Questions Discussion**	
6	1	**Previous Year ESE Questions Discussion**	
		Total no. of Hours Planned for Unit - V	**6 Hrs**

		Total Planned Hours	**40 Hrs**

Supporting Material

1. S1: Data Mining,Pieter Adriaans & Dolf zantinge pearson education.
2. Jaiwaei Han, Micheline Kamber, Jian pei, 2012 Data Mining: Concepts and Techniques, Morgan Kaufmann publishers.
3. Margaret H.Dunham, 2008, 3rd edition, Data mining introductory and advancrd topics, person education.

Websites

W1: http://www.webopedia.com
W2:www.ijarcsse.com/
W3:www.tutorialpoint.com/data-mining
W4:Shodhganya.inflibnet.ac.in
W5:https://en.wikipedia.org/machine learning/
W6: http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-3.html
W7:https://www.geeksforgeeks.org/apriori-algorithm/
W8: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
W9:https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Journal:

J1: An Overview of Association Rule Mining Algorithms, Trupti A. Kumbhare et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 927-930

SYLLABUS

UNIT-I

Overview: Predictive and descriptive data mining techniques

Introduction:

Data Mining is an important analytic process designed to explore data. Much like the real-life process of mining diamonds or gold from the earth, the most important task in data mining is to extract non-trivial nuggets from large amounts of data.

Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

Data mining tasks can be classified into two categories: descriptive and predictive.

- Descriptive mining tasks characterize the general properties of the data in the database.
- Predictive mining tasks perform inference on the current data in order to make predictions.

Concept/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. For example, in the Electronics store, classes of items for sale include computers and printers, and concepts of customers include bigSpenders and budgetSpenders.

Data characterization

Data characterization is a summarization of the general characteristics or features of a target class of data.

Data discrimination

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, are patterns that occur frequently in data. There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.

Association analysis

Suppose, as a marketing manager, you would like to determine which items are frequently purchased together within the same transactions.

buys(X,“computer”)=buys(X,“software”) [support=1%,confidence=50%]

where X is a variable representing a customer.Confidence=50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.

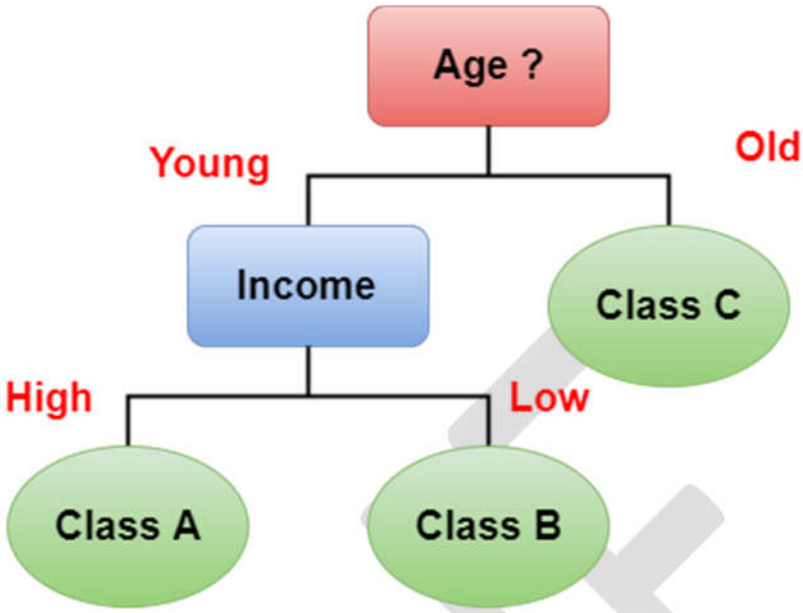
Support=1% means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

Classification and Prediction

Classification is the process of finding a model that describes and distinguishes data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

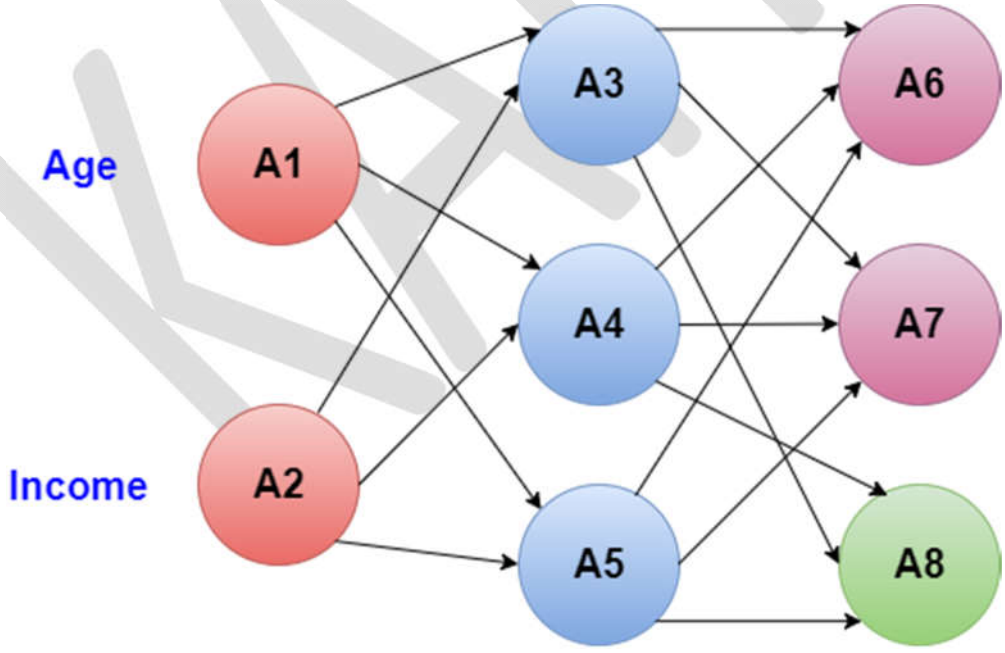
“How is the derived model presented?” The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

A **decision tree** is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.



Decision tree

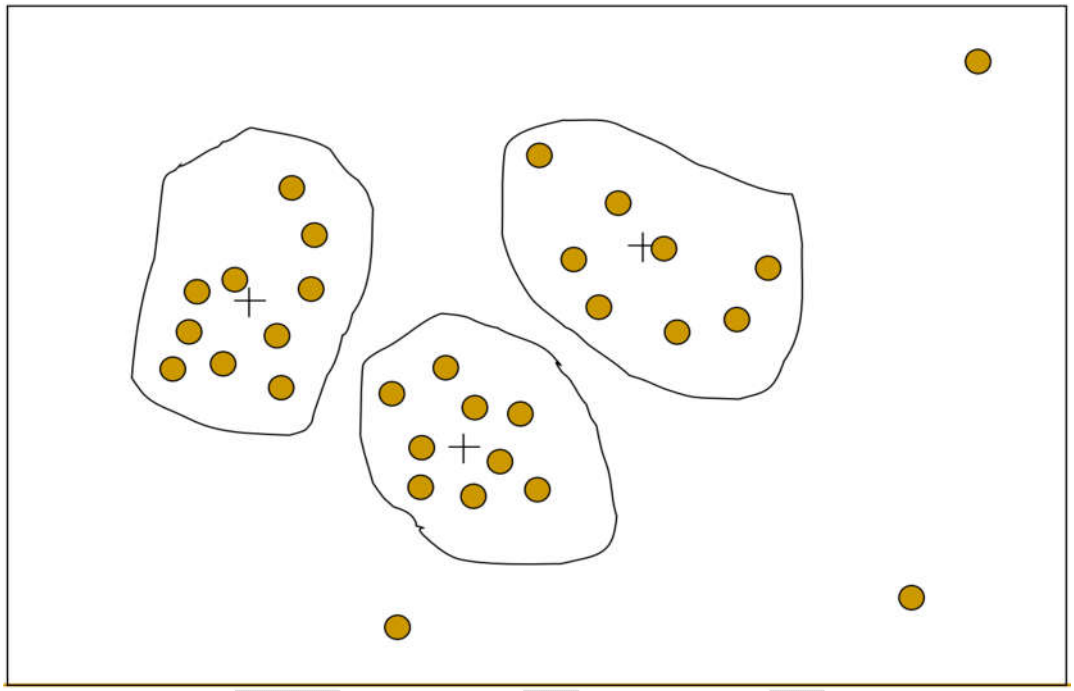
A **neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.



Neural Network

Cluster Analysis

In classification and prediction analyze class-labeled data objects, where as clustering analyzes data objects without consulting a known class label.



Cluster Analysis

The objects are grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions.The analysis of outlier data is referred to as outlier mining.

The two "high-level" primary goals of data mining, in practice, **are prediction and description.**

1. **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
2. **Description** focuses on finding human-interpretable patterns describing the data.

The goals of prediction and description are achieved by using the following primary **data mining tasks**:

1. **Classification** is learning a function that maps (classifies) a data item into one of several predefined classes.
2. **Regression** is learning a function which maps a data item to a real-valued prediction variable.
3. **Clustering** is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data.
    - Closely related to clustering is the task of probability density estimation which consists of techniques for estimating, from data, the joint multi-variate probability density function of all of the variables/fields in the database.
4. **Summarization** involves methods for finding a compact description for a subset of data.
5. **Dependency Modeling** consists of finding a model which describes significant dependencies between variables.
   Dependency models exist at two levels:
    1. The structural level of the model specifies (often graphically) which variables are locally dependent on each other, and
    2. The quantitative level of the model specifies the strengths of the dependencies using some numerical scale.

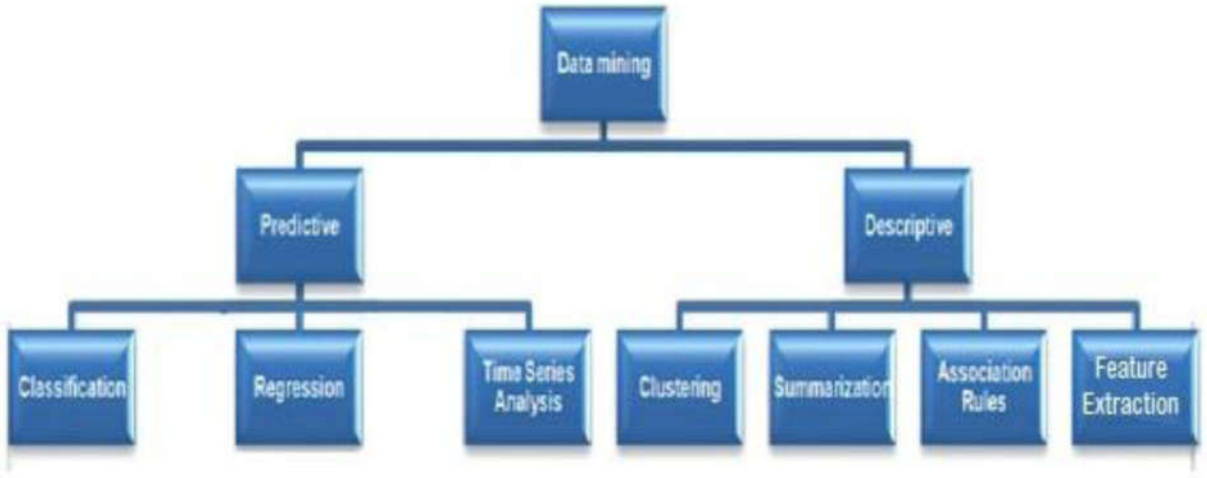
Change and Deviation Detection focuses on discovering the most significant changes in the data from previously measured or normative values.

Introduction to Data Mining Tasks

The data mining tasks can be classified generally into two types based on what a specific task tries to achieve. Those two categories are descriptive tasks and predictive tasks. The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave.

Different Data Mining Tasks

There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining.



create predictive power—using features to predict unknown or future values of the same or other feature—and **create a descriptive power**—find interesting, human-interpretable patterns that describe the data.

Predictive Data Mining:

Prediction task predicts the possible values of missing or future data.

Predictive data mining tasks come up with a model from the available data set that is helpful in predicting unknown or future values of another data set of interest. A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task.

Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest.

For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

a) Classification

- Maps data into predefined groups or classes

- It is also referred as supervised learning because the classes are defined before examining the data.

-E.g whether to make a bank loan and identifying credit risks. -Pattern recognition is a type of classification.

Classification refers to assigning cases into categories based on a predictable attribute. Each case contains a set of attributes, one of which is the *class*attribute (predictable attribute). The task requires finding a model that describes the class attribute as a function of input attributes.

In the College Plans dataset previously described,the *class* is the College Plans attribute with two states: Yes and No. To train a classification model, you need to know the class value of input cases in the training dataset, which are usually the historical data.

Data mining algorithms that require a target to learn against are considered *supervised* algorithms.

Typical classification algorithms include decision trees, neural network, and Naïve Bayes.

b) Regression

It is used to map a data item to a real valued prediction variable. In regression there is a learning of function that does mapping. Regression assumes that the target data fit into some known type of function (e.g linear, logistic,etc); For e.g A professor want to reach a certain level of savings

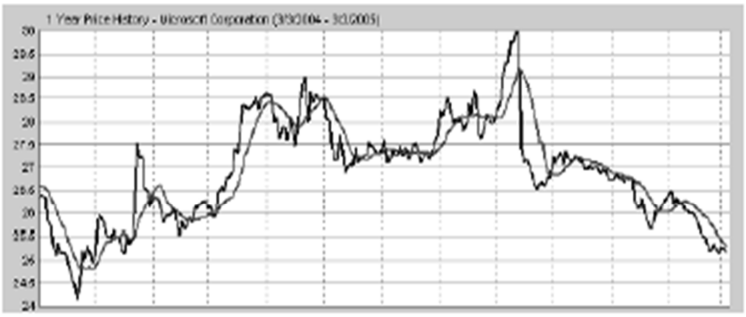
The regression task is similar to classification. The main difference is that the predictable attribute is a continuous number. Regression techniques have been widely studied for centuries in the field of statistics. Linear regression and logistic regression are the most popular regression methods. Other regressiontechniques include regression trees and neural networks. Regression tasks can solve many business problems.

For example, they can be used to predict coupon redemption rates based on the face value, distribution method, and distribution volume, or to predict wind velocities based on temperature, air pressure, and humidity.

c) Time - Series Analysis

The value of an attribute is examined as it varies over time. The values are obtained as evenly spaced(daily,weekly,hourly etc.). The time series plot is used to visualize the time series.

Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series reflects the process being measured and there are certain components that affect the behavior of a process. Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.



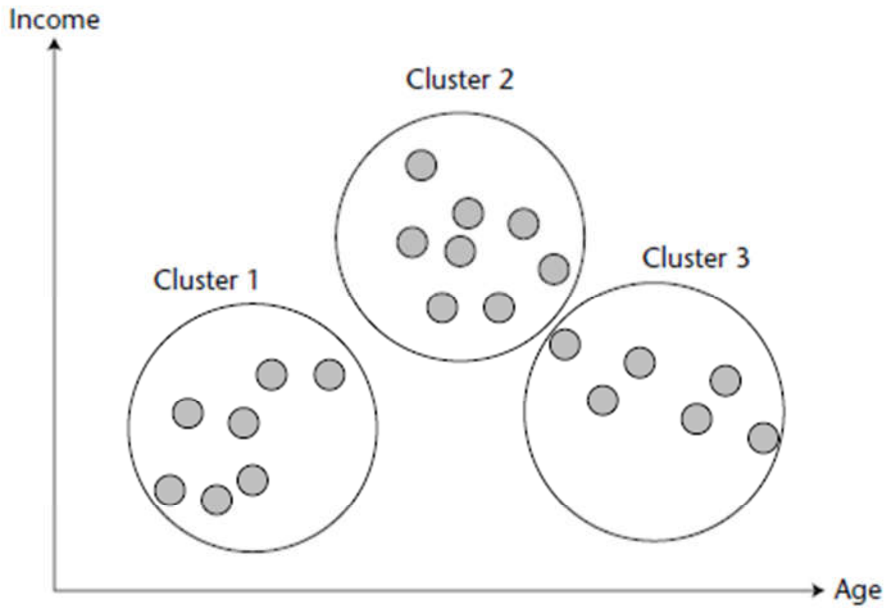
Descriptive Data Mining:

Descriptive data mining tasks usually finds data describing patterns and comes up with new, significant information from the available data set. A retailer trying to identify products that are purchased together can be considered as a descriptive data mining task.

d) Clustering

Clustering is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on. For example, an insurance company can cluster its customers based on age, residence, income etc. This group information will be helpful to understand the customers better and hence provide better customized services.

The clustering algorithm groups the dataset into three segments based on these two attributes. Cluster 1 contains the younger population with a low income. Cluster 2 contains middle-aged customers with higher incomes. Cluster 3 is a group of senior individuals with a relatively low income. Clustering is an *unsupervised* data mining task. No single attribute is used toguide the training process. All input attributes are treated equally. Most clustering algorithms build the model through a number of iterations and stop when the model converges, that is, when the boundaries of these segments are stabilized.



e) Summarization

Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data.

It maps data into subsets with associated simple descriptions. Summarization is also called characterization or generalization. It extracts or derives representative information about the database.

For e.gOne of many criteria used to compare universities by the U.S News and World Report is the average SAT or ACT score.

For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

f) Association

Association is another popular data mining task. Association is also called market basket analysis. A typical association business problem is to analyze a sales transaction table and identify those products often sold in the same shopping basket. The common usage of association is to identify common sets of items (frequent itemsets) and rules for the purpose of cross-selling. In terms of association, each product, or more generally, each attribute/value pair is considered an item.

The association task has two goals:

1) to find frequent itemsets and

2) to find association rules.

Most association type algorithms find frequent itemsets by scanning the dataset multiple times. The frequency threshold (support) is defined by the user before processing the model.

For example, support = 2% means that the model analyzes only items that appear in at least 2% of shopping carts.

Afrequent itemset may look like

{Product = “Pepsi”, Product = “Chips”, Product = “Juice”}.

Each itemset has a size, which is the number of items that it contains. The size of this particular itemset is 3.Apart from identifying frequent itemsets based on support, most association type algorithms also find rules. An association rule has the form A, B => C with a probability, where A, B, C are all frequent item sets. The probability is also referred to as the *confidence* in data mining literature. The probability is a thresholdvalue that the user needs to specify before training an association model.

For example, the following is a typical rule:

Product = “Pepsi”, Product = “Chips” => Product = “Juice” with an 80% probability.

The interpretation of this rule is straightforward. If a customer buys Pepsi and chips, there is an 80% chance that he or she may also buy juice. Each node in the figure represents a product, each edge represents the relationship. The direction of the edge represents the direction of the prediction.

For example, the edge from Milk to Cheese indicates that those who purchase milk might also purchase cheese.

Association discovers the association or connection among a set of items. Association identifies the relationships between objects. Association analysis is used for commodity management, advertising, catalog design, direct marketing etc. A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products. If a retailer finds that beer and nappy are bought together mostly, he can put nappies on sale to promote the sale of beer.

Further divide the data mining task of generating models into the following two approaches:

- Supervised or directed data mining modeling.
- Unsupervised or undirected data mining modeling.

The goal in supervised or directed data mining is to use the availabledata to build a model that describes one particular variable of interest interms of the rest of the available data. The task is to explain the values ofsome particular field. The user selects the target field and directs thecomputer to determine how to estimate, classify or predict its value.

In unsupervised or undirected data mining however variable is sigled outas the target. The goals of predictive and descriptive data mining areachieved by using specific data mining techniques that fall within certainprimary data mining tasks. The goal is rather to establish somerelationship among all the variables in the data. The user asks thecomputer to identify patterns in the data that may be significant.Undirected modeling is used to explain those patters and relationshipsone they have been found

g) Feature Extraction

Feature Extraction extract a new set of features by decomposing the original set of data. This technique describes the data with a number of features far smaller than the number of original attributes. The word feature in the technique is combination of attributes in the data that have special important characteristics of the data [12]. Feature extraction is mostly applicable to latent semantic analysis, data compression, data decomposition and projection, and pattern recognition, etc. Using feature extraction process the speed and effectiveness of supervised learning can also be improved.

For example, feature extraction is used to extract the themes/features of a document collection, where documents are represented by a set of keywords and their frequencies. Each feature is represented by a combination of keywords. The documents can then be expressed from the collection in terms of the discovered themes.

Predictive analytics

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs.

Predictive Analytics Process

1. **Define Project** : Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. **Data Collection** : Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.
3. **Data Analysis** : Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. **Statistics** : Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. **Modelling** : Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.
6. **Deployment**: Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling.
7. **Model Monitoring**: Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

In the current literature, concepts ”knowledge discovery”, ”data mining” and ”machine learning” are often used interchangeably. Sometimes the whole KDD process is called data mining or machine learning, or machine learning is considered as a subdiscipline of data mining. To avoid confusion, we follow a systematic division to descriptive and predictive modelling, which matches well with the classical ideas of data mining and machine learning1 . This approach has several advantages:

- The descriptive and predictive models can often be paired as illustrated in Table 2.1. Thus, descriptive models indicate the suitability of a given predictive model and can guide the search of models (i.e. they help in the selection of the modelling paradigm and the model structure).

- Descriptive and predictive modelling require different validation techniques. Thus the division gives guidelines, how to validate the results.

- Descriptive and predictive models have different underlying assumptions (bias) about good models. This is reflected especially by the score functions, which guide the search.

Table : Examples of descriptive and predictive modeling paradigm pairs. The descriptive models reveal the suitability of the corresponding predictive model and guide the search.

Descriptive paradigm	**Predictive paradigm**
Correlation analysis	Linear regression
Associative rules	Probabilistic rules
Clustering	Classification
Episodes	Markov models

Selecting the modelling paradigm

Selecting the modelling paradigm has a critical role in data modelling. An unsuitable modelling paradigm can produce false, unstable or trivial models, even if we use the best available learning algorithms. Often, we have to try several modelling paradigms, and learn several models, before we find the optimal model for the given data. The number of alternative modelling paradigms can be pruned by analyzing the following factors:

ˆ Properties of data.

ˆ The inductive bias in the modelling paradigm.

ˆ The desired robustness and the representational power of the model.

The problem is to find such a modelling paradigm that the properties of the data match the inductive bias of the paradigm and the resulting models are accurate. We recall that in accuracy, we often have to make a compromise between the robustness and the representational power.

Data properties

The first question is the type and the size of the given data. Some modelling paradigms require only numeric or only categorial data, while others can combine both. Numeric data can always be discretized to categorial, but categorial data cannot be transformed to numeric. The only exception is the transformation of categorial attributes to binary values. However, this transformation increases the number of attributes significantly and the model becomes easily too complex.

In educational domain, the most critical factor of data is the size of the data relative to the model complexity. As a rule of thumb, it is often suggested that we should have at least 5-10 rows of data per each model parameter. The number of model parameters depends on the number of attributes and their domain sizes. Often, we can reduce the number of attributes and/or their domain sizes in the data preprocessing phase. We should select a modelling paradigm, which produces simple models. In practice, we recommend to take the simplest modelling paradigms like linear regression or naive Bayes as a starting point, and analyze whether they can represent the essential features in the data. Only if the data requires higher representational power (e.g. non-linear dependencies, or non-linear class boundaries), more powerful modelling paradigms should be considered.

The representational power is part of data bias in the modelling paradigm. The other assumptions in data bias should be checked as well. For example, our analysis suggests that the educational data is seldom normally distributed. One reason is the large number of outliers – exceptional and unpredictable students. If we want to use a paradigm, which assumes normality, we should first check how large is the deviation from normality and how sensitive the paradigm is to the violation of this assumption.

When the goal is predictive modelling, we recommend to analyze the data properties first by descriptive modelling. According to our view (Figure 3.1), descriptive and predictive tasks are complementary phases of the same modelling process. This view is especially useful in adaptive learning environments, in which the model can be developed through several courses. The existing data is analyzed in the descriptive phase and a desirable modelling paradigm and model family are defined. An initial model is learnt, and applied to new data in the prediction phase. In the same time we can gather new features from users, because the descriptive modelling often reveals also what data we are missing. After the course, the new data is analyzed, and the old model is updated or a

new better model is constructed. As a result, our domain knowledge increases and the predictions improve in each cycle.

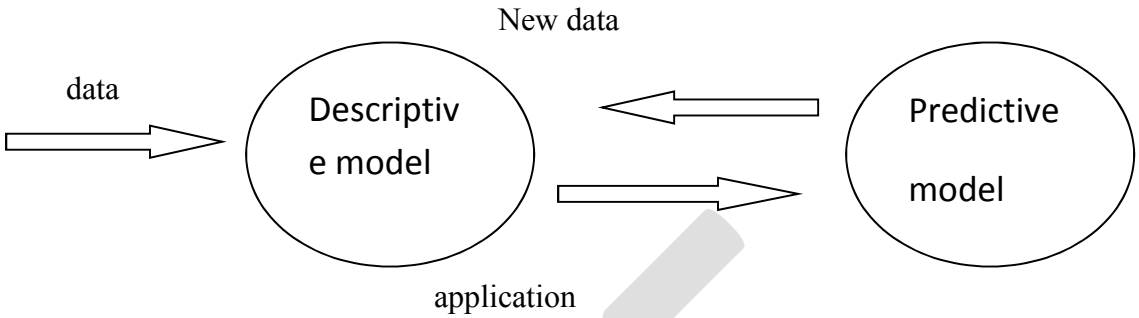


Figure: Iterative process of descriptive and predictive modeling. Descriptive modeling reveals the underlying patterns in the data and guides the selection of the most appropriate modeling paradigm and model family for the predictive modeling. When the predictive model is applied in practice, new data is gathered for new descriptive models.

Applications of Analytics

1) Predictive (forecasting)
2) Descriptive (business intelligence and data mining)

Predictive Analytics

Predictive analytics turns data into valuable, actionable information. Predictive analytics uses data to determine the probable future outcome of an event or a likelihood of a situation occurring.

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Three basic cornerstones of predictive analytics are:

1) Predictive modeling
2) Decision Analysis and Optimization
3) Transaction Profiling

An example of using predictive analytics is optimizing customer relationship management systems. They can help enable an organization to analyze all customer data therefore exposing patterns that predict customer behavior.

Another example is for an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher profitability per customer and stronger customer relationships.

An organization must invest in a team of experts (data scientists) and create statistical algorithms for finding and accessing relevant data. The data analytics team works with business leaders to design a strategy for using predictive information.

Descriptive Analytics

Descriptive analytics looks at data and analyzes past events for insight as to how to approach the future. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Almost all management reporting such as sales, marketing, operations, and finance, uses this type of post-mortem analysis.

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do.

Descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

For example, descriptive analytics examines historical electricity usage data to help plan power needs and allow electric companies to set optimal prices.

POSSIBLE QUESTIONS

2 MARK

1. Define Predictive and descriptive data mining.
2. What is clustering?
3. Mention the different data mining task.
4. What is Data discrimination?
5. What is Data characterization?
6. Define Decision trees.
7. What are the applications of analytics?

6 MARK

1. Explain the Predictive and Descriptive data mining with examples.
2. Explain Data mining functionalities.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore-641021

Department of Computer Science

III B.Sc(CS) (BATCH 2017-2020)

Data Mining (17CSU602A)

PART-A OBJECTIVE TYPE/ MULTIPLE CHOICE QUESTIONS

ONLINE EXAMINATIONS **ONE MARK QUESTIONS**

UNIT-1

S.NO	QUESTIONS	OPT1	OPT2	OPT3	OPT3	ANSWER
1	Which process multiple data sources may be combined	data cleaning	data integration	data selection	data transformation	data integration
2	_____ process is used to retrieve the relevant data from the database	data cleaning	data integration	data selection	data transformation	data selection
3	_____ process is used to identify the truly interesting patterns representing knowledge based on some interesting measures	data cleaning	data integration	data selection	pattern evaluation	pattern evaluation
4	_____ database consists of a file where each record represents a transaction	Relational	Transactional	Object Oriented	Spatial	Transactional
5	A set of _____ that describe the objects. These corresponds to attributes in the entity relationshipp and relational models	Messages	methods	variables	None of the above	variables

6	A set of _____ helps the objects communicate with other objects	Messages	methods	variables	None of the above	Messages
7	_____ database contains spatial related information	Relational	Transactional	Object Oriented	Spatial	Spatial
8	_____ database typically stores relational data that include time-related attributes	Temporal	Transactional	Object Oriented	Spatial	Temporal
9	A _____ database consists of a set of interconnected	Temporal	Heterogeneous	Spatial	Object Oriented	Heterogeneous
10	Many applications involve the generation and analysis of a new kind of data called	Stream	multimedia	Object Oriented	Transactional	Stream
11	_____ datamining task alternatively referred to as affinity analysis	Link analysis	Clustering	Summarization	Segmentation	Link analysis
12	_____ is the process of finding a model that describes and distinguishes data classes	Prediction	Classification	Analysis	Clustering	Classification
13	_____ is the statistical methodology that is most often used for numeric prediction	Prediction	Classification	Regression	Clustering	Regression
14	The data objects that do not comply with the general behavior or model of the data is called	Evolution	Classification	Regression	Outlier	Outlier

15	An approach to solve an estimation problem with incomplete data is ____	Expectation-Maximization	Maximum likelihood estimate	Root mean square	Mean squared error	Expectation-Maximization
16	_____ is the task of discovering interesting patterns from large	database	data warehouse	data mining	None of the above	data mining
17	_____ is a repository for long term storage of data from multiple	database	data warehouse	data mining	None of the above	data warehouse
18	A _____ is a collection of tables each of which is assigned a unique name	Knowledge	Data mining	Data repository	Information	Data mining
19	Use of algorithms to extract information	Prediction	Distributive measures	algebraic measures	None of the above	algebraic measures
20	In Box plots median is marked by _____ with in the box	circle	line	double line	None of the above	line
21	_____ is a random error or variance in a measured variable	smooth	binning	Regression	Noise	Noise
22	_____ works to remove noise from the data	smoothing	aggregation	generalization	normalization	smoothing
23	_____ operation is used for summarization	smoothing	aggregation	generalization	normalization	aggregation
24	_____ is used to scale attribute data so as to fall with in a small specified range	smoothing	aggregation	generalization	normalization	normalization
25	_____ performs a linear transformation on the original data	smoothing	aggregation	min-max normalization	None of the above	min-max normalization

26	_____ databases contain word description for objects	Relational	Transactional	Object Oriented	Text	Text
27	Maps can be represented in _____format in spatial database	lines	circles	vector	scalar	vector
28	_____ database store image ,audio and video data	Relational	Transactional	multimedia	Object Oriented	multimedia
29	_____ is a comparison of the general features of target class data objects	characterization	Classification	discrimination	None of the above	discrimination
30	A _____is a flow chart like tree structure where each node denotes a test on an attribute value	neural network	regression	k means	decision tree	decision tree
31	A _____ when used for classification resembled neural like processing	neural network	regression	k means	decision tree	neural network
32	_____ represents the percentage of transaction from a transactional database that the given rule satisfies in a	confidence	support	completeness	None of the above	support
33	_____ routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data	data cleaning	data integration	data selection	data transformation	data cleaning

34	_____ techniques such as data cube aggregation attribute subset selection, dimensionality reduction numerosity	data integration	data selection	data transformation	data reduction	data reduction
35	KDD stands for	Knowledge discoveryry in database Database	knowledge database iscovery	knowledge discovery in database	knowledge distributee database	knowledge discovery in database
36	In which reduction data are replaced or estimated by alternative, smaller data representation	numerosity	dimensionality	aggregation	None of the above	numerosity
37	In which reduction encoding mechanism are used to reduce the data set size	numerosity	dimensionality	aggregation	None of the above	dimensionality
38	If the original data can be reconstructed from the compressed data with out loss of any information is called	lossy	lossless	discrimination	None of the above	lossless
39	If the original data can be reconstructed from the compressed data with loss of any information is called	lossy	lossless	discrimination	aggregation	lossy
40	In _____ the data are modeled to fit a straight line	linear regression	multiple regression	log-linear regression	Discrete Wavelet Transform	linear regression
41	A technique that consists of graph and algorithms that access that graph in a processing system	linear regression	multiple regression	log-linear regression	logistics	multiple regression

42	Another name for an output attribute.	predictive variable	independent variable	estimated variable	dependent variable	dependent variable
43	Classification problems are distinguished from estimation problems in that	classification problems require the output attribute to be numeric	classification problems require the output attribute to be categorical	classification problems do not allow an output attribute	classification problems are designed to predict future outcome	classification problems require the output attribute to be categorical
44	Which statement is true about prediction problems?	The output attribute must be categorical	The output attribute must be numeric	The resultant model is designed to determine future outcomes	The resultant model is designed to classify current behavior	The resultant model is designed to determine future outcomes
45	Which statement about outliers is true?	Outliers should be identified and removed from a dataset	Outliers should be part of the training dataset but should not be present in the test data	Outliers should be part of the test dataset but should not be present in the training data	The nature of the problem determines how outliers are used	The nature of the problem determines how outliers are used
46	The average squared difference between classifier predicted output and actual output.	mean squared error	root mean squared error	mean absolute error	mean relative error	mean squared error
47	Which of the following problems is best solved using time-series analysis?	Predict whether someone is a likely candidate for having a stroke.	Determine if an individual should be given an unsecured loan.	Develop a profile of a star athlete.	Determine the likelihood that someone will terminate their cell phone contract.	Determine the likelihood that someone will terminate their cell phone contract.

SYLLABUS

UNIT-II

Supervised and unsupervised learning techniques

Introduction to Machine Learning:

DATA:

It can be any unprocessed fact, value, text, sound or picture that is not being interpreted and analyzed.

Data is the most important part of all Data Analytics, Machine Learning, and Artificial Intelligence.

Without data, we can't train any model and all modern research and automation will go unsuccessful.

Big Enterprises are spending loads of money just to gather as much certain data as possible.

Example:

Why did Facebook acquire WhatsApp by paying a huge price of $19 billion? The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information of their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

INFORMATION:

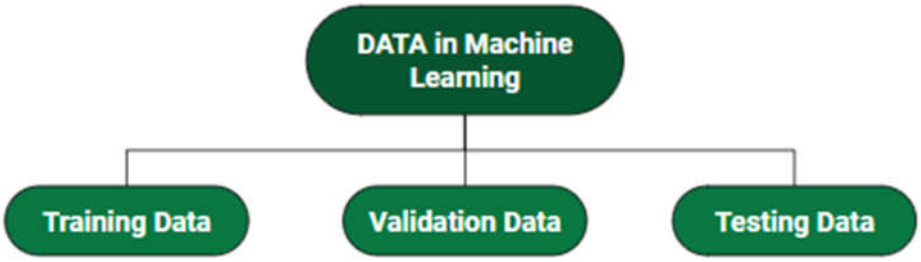
Data that has been interpreted and manipulated and has now some meaningful inference for the users.

KNOWLEDGE:

Combination of inferred information, experiences, learning and insights. Results in awareness or concept building for an individual or organization.

How we split data in Machine Learning?

- **Training Data:** The part of data we use to train our model. This is the data which your model actually sees (both input and output) and learn from.
- **Validation Data:** The part of data which is used to do a frequent evaluation of model, fit on training dataset along with improving involved hyper parameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.
- **Testing Data:** Once our model is completely trained, testing data provides the unbiased evaluation. When we feed in the inputs of testing data, our model will predict some values (without seeing actual output). After prediction, we evaluate our model by comparing it with actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.



Consider an example:

There's a Shopping Mart Owner who conducted a survey for which he has a long list of questions and answers that he had asked from the customers, this list of questions and answers is **DATA**. Now every time when he want to infer anything and can't just go through each and every question of thousands of customers to find something relevant as it would be time-consuming and not helpful. In order to reduce this overhead and time wastage and to make work easier, data is manipulated through software, calculations, graphs etc. as per own convenience, this inference from manipulated data is **Information**. So, Data is must for Information. Now **Knowledge** has its role in differentiating between two individuals having same information. Knowledge is actually not a technical content but is linked to human thought process.

Properties of Data –

1. **Volume:** Scale of Data. With growing world population and technology at exposure, huge data is being generated each and every millisecond.
2. **Variety:** Different forms of data – healthcare, images, videos, audio clippings.
3. **Velocity:** Rate of data streaming and generation.
4. **Value:** Meaningfulness of data in terms of information which researchers can infer from it.
5. **Veracity:** Certainty and correctness in data we are working on.

What is Machine Learning?

Definition 1:

Arthur Samuel, a pioneer in the field of artificial intelligence and computer gaming, coined the term **"Machine Learning"**. He defined machine learning as – **"Field of study that gives computers the capability to learn without being explicitly programmed.**

Definition 2:

A breakthrough in machine learning would be worth ten Microsofts.

— Bill Gates, Former Chairman, Microsoft

Definition 3:

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Basic Difference in ML and Traditional Programming?

- **Traditional Programming:** We feed in DATA (Input) + PROGRAM (logic), run it on machine and get output.
- **Machine Learning:** We feed in DATA (Input) + Output, run it on machine during training and the machine creates its own program (logic), which can be evaluated while testing.

Examples of Machine Learning

There are many examples of machine learning. Here are a few examples of classification problems where the goal is to categorize objects into a fixed set of categories.

1. **Face detection**: Identify faces in images (or indicate if a face is present).
2. **Email filtering**: Classify emails into spam and not-spam.
3. **Medical diagnosis**: Diagnose a patient as a sufferer or non-sufferer of some disease.
4. **Weather prediction**: Predict, for instance, whether or not it will rain tomorrow.

Applications of Machine Learning

Sample applications of machine learning:

- **Web search**: ranking page based on what you are most likely to click on.

- **Computational** biology: rational design drugs in the computer based on past experiments.
- **Finance**: decide who to send what credit card offers to. Evaluation of risk on credit offers. How to decide where to invest money.
- **E-commerce**: Predicting customer churn. Whether or not a transaction is fraudulent.
- **Space exploration**: space probes and radio astronomy.
- **Robotics**: how to handle uncertainty in new environments. Autonomous. Self-driving car.
- **Information extraction**: Ask questions over databases across the web.
- **Social networks**: Data on relationships and preferences. Machine learning to extract value from data.
- **Debugging**: Use in computer science problems like debugging. Labor intensive process. Could suggest where the bug could be.

Key elements of machine learning

There are tens of thousands of machine learning algorithms and hundreds of new algorithms are developed every year.

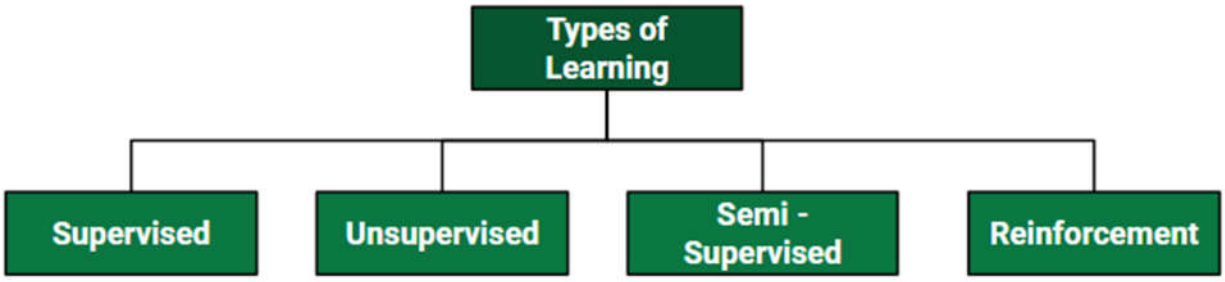
Every machine learning algorithm has three components:

- **Representation**: how to represent knowledge. Examples include decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles and others.
- **Evaluation**: the way to evaluate candidate programs (hypotheses). Examples include accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence and others.
- **Optimization**: the way candidate programs are generated known as the search process. For example combinatorial optimization, convex optimization, constrained optimization.

All machine learning algorithms are combinations of these three components. A framework for understanding all algorithms.

Types of Learning

There are four types of machine learning:



- **Supervised learning**: (also called inductive learning) Training data includes desired outputs. This is spam this is not, learning is supervised.
- **Unsupervised learning**: Training data does not include desired outputs. Example is clustering. It is hard to tell what is good learning and what is not.
- **Semi-supervised learning**: Training data includes a few desired outputs.
- **Reinforcement learning**: Rewards from a sequence of actions. AI types like it, it is the most ambitious type of learning.

Supervised learning is the most mature, the most studied and the type of learning used by most machine learning algorithms. Learning with supervision is much easier than learning without supervision.

Inductive Learning is where we are given examples of a function in the form of data (x) and the output of the function ($f(x)$). The goal of inductive learning is to learn the function for new data (x).

- **Classification**: when the function being learned is discrete.
- **Regression**: when the function being learned is continuous.
- **Probability Estimation**: when the output of the function is a probability.

Supervised Learning:

Supervised learning is when the model is getting trained on a labeled dataset. **Labelled** dataset is one which has both input and output parameters. In this type of learning both training and validation datasets are labeled as shown in the figures below.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Both the above figures have labeled data set –

- **Figure A:** It is a dataset of a shopping store which is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age and salary.
  Input : Gender, Age, Salary

Ouput : Purchased i.e. 0 or 1 ; 1 means yes the customer will purchase and 0 means that customer won't purchase it.

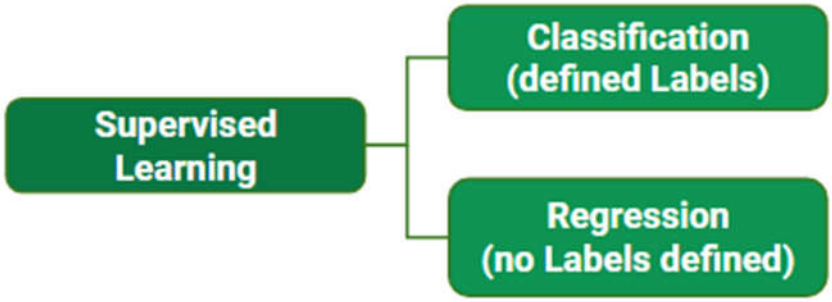
- **Figure B:** It is a Meteorological dataset which serves the purpose of predicting wind speed based on different parameters.

Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction
Output: Wind Speed

Training the system:

While training the model, data is usually split in the ratio of 80:20 i.e. 80% as training data and rest as testing data. In training data, we feed input as well as output for 80% data. The model learns from training data only. We use different machine learning algorithms (which we will discuss in detail in next articles) to build our model. By learning, it means that the model will build some logic of its own.

Once the model is ready then it is good to be tested. At the time of testing, input is fed from remaining 20% data which the model has never seen before, the model will predict some value and we will compare it with actual output and calculate the accuracy.



Types of Supervised Learning:

1. **Classification:** It is a Supervised Learning task where output is having defined labels (discrete value). For example in above Figure A, Output – Purchased has defined labels i.e. 0 or 1 ; 1 means the customer will purchase and 0 means that customer won't purchase. The goal here is to predict discrete values belonging to a particular class and evaluate on the basis of accuracy.

It can be either binary or multi class classification. In **binary** classification, model predicts either 0 or 1 ; yes or no but in case of **multi class** classification, model predicts more than one class.

Example: Gmail classifies mails in more than one classes like social, promotions, updates, forum.

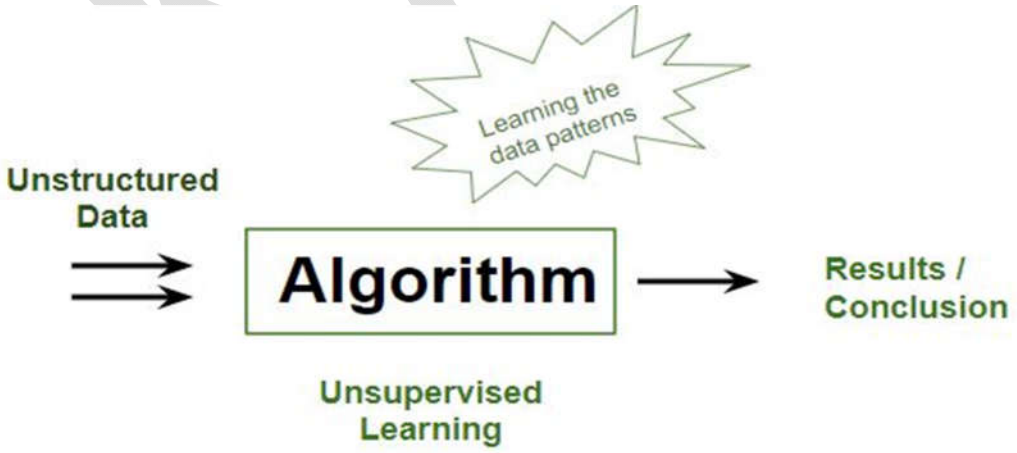
2. **Regression:** It is a Supervised Learning task where output is having continuous value.

Example in above Figure B, Output – Wind Speed is not having any discrete value but is continuous in the particular range. The goal here is to predict a value as much closer to actual output value as our model can and then evaluation is done by calculating error value. The smaller the error the greater the accuracy of our regression model.

Example of Supervised Learning Algorithms:

- Linear Regression
- Nearest Neighbor
- Guassian Naive Bayes
- Decision Trees
- Support Vector Machine (SVM)
- Random Forest

Unsupervised Learning:



It's a type of learning where we don't give target to our model while training i.e. training model has only input parameter values. The model by itself has to find which way it can learn.

Data-set in Figure A is mall data that contains information of its clients that subscribe to them.

Once subscribed they are provided a membership card and so the mall has complete information about customer and his/her every purchase.

Now using this data and unsupervised learning techniques, mall can easily group clients based on the parameters we are feeding in.

CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure A

Training data we are feeding is –

- **Unstructured data:** May contain noisy(meaningless) data, missing values or unknown data

- **Unlabeled data:** Data only contains value for input parameters, there is no targeted value (output). It is easy to collect as compared to labeled one in supervised approach.

Types of Unsupervised Learning:

- **Clustering:** Broadly this technique is applied to group data based on different patterns, our machine model finds. For example in above figure we are not given output parameter value, so this technique will be used to group clients based on the input parameters provided by our data.
- **Association:** This technique is a rule based ML technique which finds out some very useful relations between parameters of a large data set. For e.g. shopping stores use algorithms based on this technique to find out relationship between sale of one product w.r.t to others sale based on customer behavior. Once trained well, such models can be used to increase their sales by planning different offers.

Some algorithms:

- K-Means Clustering
- DBSCAN – Density-Based Spatial Clustering of Applications with Noise
- BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies
- Hierarchical Clustering

Semi-supervised Learning

Its working lies between Supervised and Unsupervised techniques. We use these techniques when we are dealing with a data which is a little bit labeled and rest large portion of it is unlabeled.

We can use unsupervised technique to predict labels and then feed these labels to supervised techniques.

This technique is mostly applicable in case of image data-sets where usually all images are not labeled.

Reinforcement Learning:

In this technique, model keeps on increasing its performance using a Reward Feedback to learn the behavior or pattern. These algorithms are specific to a particular problem e.g. Google Self Driving car, AlphaGo where a bot competes with human and even itself to getting better and better performer of Go Game. Each time we feed in data, they learn and add the data to its knowledge that is training data. So, more it learns the better it get trained and hence experienced.

- Agents observe input.
- Agent performs an action by making some decisions.
- After its performance, agent receives reward and accordingly reinforce and the model stores in state-action pair of information.

Some algorithms:

- Temporal Difference (TD)
- Q-Learning

- Deep Adversarial Networks

Differences between Supervised Learning and Unsupervised Learning

(http://www.differencebetween.net/technology/differences-between-supervised-learning-and-unsupervised-learning/)

1. **Input Data in Supervised Learning and Unsupervised Learning**

The primary difference between supervised learning and unsupervised learning is the data used in either method of machine learning. It is worth noting that both methods of machine learning require data, which they will analyze to produce certain functions or data groups. However, the input data used in supervised learning is well known and is labeled. This means that the machine is only tasked with the role of determining the hidden patterns from already labeled data. However, the data used in unsupervised learning is not known or labeled. It is the work of the machine to categorize and label the raw data before determining the hidden patterns and functions of the input data.

2. **Computational Complexity in Supervised Learning and Unsupervised Learning**

Machine learning is a complex affair and any person involved must be prepared for the task ahead. One of the stands out differences between supervised learning and unsupervised learning is computational complexity. Supervised learning is said to be a complex method of learning while unsupervised method of learning is less complex. One of the reasons that makes supervised learning affair is the fact that one has to understand and label the inputs while in unsupervised learning, one is not required to understand and label the inputs. This explains why many people have preferred unsupervised learning as compared to the supervised method of machine learning.

3. **Accuracy of the Results of Supervised Learning and Unsupervised Learning**

The other prevailing difference between supervised learning and unsupervised learning is the accuracy of the results produced after every cycle of machine analysis. All the results generated from supervised method of machine learning are more accurate and reliable as compared to the results generated from the unsupervised method of machine learning. One of the factors that explain why supervised method of machine learning produces accurate and reliable results is because the input data is well known and labeled which means that the machine will only

analyze the hidden patterns. This is unlike unsupervised method of learning where the machine has to define and label the input data before determining the hidden patterns and functions.

4. Number of Classes in Supervised Learning and Unsupervised Learning

It is also worth noting that there is a significant difference when it comes to the number of classes. It is worth noting that all the classes used in supervised learning are known which means that also the answers in the analysis are likely to be known. The only goal of supervised learning is therefore to determine the unknown cluster. However, there is no prior knowledge in unsupervised method of machine learning. In addition, the numbers of classes are not known which clearly means that no information is known and the results generated after the analysis cannot be ascertained. Moreover, the people involved in unsupervised method of learning are not aware of any information concerning the raw data and the expected results.

5. Real Time Learning in Supervised Learning and Unsupervised Learning

Among other differences, there exist the time after which each method of learning takes place. It is important to highlight that supervised method of learning takes place off-line while unsupervised method of learning takes place in real time. People involved in preparation and labeling of the input data do so off-line while the analysis of the hidden pattern is done online which denies the people involved in machine learning an opportunity to interact with the machine as it analyzes the discrete data. However, unsupervised method of machine learning takes place in real time such that all the input data is analyzed and labeled in the presence of learners which helps them to understand different methods of learning and classification of raw data. Real time data analysis remains to be the most significant merit of unsupervised method of learning.

TABLE SHOWING DIFFERENCES BETWEEN SUPERVISED LEARNING AND UNSUPERVISED LEARNING: COMPARISON CHART

	Supervised Learning	**Unsupervised Learning**
Input Data	Uses Known and Labeled Input Data	Uses Unknown Input Data
Computational Complexity	Very Complex in Computation	Less Computational Complexity

Real Time	Uses off-line analysis	Uses Real Time Analysis of Data
Number of Classes	Number of Classes is Known	Number of Classes is not Known
Accuracy of Results	Accurate and Reliable Results	Moderate Accurate and Reliable Results

Summary of Supervised Learning and Unsupervised Learning

- Data mining is becoming an essential aspect in the current business world due to increased raw data that organizations need to analyze and process so that they can make sound and reliable decisions.
- This explains why the need for machine learning is growing and thus requiring people with sufficient knowledge of both supervised machine learning and unsupervised machine learning.

It is worth understanding that each method of learning offers its own advantages and disadvantages. This means that one has to be conversant with both methods of machine learning before determine which method one will use to analyze data.

POSSIBLE QUESTIONS

2 MARK

1. Define Data, Information and Knowledge.
2. How We Split Data In Machine Learning?
3. Mention the Properties Of Data.
4. What Is Machine Learning?
5. Basic Difference between Machine Learning and Traditional Programming?
6. List any four Examples of Machine Learning.
7. List the Applications Of Machine Learning.
8. What are Types of Learning?
9. What are Key Elements of Machine Learning?

6 MARK

1. Explain Supervised Learning: with example.
2. State Differences between Supervised Learning and Unsupervised Learning.
3. Write in detail about Unsupervised Learning with example.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore-641021

Department of Computer Science

III BSc(CS) (BATCH 2017-2020)

Data Mining (17CSU602A)

PART-A OBJECTIVE TYPE/ MULTIPLE CHOICE QUESTIONS

ONLINE EXAMINATIONS **ONE MARK QUESTIONS**

UNIT-2

sno	Question	option1	option2	option3	option4	Answer
1	_____ approximates discrete multidimensional probability distributions	linear regression	multiple regression	log-linear regression	aggregation	log-linear regression
2	In _____ the width of each bucket range is uniform	Equal-width	Equal – frequency	V-optimal	MaxDiff	Equal-width
3	In _____ histogram the bucket created so that roughly the frequency of each bucket is constant	Equal-width	Equal – frequency	V-optimal	MaxDiff	Equal – frequency
4	Classification frequently performed by simply applying ____	Data	Information	Knowledge of the data	Report	Knowledge of the data
5	In _____ histogram the difference between each pair of adjacent values	Equal-width	Equal – frequency	V-optimal	MaxDiff	MaxDiff
6	_____ can be used as a data reduction technique	Equal-width	Equal – frequency	V-optimal	sample	sample
7	SRSWOR stands for	Simple Random sample without replacement	Simple Random Sample with replacement	Sort Random simple with replacement	None	Simple Random sample without replacement
8	SRSWR stands for	Simple Random sample without replacement	Simple Random Sample with replacement	Sort Random simple with replacement	None	Simple Random Sample with replacement
9	_____ is not a classification alforithm category	Decision tree	Distance	Neural network	Query	Query

10	A _____ is a collection of data objects that are similar to one another	cluster	outlier	predictio n	none	cluster
11	_____ represent object by object structure	Data Matrix	Dissimila r matrix	Singular matrix	None	Dissimila r matrix
12	_____ represent object by variable structure	Data Matrix	Dissimila r matrix	Singular matrix	None	Data Matrix
13	Data Matrix is often called	one mode	two mode	three mode	none	two mode
14	Dissimilarity matrix is often called	one mode	two mode	three mode	none	one mode
15	Manhattan distance is also called as	Euclidea n distance	Minkows ki distance	city block distance	none	city block distance
16	_____ us a generalization of both Euclidean distance and Manhattan distance	Euclidea n distance	Minkows ki distance	city block distance	none	Minkows ki distance
17	Association rule is also called as _____	Market basket analysis	Euclidea n distance	Minkows ki distance	city block distance	Market basket analysis
18	In Association rule X- Y , X represents	Antecede nts	conseque nts	confidenc e	predictab ility	Antecede nts
19	In Association rule X- Y , Y represents	Antecede nts	conseque nts	confidenc e	predictab ility	conseque nts
20	A frequent item set Y is _____	minimal	maximal	predictio n	none	maximal
21	Expansion of DHP is _____	Direct Hashing and Pruning	Dynamic itemset Counting	Decision hashing and pruning	None	Direct Hashing and Pruning
22	Expansion of DIC is_____	Direct Hashing and Pruning	Dynamic item set Counting	Decision hashing and pruning	None	Dynamic item set Counting
23	_____algorithm does not generate candidate itemset	Apriori algorithm	Direct hashing and pruning	FP-Growth	None	FP-Growth
24	A _____ has only two states 0 or 1	Trinary variable	Binary variable	Unary Variable	none	Binary variable
25	A binary variable is _____ if the outcomes of the states are not equally important	symmetri c	asymmetr ic	different	none	asymmetr ic
26	An if-then approach to perform classification is called ___ algorithm	Rule based	Decision tree based	different	none	symmetri c

27	Moving from sink to source nodes in a learning technique is ___	Propagation	Back propagation	Jaccard	differential	Jaccard
28	Which algorithm comes under Hierarchical methods	K-means	K-mediods	Categorical	Ordinal	Categorical
29	SPRINT abbrevation for Scalable PaRallelizable Induction of decision____	Trees	Tables	Categorical	Ordinal	Ordinal
30	_____ method quantize the object space into a finite number of cells that form a grid structure	Partitioning	Hierarchical	Categorical	Grid-based	Grid-based
31	_____ method hypothesize a model for each of the clusters and find the best fit of the data to the given model	Model Based Clustering	Hierarchical	Neural network based	Distance based	Model Based Clustering
32	CART is an abbrevation for classification and __ trees	Relation	Regression	Vertex	Arcs	Regression
33	____ technique minimizes the expected number of comparisons	CART	C45	BIRCH	DBSCAN	BIRCH
34	Decision tree approach divides the search space into ___ regions	Two	Squared	Techniques	Threshold	Two
35	The divisive approach also called _____ approach	top-down	bottom up	Density based	Grid-based	Grid-based
36	Construction of a tree model to sovle a classification problem is ___	Neural network	decision tree	Density based	Model –based	Model –based
37	____ uses information gain as its attribute selection measure	partition rule	decision rule	Repeated	Reciprocal	decision rule
38	A _____ is a heuristic for selecting the splitting criterion that best separates a given data partition	partition rule	decision rule	ID3	Neural network	ID3
39	The cost complexity pruning algorithm used in _____	CART	ID3	Triangular	Rectangle	Rectangle
40	DSS stands for _____	Direct Support System	Decision Support System	partition	none	Decision Support System
41	Common classification scheme based on use of distance measure	K-nearest neighbor	Baye's theorem	Association rule	Classification by decision tree	Classification by decision tree
42	Relationship between false positives and true positives is represented as ___ curve	Operating characteristic	Regression	splitting rule	ID3	ID3

43	In which stage companies began to build more sophisticated systems intended to	Decision Support system	Information system	splitting rule	attribute selection measure	attribute selection measure
44	Operational systems are _____ systems	Decision Support system	information system	splitting rule	none	Decision Support system
45	Regression is a ___ based algorithm	Statistical	Rule	Dynamic Support system	none	Rule
46	In which stage companies began to build more sophisticated systems intended to	Decision Support system	information system	Characteristic curve	Activation function	Decision Support system
47	Operational systems are _____ systems	Decision Support system	information system	Missing data	Relationship	Operating characterisitic
48	Database designed for _____ task	Historical	analytical	Special Extract Programs	none	analytical
49	Data used to build a data mining model	validation data	training data	test data	hidden data	training data
50	Supervised learning and unsupervised clustering both require at least one	hidden attribute	output attribute	input attribute	categorical attribute	input attribute
51	Supervised learning differs from unsupervised clustering in that supervised learning requires	at least one input attribute	input attributes to be categorical	at least one output attribute	ouput attriubutes to be categorical	at least one output attribute
52	Which is the best approach of this problem:Determine whether a credit card transaction is valid or fraudulent	supervised learning	unsupervised clustering	data query	predictive	supervised learning
53	Which is the best approach of this problem:What is the average weekly salary of all female employees under forty years of age?	supervised learning	unsupervised clustering	data query	Descriptive	data query
54	Which is the best approach of this problem:Do meaningful attribute relationships exist in a database containing information about credit card customers?	supervised learning	unsupervised clustering	data query	regression	unsupervised clustering

55	Which is the best approach of this problem:What attribute similarities group customers holding one or several insurance policies?	supervised learning	unsupervised clustering	data query	classification	supervised learning
56	Which of the following is a common use of unsupervised clustering?	detect outliers and determine if meaningful relationships can be found in a dataset	determine a best set of input attributes for supervised learning	evaluate the likely performance of a supervised learner model	All of a,b,c, are common uses of unsupervised clustering	All of a,b,c, are common uses of unsupervised clustering
57	_____ need the control of human operation during their execution	noise	data minig	supervised algorithm	none	supervised algorithm
58	Data used to optimize the parameter settings of a supervised learner model	training	test	verification	validation	validation
59	We have performed a supervised classification on a dataset containing 100 test set instances Eighty of the test set instances were correctly classified The 95% test set accuracy confidence boundaries are:	76% and 84%	72% and 88%	78% and 82%	70% and 90%	72% and 88%
60	Bootstrapping allows us to	choose the same training instance several times	choose the same test set instance several times	build models with alternative subsets of the training data several times	test a model with alternative subsets of the test data several times	choose the same training instance several times

61	We have built and tested two supervised learner models¾M1 and M2 We compare the test set accuracy of the models using the classical hypothesis testing paradigm using a 95% confidence setting The computed value of P is 253 What can we say about this result?	Model M1 performs significantly better than M2	Model M2 performs significantly better than M1	Both models perform at the same level of accuracy	The models differ significantly in their performance	The models differ significantly in their performance
62	Unsupervised evaluation can be internal or external Which of the following is an internal method for evaluating alternative clusterings produced by the K-Means algorithm?	Use a production rule generator to compare the rule sets generated for each clustering	Compute and compare class resemblance scores for the clusters formed by each clustering	Compare the sum of squared error differences between instances and their corresponding cluster centers for each alternative clustering	Create and compare the decision trees determined by each alternative clustering	Compare the sum of squared error differences between instances and their corresponding cluster centers for each alternative clustering
63	A two-layered neural network used for unsupervised clustering	backpropagation network	Kohonen network	perceptron network	aggomerative network	Kohonen network
64	This type of supervised network architecture does not contain a hidden layer	backpropagation	perceptron	self-organizing map		perceptron
65	This supervised learning technique can process both numeric and categorical input attributes	linear regression	Bayes classifier	logistic regression	backpropagation learning	Bayes classifier
66	This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration	agglomerative clustering	conceptual clustering	K-Means clustering	AIS	K-Means clustering

SYLLABUS

UNIT-III

Process of knowledge discovery in databases, pre-processing methods

What is Data Mining?

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.

The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc.

Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc.

Types of Data

Data mining can be performed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
- Object-oriented and object-relational databases
- Transactional and Spatial databases
- Heterogeneous and legacy databases
- Multimedia and streaming database
- Text databases
- Text mining and Web mining

Data Mining Implementation Process / Data mining Life cycle

Let's study the Data Mining implementation process in detail

Business understanding:

In this phase, business and data-mining goals are established.

- First, you need to understand business and client objectives. You need to define what your client wants (which many times even they do not know themselves)
- Take stock of the current data mining scenario. Factor in resources, assumption, constraints, and other significant factors into your assessment.
- Using business objectives and current scenario, define your data mining goals.
- A good data mining plan is very detailed and should be developed to accomplish both business and data mining goals.

Data understanding:

In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

- First, data is collected from multiple data sources available in the organization.
- These data sources may include multiple databases, flat filer or data cubes. There are issues like object matching and schema integration which can arise during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily. For example, table A contains an entity named cust_no whereas another table B contains an entity named cust-id.
- Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not. Here, Metadata should be used to reduce errors in the data integration process.
- Next, the step is to search for properties of acquired data. A good way to explore the data is to answer the data mining questions (decided in business phase) using the query, reporting, and visualization tools.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

Data preparation:

In this phase, data is made production ready.

- The data preparation process consumes about 90% of the time of the project.
- The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).
- Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values.

For example, for a customer demographics profile, age data is missing. The data is incomplete and should be filled. In some cases, there could be data outliers. For instance, age has a value 300. Data could be inconsistent. For instance, name of the customer is different in different tables.

Data transformation operations change the data to make it useful in data mining. Following transformation can be applied

Data transformation:

Data transformation operations would contribute toward the success of the mining process.

- **Smoothing:** It helps to remove noise from the data.
- **Aggregation:** Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.
- **Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.
- **Normalization:** Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.
- **Attribute construction**: these attributes are constructed and included the given set of attributes helpful for data mining.

The result of this process is a final data set that can be used in modeling.

Modelling

In this phase, mathematical models are used to determine data patterns.

- Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

Evaluation:

In this phase, patterns identified are evaluated against the business objectives.

- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- A go or no-go decision is taken to move the model in the deployment phase.

Deployment:

In the deployment phase, you ship your data mining discoveries to everyday business operations.

- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.
- A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.
- A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

Data Mining Techniques

1. Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction:

Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

Classification of Data Mining System

Data mining systems can be categorized according to various criteria as follows:

1. Classification of data mining systems according to the type of data sources mined:

This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

2. Classification of data mining systems according to the database involved:

This classification based on the data model involved such as relational database, object oriented database, data warehouse, transactional database, etc.

3. Classification of data mining systems according to the kind of knowledge discovered:

This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

4. Classification of data mining systems according to mining techniques used:

This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

Process of Knowledge Discovery in Databases (KDD)

Why we need Data Mining?

Volume of information is increasing everyday that we can handle from business transactions, scientific data, sensor data, Pictures, videos, etc. So, we need a system that will be capable of extracting essence of information available and that can automatically generate report, views or summary of data for better decision-making.

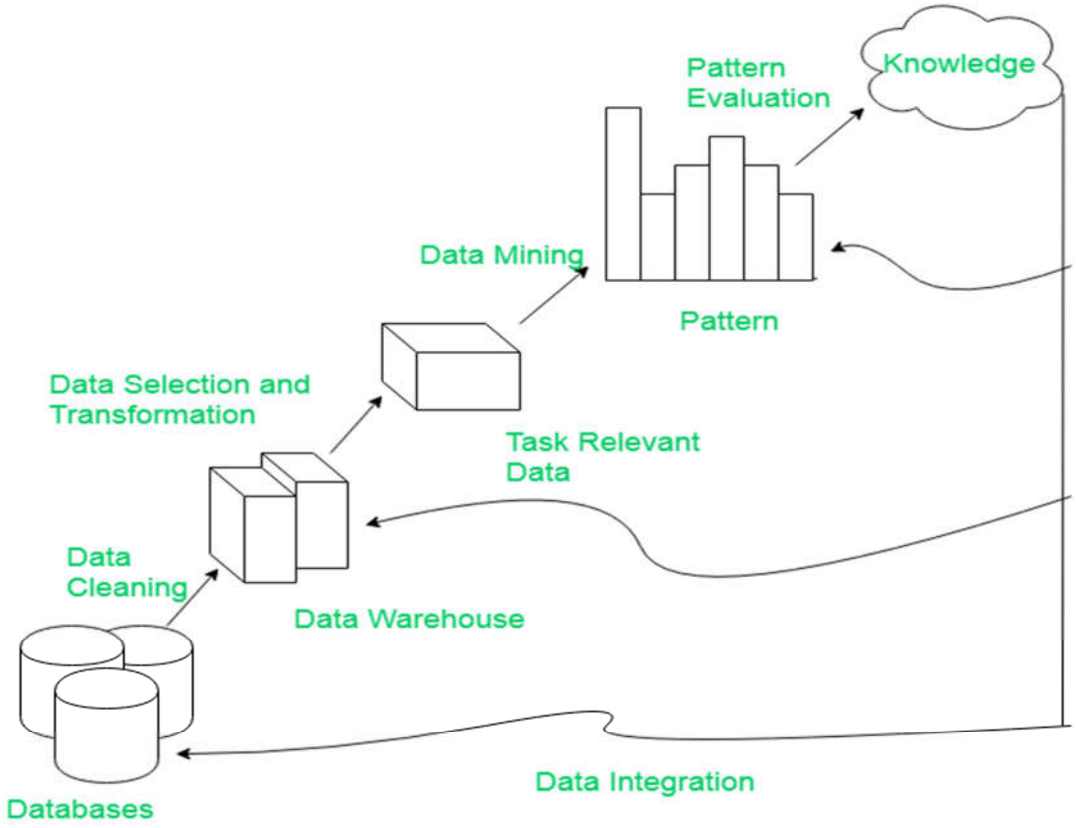
Why Data Mining is used in Business?

Data mining is used in business to make better managerial decisions by:

- Automatic summarization of data
- Extracting essence of information stored.
- Discovering patterns in raw data.

Data mining also known as Knowledge Discovery in Databases refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

Steps Involved in KDD Process:



KDD process

1. Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.

- Cleaning in case of Missing values.
- Cleaning noisy data, where noise is a random or variance error.
- Cleaning with Data discrepancy detection and Data transformation tools.

2. Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source (DataWarehouse).

- Data integration using Data Migration tools.
- Data integration using Data Synchronization tools.
- Data integration using ETL(Extract-Load-Transformation) process.

3. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- Data selection using Neural network.
- Data selection using Decision Trees.
- Data selection using Naive bayes.
- Data selection using Clustering, Regression, etc.

4. Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two step process:

- Data Mapping: Assigning elements from source base to destination to capture transformations.
- Code generation: Creation of the actual transformation program.

5. Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

- Transforms task relevant data into patterns.
- Decides purpose of model using classification or characterization.

6. Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.

- Find interestingness score of each pattern.
- Uses summarization and Visualization to make data understandable by user.

7. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

- Generate reports.
- Generate tables.
- Generate discriminant rules, classification rules, characterization rules, etc.

Note:

- KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.
- Preprocessing of databases consists of Data cleaning and Data Integration.

Pre-processing methods

Why preprocessing?

1. Real world data are generally

    - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - Noisy: containing errors or outliers
    - Inconsistent: containing discrepancies in codes or names

2. Tasks in data preprocessing

    - Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
    - Data integration: using multiple databases, data cubes, or files.
    - Data transformation: normalization and aggregation.
    - Data reduction: reducing the volume but producing the same or similar analytical results.
    - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Data cleaning

1. Fill in missing values (attribute or class value):

    - Ignore the tuple: usually done when class label is missing.
    - Use the attribute mean (or majority nominal value) to fill in the missing value.
    - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
    - Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

2. Identify outliers and smooth out noisy data:

    - Binning
        - Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);
        - Then smooth by bin means, bin median, or bin boundaries.
    - Clustering: group values in clusters and then detect and remove outliers (automatic or manual)
    - Regression: smooth by fitting the data into regression functions.

3. Correct inconsistent data: use domain knowledge or expert decision.

Data transformation

1. Normalization:
   - Scaling attribute values to fall within a specified range.
     - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
   - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev
2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalization: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes.

Data reduction

1. Reducing the number of attributes
   - Data cube aggregation: applying roll-up, slice or dice operations.
   - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space (see Lecture 5: Attribute-oriented analysis).
   - Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data..
2. Reducing the number of attribute values
   - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
   - Clustering: grouping values in clusters.
   - Aggregation or generalization
3. Reducing the number of tuples
   - Sampling

Discretization and generating concept hierarchies

1. Unsupervised discretization - class variable is not used.
   - Equal-interval (equiwidth) binning: split the whole range of numbers in intervals with equal size.
   - Equal-frequency (equidepth) binning: use intervals containing equal number of values.
2. Supervised discretization - uses the values of the class variable.
   - Using class boundaries. Three steps:
     - Sort values.
     - Place breakpoints between values belonging to different classes.
     - If too many intervals, merge intervals with equal or similar class distributions.
   - Entropy (information)-based discretization. Example:

- Information in a class distribution:
  - Denote a set of five values occurring in tuples belonging to two classes (+ and -) as [+,+,+,-,-]
  - That is, the first 3 belong to "+" tuples and the last 2 - to "-" tuples
  - Then, Info([+,+,+,-,-]) = -(3/5)*log(3/5)-(2/5)*log(2/5) (logs are base 2)
  - 3/5 and 2/5 are relative frequencies (probabilities)
  - Ignoring the order of the values, we can use the following notation: [3,2] meaning 3 values from one class and 2 - from the other.
  - Then, Info([3,2]) = -(3/5)*log(3/5)-(2/5)*log(2/5)
- Information in a split (2/5 and 3/5 are weight coefficients):
  - Info([+,+],[+,-,-]) = (2/5)*Info([+,+]) + (3/5)*Info([+,-,-])
  - Or, Info([2,0],[1,2]) = (2/5)*Info([2,0]) + (3/5)*Info([1,2])
- Method:
  - Sort the values;
  - Calculate information in all possible splits;
  - Choose the split that minimizes information;
  - Do not include breakpoints between values belonging to the same class (this will increase information);
  - Apply the same to the resulting intervals until some stopping criterion is satisfied.

3. Generating concept hierarchies: recursively applying partitioning or discretization methods.

POSSIBLE QUESTIONS

2 MARK

1. What is Data Mining?
2. Mention some Types of Data.
3. Define Data mining Life cycle.
4. Why we need Data Mining?
5. Why Data Mining is used in Business?
6. Why preprocessing?
7. What is KDD?

6 MARK

1. Explain the Steps Involved in KDD Process.
2. Elaborate Classification of Data Mining System.
3. Explain Data Mining Techniques in detail.
4. Enlighten the Data Mining Implementation Process / Data mining Life cycle.

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act, 1956)
Coimbatore-641021
Department of Computer Science
III BSc(CS) (BATCH 2017-2020)
Data Mining (17CSU602A)
PART-A OBJECTIVE TYPE/ MULTIPLE CHOICE QUESTIONS
ONLINE EXAMINATIONS **ONE MARK QUESTIONS**
UNIT-3

S.NO	Questions	option1	option2	option3	option4	Answers
1	Which system the data is stored based on individual applications	Subject oriented data	Integrated data	Time Stamped data	Nonvolatile data	Subject oriented data
2	In which system data is pull together from the various applications	Subject oriented data	Integrated data	Time Stamped data	Nonvolatile data	Integrated data
3	_____ refers to the level of details	Subject oriented data	Integrated data	Time Stamped data	Data granularity	Data granularity
4	_____ data comes from the various operational systems of the enterprise.	production	Internal	Archived	External	production
5	_____ data is used to keep the private details of the users in every Organization	production	Internal	Archived	External	Internal
6	_____ data is mostly dependent on the external resources for a high percentage of information.	production	Internal	Archived	External	External
7	_____ function is used to extract information from other data sources	Data Extraction	Data Transformation	Data loading	None	Data Extraction
8	In expert systems the knowledge of domain is represented in terms of _____.	If-then rules	Planners	Samples	Charts	If-then rules
9	_____ is used exclusively for the discovery stage of the KDD process.	OLAP	Cleaning	Enriching	Data mining	Data mining
10	If you know exactly what you are looking for the use _____.	genetic algorithm	SQL	neural network	clustering	SQL
11	Data mart is _____	departmental	corporate	organized	none	departmental
12	Data warehouse is _____	departmental	corporate	organized	none	corporate

13	_____ data comes from various operational system	Internal	production	archived	external	production
14	_____ in a data warehouse is similar to the data dictionary	internal	production	metadata	external	metadata
15	---------- provides support for concurrent data access, data integrity, and access security	Transperancy	Multi user	Client / server Architecture	Flexible reporting	Multi user
16	The _____ algorithm for portioning where each clusters center is represented by the mean value of the objects in the cluster	k-means	k-mediods	star	none	k-means
17	PAM refers to	Partitioning Around Mediods	Partial around mediod	Partitioning Against Mediods	None	Partitioning Around Mediods
18	Expansion of CLARA	Centre LARge Application	Cluster LARge Application	LARge Application	none	Cluster LARge Application
19	_____ algorithm combines sampling text with PAM	Cluster	K-mediods	K-means	CLARANS	CLARANS
20	A classification type where data are grouped but are not predefined is _____	Decision tree	Neural network	Clustering	Tables	Clustering
21	Sample points with values much different from those of the remaining set of data are called _____	Outliers	Point estimation	Correlation	Discarded	Outliers
22	To measure the distance between clusters it is sometimes called a _____	single linkage algorithm	minimal spanning algorithm	nearest-neighbor clustering algorithm	farthest neighbor clustering algorithm	nearest-neighbor clustering algorithm
23	In agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called a	single linkage algorithm	minimal spanning algorithm	neighbor clustering algorithm	neighbor clustering algorithm	minimal spanning algorithm
24	Tree data structure to illustrate hierarchical clustering technique is ____	Histogram	Decision tree	Dendrogram	Classification	Dendrogram
25	BIRCH is Balanced iterative reducing and ___ using hierarchies	Combining	Clustering	Correlative	Conclusive	Clustering

26	Algorithm used in database design as to how to physically place data on disk is ____	Cluster	Bondenergy elgorithm	BOND	Classification	Bond energy algorithm
27	Iterative merging of items into existing clusters that are closed is ___ algorithm	Nearest neighbor	Spanning tree	Decision tree	Neural network	Nearest neighbor
28	Association rules are frequently used by ____	Medical representative	Retail store	Bank	Manufacturer	Retail store
29	The percentage of transactions in which that item occurs	Support	Percentile	Subset	Periodicity	Support
30	___ algorithm is used to generate the candidte itemsets for each pass after the first	Apriori gen	Apriori	Sampling	Partitioning	Apriori gen
31	A quantitative association rule involves categorical and ___ data	Quantitative	Qualitative	Statitical	Associated	Quantitative
32	A correlation rule is defined as a set of itemsets that are ____	Correlated	Regression	neural networks	genetic algorithm	Correlated
33	___ itemsets are said to be downward closed	Large	Small	Medium	Verylarge	Large
34	Where are genetic algorithm applicable?	Realtime application	Biology	Selection	Reproduction	Biology
35	Which of the following is found in genetic algorithm?	Evolution	Crossover	Economics	Representation	Evolution
36	Which new states are generated in genetic algorithm	Composition	Mutation	Crossover	Crossover and mutation	Crossover and mutation
37	Which of the following is pre-requisite when genetic algorithm are applied to solve problems	Encoding of solutions	Well understood search space	Only one optimal solution	Artificial life	Encoding of solutions
38	_____ gives the relationship and patterns between data elements	KDD	data	objects	attributes	KDD
39	_____ is used to describe the whole process of extraction of knowledge from data	data	data mining	KDD	algorithm	data mining
40	KDD has been described as the application of ___ to data mining.	the waterfall model	object-oriented programming	the scientific method	procedural intuition	the scientific method
41	The choice of a data mining tool is made at this step of the KDD process	goal identification	creating a target dataset	data preprocessing	data mining	goal identification
42	Attibutes may be eliminated from the target dataset during this step of the KDD process	creating a target dataset	data preprocessing	data transformation		data transformation

43	This step of the KDD process model deals with noisy data	Creating a target dataset	data preprocessing	data transformation	data mining	data preprocessing
44	KDD stands for _____	Knowledge discovery in database	Knowledge development in data	knowledge data	knowledge extractor	Knowledge discovery in database
45	Creative coding is the heart of _____ process	data	data mining	KDD	algorithm	KDD
46	Which is not present in the six stages of KDD	enrichment	coding	selection	replying	replying
47	At every state of the KDD process the data miner can step back	one phase	two phases	one or more phase	only three	one phase
48	knowledge discovery process consists of _____ stages	six	three	four	five	six
49	Data mining is best described as the process of_____	identifying patterns in data	deducing relationships in data	representing data	simulating trends in data	identifying patterns in data
50	A person trained to interact with a human expert in order to capture their knowledge.	knowledge programmer	knowledge developer	knowledge engineer	knowledge extractor	knowledge engineer

SYLLABUS

UNIT-IV

Data Mining Techniques: Association Rule Mining, classification and regression techniques, clustering

Overview of Data Mining Techniques:

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses. Three steps involved are

- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.

Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and creditrisk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as

preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks

- Back Propagation

Association Rule Mining:

Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.

An association rule has two parts:

- an antecedent (if) and
- a consequent (then).

An antecedent is something that's found in data, and a consequent is an item that is found in combination with the antecedent. Have a look at this rule for instance:

"If a customer buys bread, he's 70% likely of buying milk."

In the above association rule, bread is the antecedent and milk is the consequent.

Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following two parameters, the important relationships are observed:

1. **Support**: Support indicates how frequently the if/then relationship appears in the database.
2. **Confidence**: Confidence tells about the number of times these relationships have been found to be true.

Association rules are if/then statements that help to uncover relationships between unrelated data in a database, relational database or other information repository. Association rules are used to find the relationships between the objects which are frequently used together. **Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalog design, and loss-leader analysis etc.**

$$Rule: \ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

Support(S)

Support(S) of an association rule is defined as the percentage/fraction of records that contain X∪Yto the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1percent of the transaction contain purchasing of this item.

Support (XY) = Support count of (XY) / Total number of transaction in D

Confidence(C)

Confidence(C) of an association rule is defined as the percentage/fraction of the number of transactions that contain X∪Y to the total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X⇒Y is 80%, it means that 80% of the transactions that contain X also contain Y together.

Confidence (X|Y) = Support (XY) / Support (X)

Association Rules Goals

•Find all sets of items (item-sets) that have support (number of transactions) greater than the minimum support (large item-sets).

•Use the large item-sets to generate the desired rules that have confidence greater than the minimum confidence.

For example, if the customer buys bread then he may also buy butter. If the customer buys laptop then he may also buy memory card.

There are two basic criteria that association rules uses, support and confidence. It identifies the relationships and rules generated by analyzing data for frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user -specified minimum confidence at the same time.

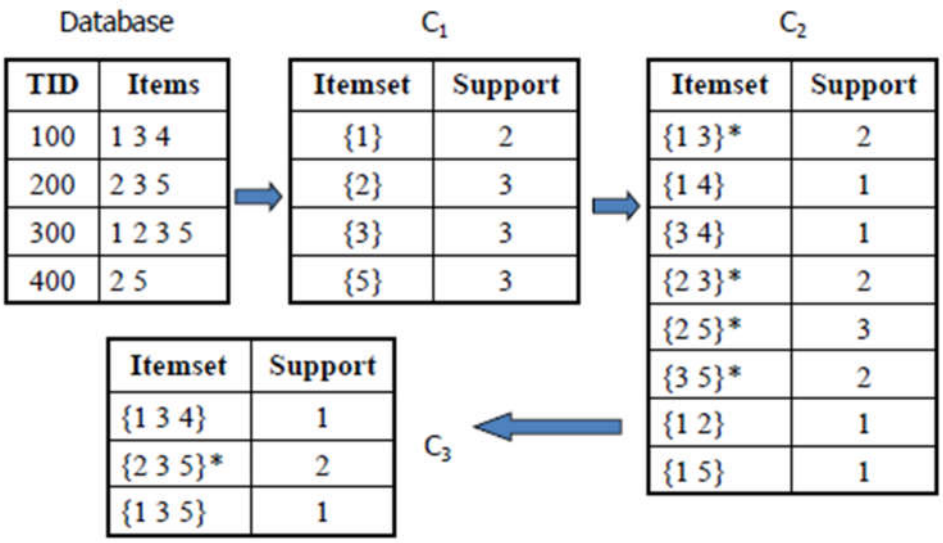
AIS ALGORITHM

1. Candidate itemsets are generated and counted on-the-fly as the database is scanned.
2. For each transaction, it is determined which of the large itemsets of the previous pass are contained in this transaction.
3. New candidate itemsets are generated by extending these large itemsets with other items in this transaction.

The drawback of this algorithm

1.If it has too many candidate itemsets that finally turned out to be small are generated.
2.It requires more space and it wastes much effort.
3.As well as this algorithm requires too many passes over the whole database.

Example of AIS algorithm



TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

SETM ALGORITHM

1. Candidate itemsets are generated on-the-fly as the database is scanned, but counted at the end of the pass.
2. New candidate itemsets are generated the same way as in AIS algorithm, but the TID of the generating transaction is saved with the candidate itemset in a sequential structure.
3. At the end of the pass, the support count of candidate itemsets is determined by aggregating this sequential structure.

Example of SETM algorithm

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	TID
{1 3}	100
{1 4}	100
{3 4}	100
{2 3}	200
{2 5}	200
{3 5}	200
{1 2}	300
{1 3}	300
{1 5}	300
{2 3}	300
{2 5}	300
{3 5}	300
{2 5}	400

C_3

Itemset	TID
{1 3 4}	100
{2 3 5}	200
{1 3 5}	300
{2 3 5}	300

The SETM algorithm has the same disadvantage of the AIS algorithm. Another disadvantage is that for each candidate itemset, there are as many entries as its support value.

APRIORI ALGORITHM

Apriori is a classic algorithm for mining frequent items for boolean Association rule.It uses a bottom-up approach, designed for finding Association rules in a database that contains transactions.

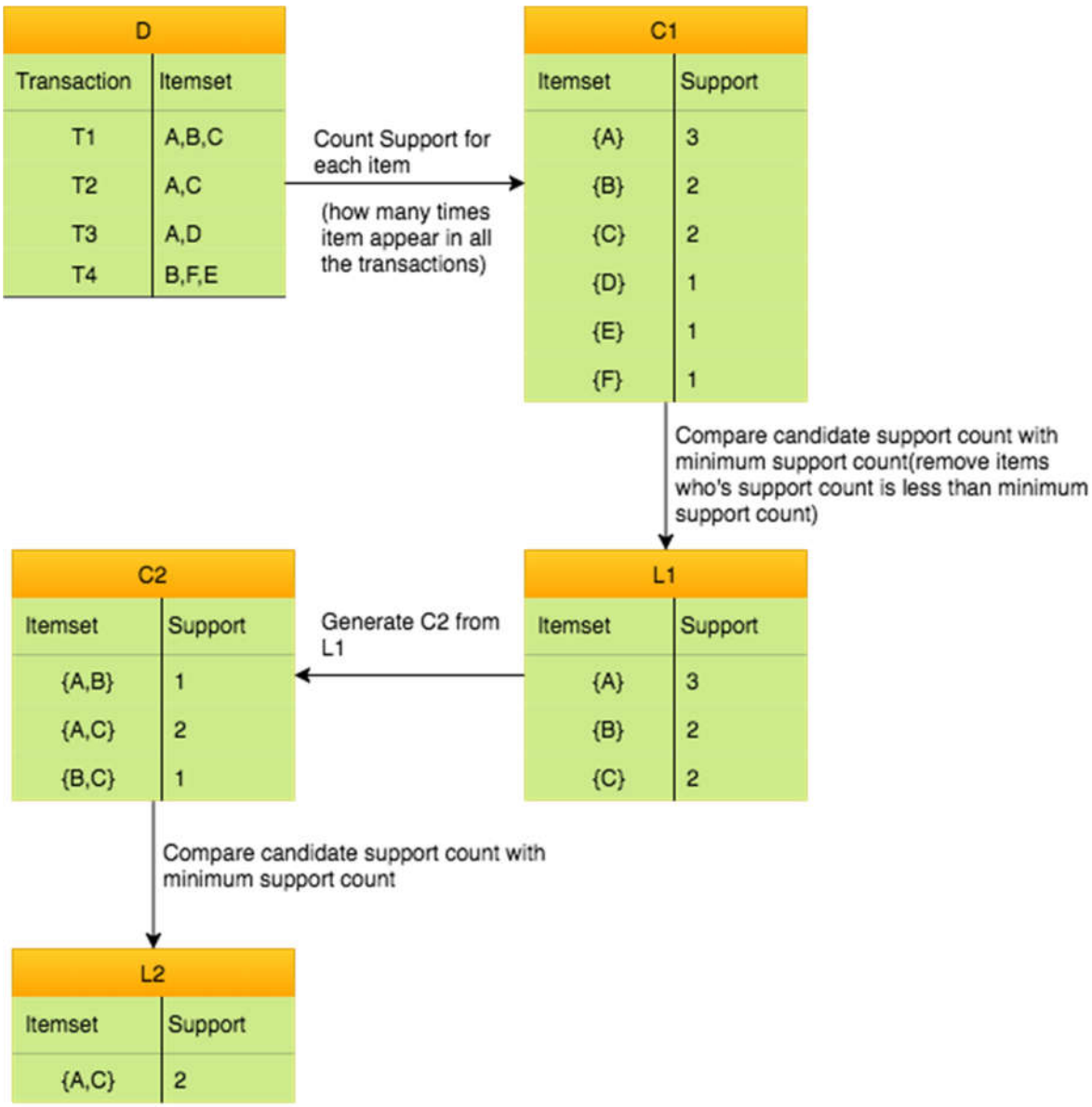
Following is the procedure for Apriori algorithm:

CIk : Candidate itemset having size k
FIk : Frequent itemset having size k
FI1 = {frequent items};
For (k=1; FIk != null; k++) do begin
CIk+1 = candidates generated from FIk;
For each transaction t in database D do
Increment the count value of all candidates in CIk+1 that are contained in t
FIk+1 = candidates in CIk+1 with min_support
End Return FIk;

Apriori algorithm example

Consider the following database(D) with 4 transactions (T1,T2,T3 and T4).

Let minimum support = 2%
Let minimum confidence = 50%



Now generate Association rule:-

Association Rule	Support	Confidence	Confidence %
A ⟶ C	2	2/3 = 0.66	66%
C ⟶ A	2	2/2 = 1	100%

To Find Confidence:-

$$\text{Confidence (A -> C)} \longrightarrow \frac{\text{Number of transaction contain both A and C (in table D)}}{\text{Number of transaction contain Only A (in table D)}} = \frac{2}{3} = 0.66$$

To convert confidence into confidence%, multiply confidence with 100

Example

confidence%(A -> C) = 0.66 x 100 = 66%

Now compare confidence% of association rule with a minimum confidence threshold (50%). Since confidence% of both Association rule is greater than minimum confidence threshold, so both Association rules are final rules.

Advantage of the Apriori algorithm

- Any subset of a frequent itemset is also a frequent itemset.
- Reduce the number of candidates being considered by only exploring the itemsets whose support count is greater than the minimum support count.
- All infrequent itemsets can be pruned if it has an infrequent subset.

Drawbacks of the Apriori algorithm

- First is the complex candidate generation process which uses most of the time, space and memory.
- It requires multiple scans of the database.

APRIORITID ALGORITHM

1. The database is not used at all for counting the support of candidate itemsets after the first pass.
2. The candidate itemsets are generated the same way as in Apriori algorithm.
3. Another set C' is generated of which each member has the TID of each transaction and the large itemsets present in this transaction. This set is used to count the support of each candidate itemset.

The advantage of this algorithm is that, in the later passes the performance of Aprioritid is better than Apriori.

Example of Aprioritid algorithm

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 2}	1
{1 3}*	2
{1 5}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2

C_3

Itemset	Support
{2 3 5}*	2

C'_2

100	{1 3}
200	{2 3}, {2 5}, {3 5}
300	{1 2}, {1 3}, {1 5}, {2 3}, {2 5}, {3 5}
400	{2 5}

C'_3

200	{2 3 5}
300	{2 3 5}

APRIORIHYBRID ALGORITHM

As Apriori does better than Aprioritid in the earlier passes and Aprioritid does better than Apriori in the later passes. A new algorithm is designed that is Apriori hybrid which uses features of both the above algorithms. It uses Apriori algorithm in earlier passes and Apriori tid algorithm in later passes.

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 itemset in a transaction database without candidate generation.

FP growth represents frequent items in frequent pattern trees or FP-tree.

Advantages of FP growth algorithm:-

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

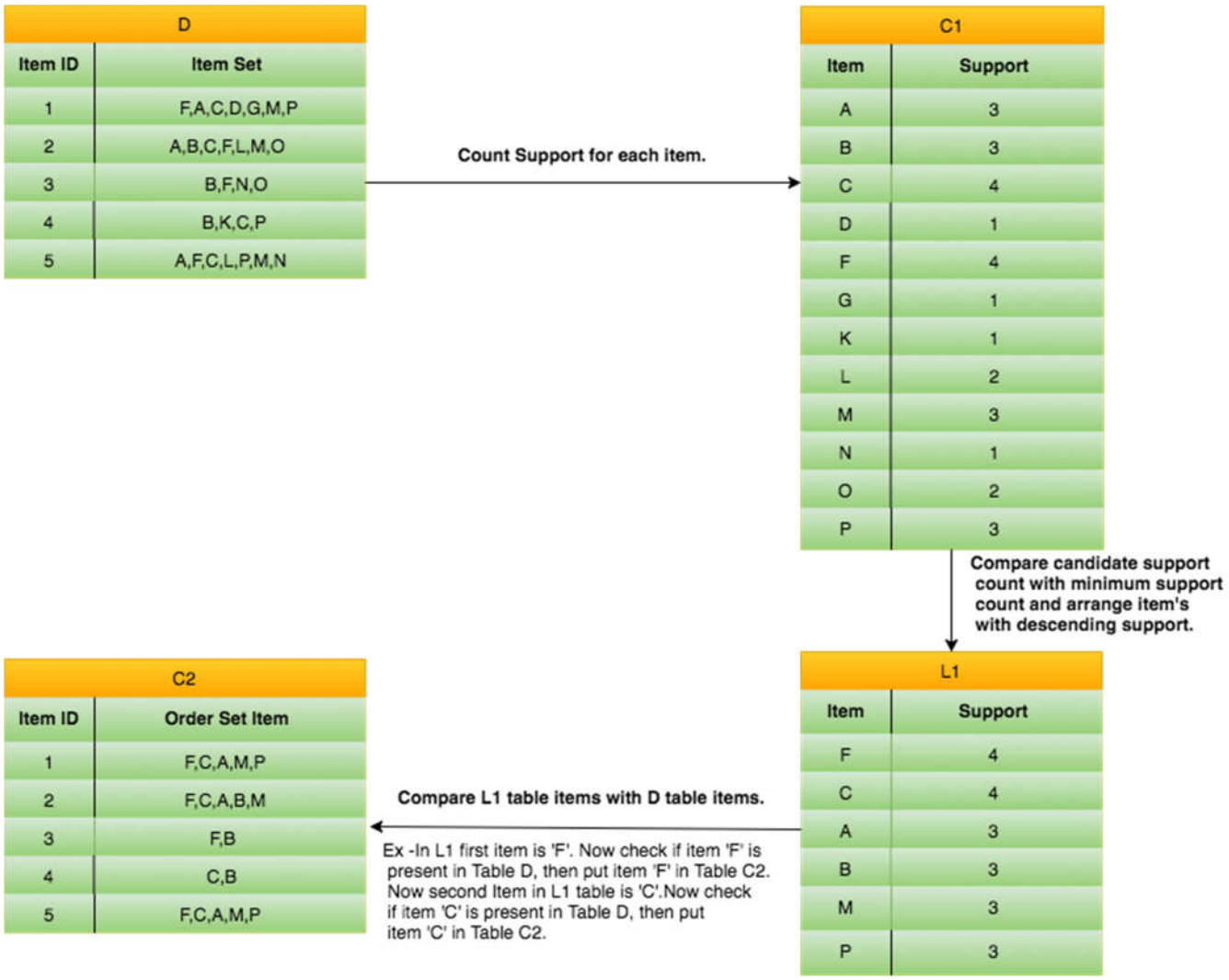
Disadvantages of FP growth algorithm:-

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

Fp growth algorithm example

Consider the following database(D)

Let 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,P
5
A,F,C,L,P,M,N
Count Support for each item.
C1
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.
L1
Item
Support
F
4
C
4
A
3
B
3
M
3
P
3
Compare L1 table items with D table items.
Ex -In L1 first item is 'F'. Now check if item 'F' is present in Table D, then put item 'F' in Table C2. Now second Item in L1 table is 'C'.Now check if item 'C' is present in Table D, then put item 'C' in Table C2.
C2
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

Difference between Fp growth and Apriori Algorithm:

FP growth algorithm and Apriori algorithm they both are used for mining frequent items for boolean Association rule.

FP growth algorithm	Apriori algorithm
FP growth algorithm is faster than Apriori algorithm.	It is slower than FP growth algorithm.
FP growth algorithm is an array based algorithm.	Apriori algorithm is a tree-based algorithm.
FP growth algorithm required only two database scan.	It requires multiple database scan to generate a candidate set.
It uses depth-first search	It uses breadth-first search.

The main applications of association rule mining:

- Basket data analysis - is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.
- Cross marketing - is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.
- Catalog design - the selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

Classification Techniques:

Classification: It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

Example: Before starting any Project, we need to check it's feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it. It is a two step process such as :

1. **Learning Step (Training Phase)**: Construction of Classification Model
   Different Algorithms are used to build a classifier by making the model learn using the training set available. Model has to be trained for prediction of accurate results.
2. **Classification Step**: Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

Training and Testing:

Suppose there is a person who is sitting under a fan and the fan starts falling on him, he should get aside in order not to get hurt. So, this is his training part to move away. While testing if the person sees any heavy object coming towards him or falling on him and moves aside then system is tested positively and if the person do not moves aside then the system is negatively tested.
Same is the case with the data; it should be trained in order to get the accurate and best results.

There are certain data types associated with data mining that actually tells us the format of the file (whether it is in text format or in numerical format).
Attributes – Represents different features of an object.

Different types of attributes are:

1. **Binary**: Possesses only two values i.e. True or False
   Example: Suppose there is a survey of evaluating some product. We need to check whether it's useful or not. So, the Customer have to answer it in Yes or No.
   Product usefulness: Yes / No
- **Symmetric**: Both values are equally important in all aspects
- **Asymmetric**: When both the values may not be important.

2. **Nominal**: When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.
   Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.
   Different Colors: Red, Green, Black, Yellow

- **Ordinal**: Values that must have some meaningful order.
  Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D
  Grades: A, B, C, D
- **Continuous**: May have infinite number of values, it is in float type
  Example: Measuring weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53
  Weight: 50, 51, 52, 53

- **Discrete**: Finite number of values.
  Example: Marks of a Student in few subjects: 65, 70, 75, 80, 90
  Marks: 65, 70, 75, 80, 90

Syntax:

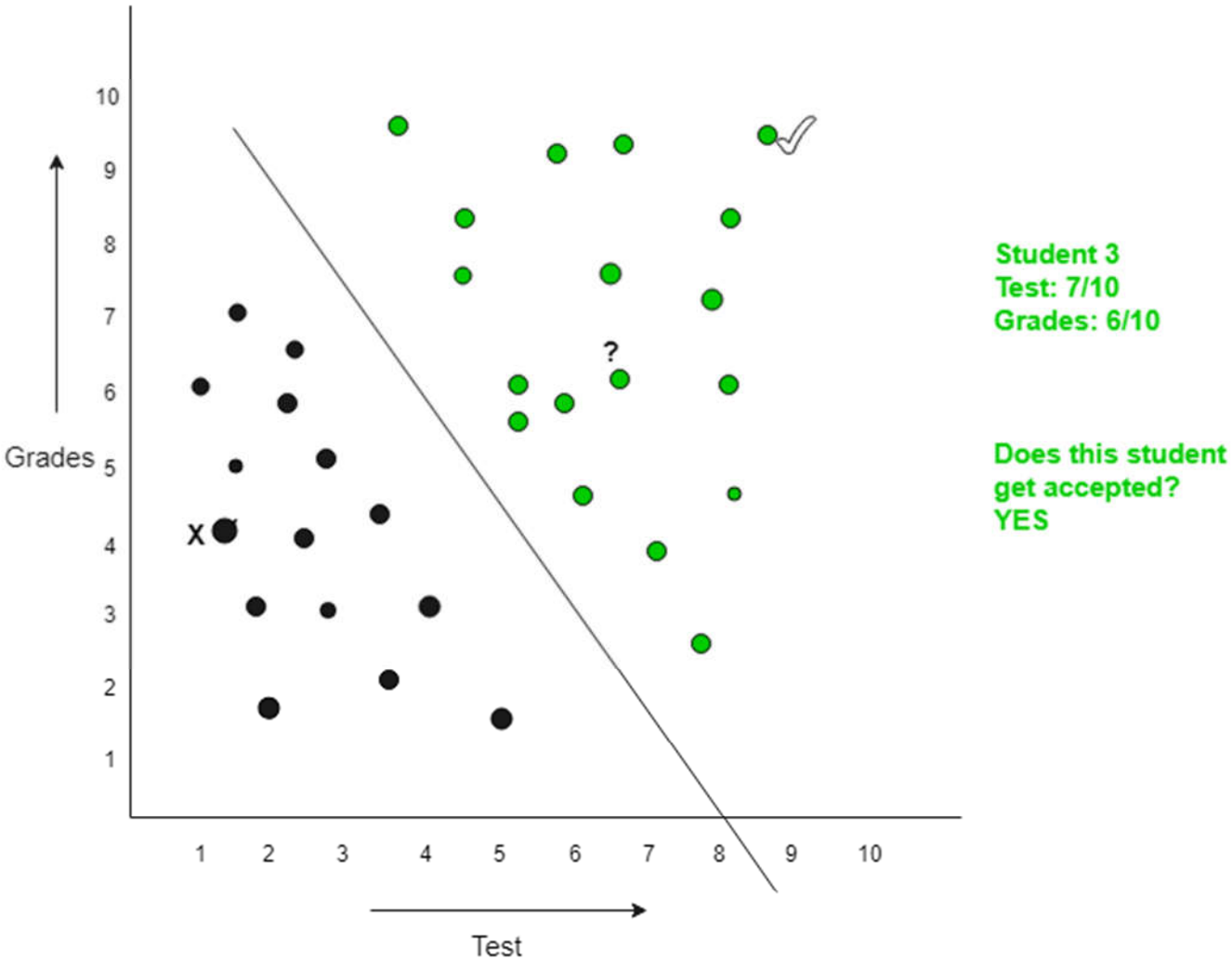
- Mathematical Notation: Classification is based on building a function taking input feature vector "X" and predicting its outcome "Y" (Qualitative response taking values in set C)
- Here Classifier (or model) is used which is a Supervised function, can be designed manually based on expert's knowledge. It has been constructed to predict class labels (Example: Label – "Yes" or "No" for the approval of some event).

Classifiers can be categorized on two major types:

1. **Discriminative**: It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on quality of data rather than on distributions.
   Example: Logistic Regression
   Acceptance of a student at a University (Test and Grades need to be considered)

Suppose there are few students and the Result of them are as follows:

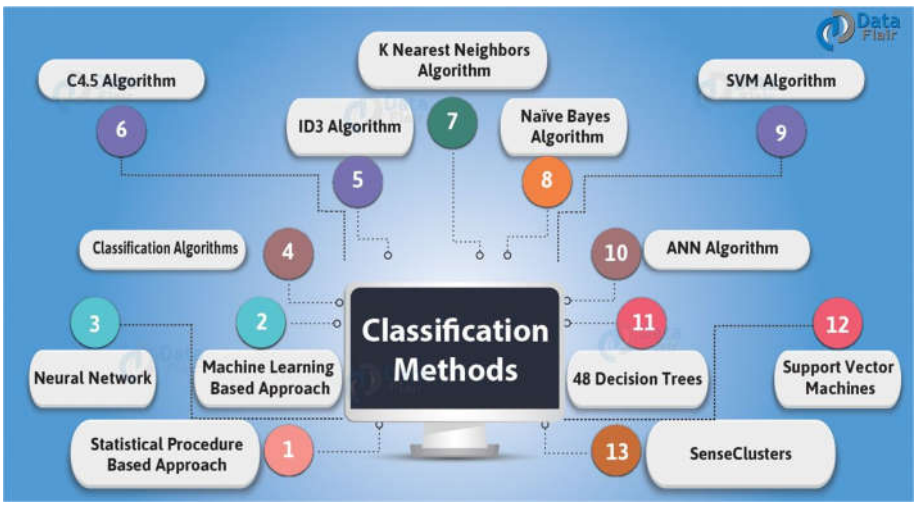
Student 1 : Test Score: 9/10, Grades: 8/10 Result: Accepted
Student 2 : Test Score: 3/10, Grades: 4/10, Result: Rejected
Student 3 : Test Score: 7/10, Grades: 6/10, Result: to be tested



2. **Generative**: It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.

   Example: Naive Bayes Classifier
   Detecting Spam emails by looking at the previous data. Suppose 100 emails and that too divided in 1:4 i.e. Class A: 25%(Spam emails) and Class B: 75%(Non-Spam emails). Now if a user wants to check that if any email contains the word cheap, then that may be termed as Spam.
   It seems to be that in Class A(i.e. in 25% of data), 20 out of 25 emails are spam and rest not.
   And in Class B(i.e. in 75% of data), 70 out of 75 emails are not spam and rest are spam.
   So, if the email contains the word cheap, what is the probability of it being spam ?? (= 80%)

Classification Algorithms:



Real Life Examples:

- **Market Basket Analysis:**
  It is a modeling technique that has been associated with frequent transactions of buying some combination of items.
  Example: Amazon and many other Retailers use this technique. While viewing some product, certain suggestions for the commodities are shown that some people have bought in the past.
- **Weather Forecasting:**
  Changing Patterns in weather conditions needs to be observed based on the parameters such as temperature, humidity, wind direction. This keen observation also requires use of previous records in order to predict it accurately.

Advantages:

- Mining Based Methods are cost effective and efficient
- Helps in identifying criminal suspects
- Helps in predicting risk of diseases
- Helps Banks and Financial Institutions to identify defaulters so that they may approve Cards, Loan, etc.

Disadvantages:

Privacy: When the data is there are chances that a company may give some information of their customers to other vendors or use this information for their profit.
Accuracy Problem: Selection of Accurate model must be there in order to get best accuracy and result.

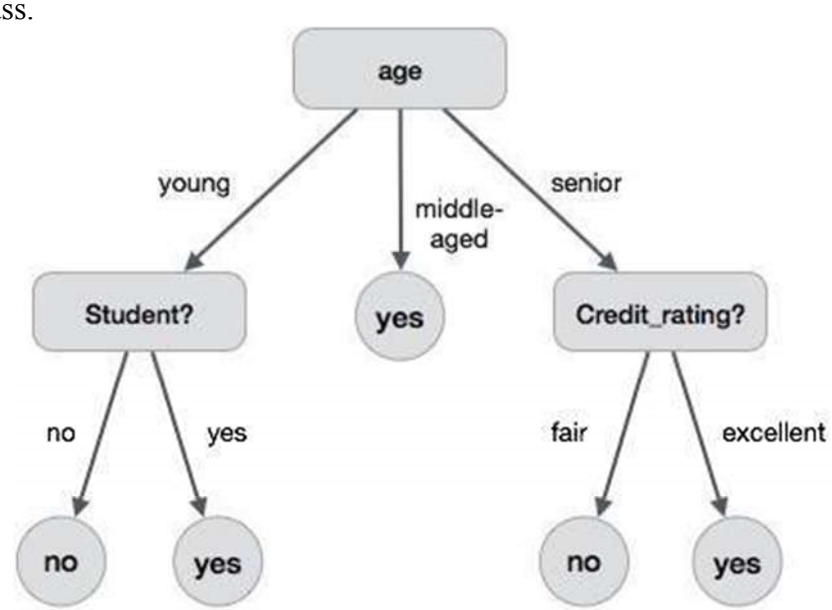
APPLICATIONS:

- Marketing and Retailing
- Manufacturing
- Telecommunication Industry
- Intrusion Detection
- Education System
- Fraud Detection

Classification: Decision Tree

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and 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 attribute. Each leaf node represents a class.



The 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.

Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches

There are two approaches to prune a tree −

- **Pre-pruning** − The tree is pruned by halting its construction early.
- **Post-pruning** - This approach removes a sub-tree from a fully grown tree.

Cost Complexity

The cost complexity is measured by the following two parameters −

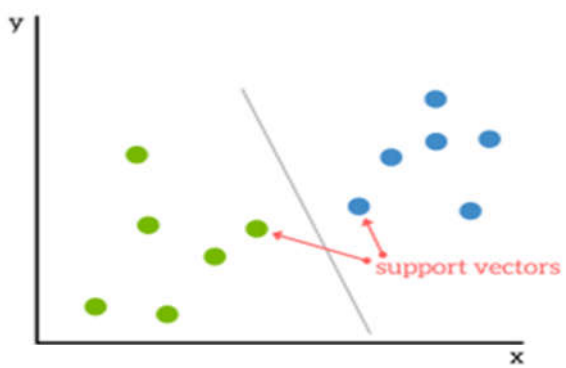
- Number of leaves in the tree, and
- Error rate of the tree.

Support Vector Machine - Classification (SVM)

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems and as such, this is what we will focus on in this post.

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the image below.



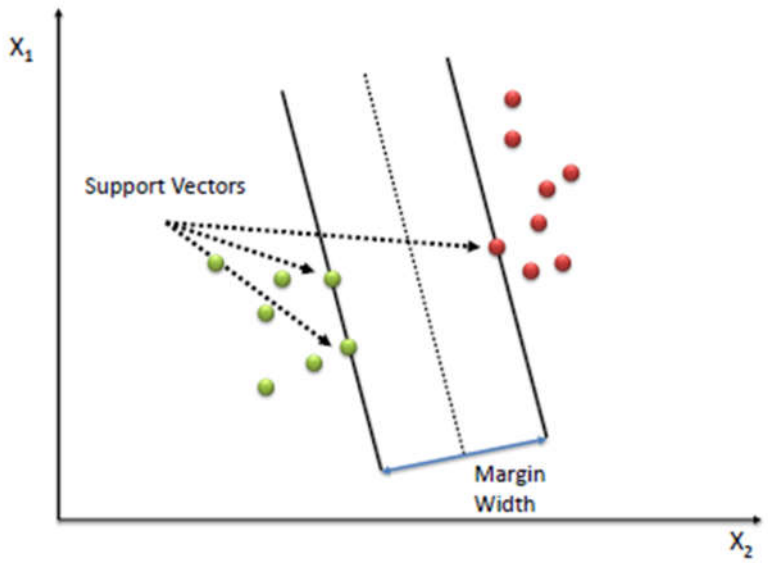
Support Vectors

Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set.

What is a hyperplane?

As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.
So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.



Algorithm

1. Define an optimal hyper plane: maximize margin
2. Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications.

Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.

Clustering:

- Generally, a group of abstract objects into classes of similar objects is made.
- We treat a cluster of data objects as one group.
- While doing cluster analysis, we first partition the set of data into groups. That based on data similarity and then assigns the labels to the groups.
- The main advantage of over-classification is that it is adaptable to changes. And helps single out useful features that distinguish different groups.

Applications of Data Mining Cluster Analysis:

- Data clustering analysis is used in many applications. Such as market research, pattern recognition, data analysis, and image processing.

- Data Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies. Categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering in Data Mining helps in identification of areas. That is of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city. That is according to house type, value, and geographic location.
- Clustering in Data Mining also helps in classifying documents on the web for information discovery
- Also, we use Data clustering in outlier detection applications. Such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool. That is to gain insight into the distribution of data. Also, need to observe characteristics of each cluster.

Requirements of Clustering in Data Mining:

The following points state us the requirement of clustering in Data Mining:

a. Scalability

We need highly scalable clustering algorithms to deal with large databases.

b. Ability to deal with different kinds of attributes

Algorithms should be capable to be applied to any kind of data. Such as interval-based data, categorical, and binary data.

c. Discovery of clusters with attribute shape

The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded by only distance measures. That tends to find a spherical cluster of small sizes.

d. High dimensionality

The clustering algorithm should not only be able to handle low-dimensional data. Although, need to handle the high dimensional space.

e. Ability to deal with noisy data

Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

f. Interpretability

The clustering results should be interpretable, comprehensible, and usable.

Data Mining Clustering Methods:

Data Mining Clustering Methods are classified into the following categories –

a. Partitioning Clustering Method:

Suppose we are given a database of 'n' objects. And the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups. That must need to satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember −

- If we have a given number of partitions (say k). Then the partitioning method will create an initial partitioning.
- Further, it uses the iterative relocation technique. That is to improve the partitioning by moving objects from one group to other.

b. Hierarchical Clustering Methods:

The hierarchical method creates a hierarchical decomposition of the given set of data objects. We can classify methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

i. Agglomerative Approach:

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.

ii. Divisive Approach:

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. Then, in the continuous iteration, a cluster is split up into smaller clusters. Also, it is down until each object in one cluster or the termination condition holds. Hence, this method is rigid, i.e., once a merging or splitting is done, it can never be undone.

c. Density-Based Clustering Method:

This Data Mining Clustering method is based on the notion of density. The idea is to continue growing the given cluster. That is exceeding as long as the density in the neighbourhood threshold.
For each data point within a given cluster, the radius of a given cluster has to contain at least number of points.

d. Grid-Based Clustering Method:

In this, the objects together form a grid. The object space is quantized into a finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is a fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

e. Model-Based Clustering Methods:

In this Data Mining Clustering method, a model is hypothesized for each cluster to find the best fit of data for a given model. Also, this method locates the clusters by clustering the density function. Thus, it reflects the spatial distribution of the data points.
This method also provides a way to determine the number of clusters. That was based on standard statistics, taking outlier or noise into account. It, therefore, yields robust clustering methods.

f. Constraint-Based Clustering Method:

The clustering is performed by the incorporation of a user or application-oriented constraints. A constraint refers to the user expectation. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application need.

POSSIBLE QUESTIONS

2 MARK

1. Define Association rule with example.

2. Construct a decision tree with root node *Type* from the data in the table below. The first row contains attribute names. Each row after the first represents the values for one data instance. The output attribute is *Class*.

Scale	Type	Shade	Texture	Class
One	One	Light	Thin	A
Two	One	Light	Thin	A
Two	Two	Light	Thin	B
Two	Two	Dark	Thin	B

Two	One	Dark	Thin	C
One	One	Dark	Thin	C
One	Two	Light	Thin	C

3. What are the Types of classification models?

4. Name the Types of clustering methods.

5. What are the Types of regression methods?

6. Mention the Types of association rule.

7. What is Association rule?

8. Define neural networks.

9. define the Association Rule Goals.

10. What are the advantages of clustering?

8 MARKS

1. Elaborate Association rule with Apriori Algorithm.

2. Explain Clustering Technique.

3. Write about Classification Technique with example.

4. Write a brief note on Regression with example.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore-641021

Department of Computer Science

III BSc(CS) (BATCH 2017-2020)

Data Mining (17CSU602A)

PART-A OBJECTIVE TYPE/ MULTIPLE CHOICE QUESTIONS

ONLINE EXAMINATIONS **ONE MARK QUESTIONS**

UNIT-4

S.No	Question	option1	option2	option3	option4	Answer
1	_____ is the percentage of retrieved documents that re in fact relevant to the query	Precision	Recall	Retrieval	information	Precision
2	A _____ database stores and manages a large collection of multimedia data	spatial	temporal	web	multimedia	multimedia
3	The object oriented data model inherits the essential concepts of _____ database	Multimedia	Object oriented	Spatial	None	Object oriented
7	_____ takes all the data at once and tries to create hyposthesis based on the data	noise	data minig	supervised algorithm	batch learning algorithm	batch learning algorithm
8	_____ is the random disturbance of a transmitted signal in the context of KDD	noise	data minig	supervised algorithm	batch learning algorithm	noise
9	_____ is used at thebegining of datamining process to get a good quality of dataset	gentic	data mining	visualization	none	visualization
10	Important elements in the cleaning operation is the _____	k-nearest	inconsistency	deduplicaiton of records	none	deduplicaiton of records
11	Including records that are closed to each other or living each other's neighbour hood is called as _____	genetic	neural networks	k-nearest neighbour	visualizaiton	k-nearest neighbour
12	_____ was modeled based on human brain	genetic	neural networks	k-nearest neighbour	visualizaiton	neural networks
14	_____ refers to elements of the message that can be derived from other part of the other message	machine language	neural	k-nearest	redundancy	redundancy
15	_____ program must be unable to read it by the users	nerual networks	genetic algorihtm	machine language	none	machine language
16	_____ is not realistic for most learning situations	data mining	neural networks	background knowledge	none	background knowledge
17	_____ are always defined on binary attributes	association rules	genetic algorithm	neural networks	KDD	association rules

19	_____ is a interaction between technology and nature	genetic algorithm	neural networks	k-nearest	none	genetic algorithm
20	_____ contains the visual map of data sets	kohonen self organizing map	neural networks	k-nearest	none	kohonen self organizing map
21	A _____ network not only has input and outputnodes but also hidden nodes	gentic algorithm	neural network	back propogation	k-nearest	back propogation
22	In _____ information can be easily retrived from the database using query tools	Mulit dimenisonal knowledge	shallow knowledge	hidden knowledge	deep knowledge	shallow knowledge
23	In _____ information that can be analyzed using online analytical processing tools	Mulit dimenisonal knowledge	shallow knowledge	hidden knowledge	deep knowledge	Mulit dimenisonal knowledge
24	In _____ data can be found relatively easily by using pattern recognition or machine learning algorithm	Mulit dimenisonal knowledge	shallow knowledge	hidden knowledge	deep knowledge	hidden knowledge
25	In _____ information that is stored in the database but can only be located if we have a clue that tells us	Mulit dimenisonal knowledge	shallow knowledge	hidden knowledge	deep knowledge	deep knowledge
26	A _____ consists of a simple three-layered network	responders	photo receptors	perceptrons	none	none
27	In perceptron input unit is called	responders	photo receptors	perceptrons	none	photo receptors
28	In perceptron intermediate units called	responders	photo receptors	perceptrons	associators	associators
29	In perceptron network output unit is called	Responders	photo receptors	perceptrons	none	Responders
30	The _____ is very useful in a data mining context	responders	photo receptors	perceptrons	space-metaphor	space-metaphor
31	In which process pollution is detected	data cleaning	data selection	data enrichment	data coding	data cleaning
32	In which process data are collected from the operation database	data cleaning	data selection	data enrichment	data coding	data selection
33	In which process additional information are included in the existing database	data cleaning	data selection	data enrichment	data coding	data enrichment

34	_____ is a creative process	data cleaning	data selection	data enrichment	data coding	data coding
35	Learning tasks can be divided into _____ areas	None	two	three	four	three
36	A _____ contains historic data and is subject oriented and static	cleaning	data warehousing	database	operational data	data warehousing
37	_____ algorithm comes under classification tasks	neural networks	association rules	inductive logic programming	genetic algorithm	neural networks
38	_____ algorithm comes under problem solving tasks	neural networks	association rules	inductive logic programming	genetic algorithm	genetic algorithm
39	_____ algorithm come under knowledge engineering tasks	neural networks	association rules	inductive logic programming	genetic algorithm	inductive logic programming
40	The visual map of given sets and kohenen sel organizing map is a collection of _____	neurons and units	data base	KDD	none	neurons and units
41	_____ tool is not a single technique it holds variety of different techniques	neurons and units	data base	Data Mining	none	Data Mining
42	A very important element in a cleanning operation is that	de duplicaiton of data	replicaiton	inconsistency	none	de duplicaiton of data
43	_____ starts with number of observations	analysis	observation	genetic	none	observation
44	_____ try to find the patterns in the observations	analysis	observation	genetic	none	analysis
45	The theory which give new phenomena that can be verified by new observations is known as	observations	analysis	theory	predection	predection
46	_____ is a form of adaptation	learning	discovery	finding	none	learning
47	Geological sample could be represented in terms of _____ rules	do while	for	if-then	none	if-then
48	The _____ revolution gives the individual knowledge worker access to central information systems	client/server	machine learning	SQL	none	client/server
49	Data can be classified into _____ categories	2	3	4	5	2

50	____ which is the basis for online analytical processing tools, prepared periodically but is directly based on detailed	reference and transaction	derived data	denormalised data	data warehouse	denormalised data
51	A repository of subjectively selected and adapted operational data	Reference and transaction	Derived	Meta	Denormalized	Reference and transaction
52	Data that once put in datawarehouse cannot be modified is ____ data	Reference	Denormalized	Transaction	Derived	Denormalized
53	Flow of data to various departments as departmental components is called ____	Processing	Procedure	Publicity	Production	Processing
54	Aggregation of data at different levels of hierarchies in a given dimension	Powerplay	Powercubes	Powerstake	Pointercube	Powerplay
55	Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that	Y is true when X is known to be true	X is true when Y is known to be true	Y is false when X is known to be false	X is false when Y is known to be false	Y is true when X is known to be true
56	Associatio rule support is defined as _____	the percentage of instances that contain the antecendent conditional items listed in the association rule	the percentage of instances that contain the consequent conditions listed in the association rule	the percentage of instances that contain all items listed in the association rule	the percentage of instances in the database that contain at least one of the antecendent conditional items listed in the association rule	the percentage of instances that contain all items listed in the association rule
57	This approach is best when we are interested in finding all possible interactions among a set of attributes.	decision tree	association rules	K-Means algorithm	genetic learning	association rules
58	An evolutionary approach to data mining.	backpropagation learning	genetic learning	decision tree learning	linear regression	genetic learning
59	The computational complexity as well as the explanation offered by a genetic algorithm is largely determined by the	fitness function	techniques used for crossover and mutation	training data	population of elements	fitness function
60	This technique uses mean and standard deviation scores to transform real-valued attributes.	decimal scaling	min-max normalization	z-score normalization	logarithmic normalization	z-score normalization
61	With Bayes classifier, missing data items are	treated as equal compares.	treated as unequal compares.	replaced with a default value.	ignored.	ignored.

62	This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.	agglomerative clustering	expectation maximization	conceptual clustering	K-Means clustering	conceptual clustering
63	This clustering algorithm initially assumes that each data instance represents a single cluster.	agglomerative clustering	conceptual clustering	K-Means clustering	expectation maximization	agglomerative clustering

SYLLABUS

UNIT-V

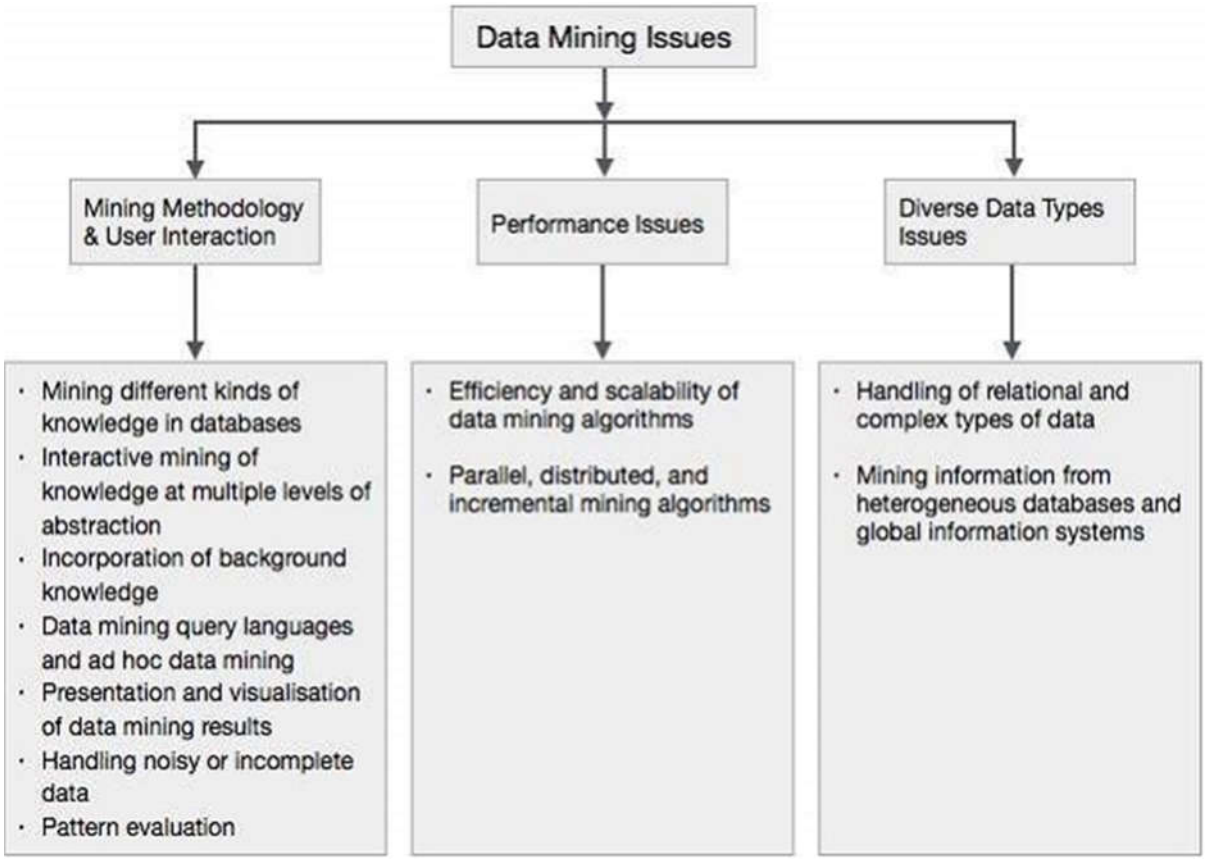
Scalability and data management issues in data mining algorithms, measures of interestingness.

Data Mining - Issues

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − to guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − the patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions

are merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − the database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

Common problems in Data Mining:

1. Poor data quality such as noisy data, dirty data, missing values, inexact or incorrect values, inadequate data size and poor representation in data sampling.
2. Integrating conflicting or redundant data from different sources and forms: multimedia files (audio, video and images), geo data, text, social, numeric, etc…
3. Proliferation of security and privacy concerns by individuals, organizations and governments.
4. Unavailability of data or difficult access to data.
5. Efficiency and scalability of data mining algorithms to effectively extract the information from huge amount of data in databases.
6. Dealing with huge datasets that require distributed approaches.
7. Dealing with non-static, unbalanced and cost-sensitive data.
8. Mining information from heterogeneous databases and global information systems.
9. Constant updating of models to handle data velocity or new incoming data.
10. High cost of buying and maintaining powerful software's, servers and storage hardware's that handle large amounts of data.
11. Processing of large, complex and unstructured data into a structured format.
12. Sheer quantity of output from many data mining methods.

Interestingness Measures for Data Mining:

Interestingness measures play an important role in data mining, regardless of the kind of patterns being mined. These measures are intended for selecting and ranking patterns according to their potential interest to the user. Good measures also allow the time and space costs of the mining process to be reduced.

Data mining can be regarded as an algorithmic process that takes data as input and yields patterns such as classification rules, association rules, or summaries as output. An association rule is an implication of the form $X \rightarrow Y$, where X and Y are nonintersecting sets of items.

For example, {milk, eggs} → {bread} is an association rule that says that when milk and eggs are purchased, bread is likely to be purchased as well. A classification rule is an implication of the form X_1 op x_1, X_2 op x_2, ... , X_n op $x_n \rightarrow Y = y$, where X_i is a conditional attribute, x_i is a value that belongs to the domain of X_i, Y is the class attribute, y is a class value, and op is a relational operator such as = or >.

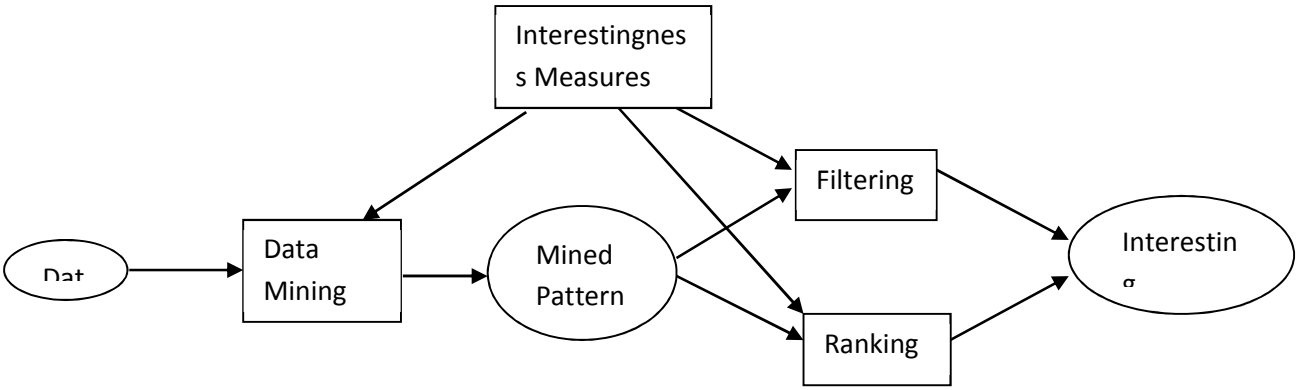


Fig: Roles of interestingness measures in the data mining process

For example, Job = Yes, AnnualIncome > 50,000 → Credit = Good, is a classification rule which says that a client who has a job and an annual income of more than $50,000 is classified as having good credit. A summary is a set of attribute-value pairs and aggregated counts, where the values may be given at a higher level of generality than the values the students majoring in computer science in terms of two attributes: nationality and program. In this case, the value of "Foreign" for the nationality attribute is given at a higher level of generality in the summary than in the input data, which gives individual nationalities.

Summary of Students Majoring in Computer Science

Program	Nationality	# of Students	Uniform Distribution	Expected
Graduate	Canadian	15	75	20
Graduate	Foreign	25	75	30
Undergraduate	Canadian	200	75	180
Undergraduate	Foreign	60	75	70

Measuring the interestingness of discovered patterns is an active and important area of data mining research. Although much work has been conducted in this area, so far there is no widespread agreement on a formal definition of interestingness in this context. Based on the diversity of definitions presented to-date, interestingness is perhaps best treated as a broad concept that emphasizes conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, and actionability. These nine specific criteria are used to determine whether or not a pattern is interesting. They are described as follows.

Conciseness:

A pattern is concise if it contains relatively few attribute-value pairs, while a set of patterns is concise if it contains relatively few patterns. A concise pattern or set of patterns is relatively easy to understand and remember and thus is added more easily to the user's knowledge (set of beliefs).
Generality/Coverage:

A pattern is general if it covers a relatively large subset of a dataset. Generality (or coverage) measures the comprehensiveness of a pattern, that is, the fraction of all records in the dataset that matches the pattern. If a pattern characterizes more information in the dataset, it tends to be more interesting frequent itemsets are the most studied general patterns in the data mining literature. An itemset is a set of items, such as some items from a grocery basket. An itemset is frequent if its support, the fraction of records in the dataset containing the itemset, is above a given threshold. The best known algorithm for finding frequent itemsets is the Apriori algorithm. Some generality measures can form the bases for pruning strategies; for example, the support measure is used in the Apriori algorithm as the basis for pruning itemsets. For classification rules, gave an empirical evaluation showing how generality affects classification results. Generality frequently coincides with conciseness because concise patterns tend to have greater coverage.

Reliability:

A pattern is reliable if the relationship described by the pattern occurs in a high percentage of applicable cases. For example, a classification rule is reliable if its predictions are highly accurate, and an association rule is reliable if it has high confidence. Many measures from probability, statistics, and information retrieval have been proposed to measure the reliability of association rules.

Peculiarity:

A pattern is peculiar if it is far away from other discovered patterns according to some distance measure. Peculiar patterns are generated from peculiar data (or outliers), which are relatively few in number and significantly different from the rest of the data. Peculiar patterns may be unknown to the user, hence interesting.

Diversity:

A pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set differ significantly from each other. Diversity is a common factor for measuring the interestingness of summaries. According to a simple point of view, a summary can be considered diverse if its probability distribution is far from the uniform distribution. A diverse summary may be interesting because in the absence of any relevant knowledge, a user commonly assumes that the uniform distribution will hold in a summary. According to this reasoning, the more diverse the summary is, the more interesting it is. We are unaware of any existing research on using diversity to measure the interestingness of classification or association rules.

Novelty:

A pattern is novel to a person if he or she did not know it before and is not able to infer it from other known patterns. No known data mining system represents everything that a user knows, and thus, novelty cannot be measured explicitly with reference to the user's knowledge. Similarly, no known data mining system represents what the user does not know, and therefore, novelty cannot be measured explicitly with reference to the user's ignorance. Instead, novelty is detected by having the user either explicitly identifies a pattern as novel or notice that a pattern cannot be deduced from and does not contradict previously discovered patterns. In the latter case, the discovered patterns are being used as an approximation to the user's knowledge.

Surprisingness:

A pattern is surprising (or unexpected) if it contradicts a person's existing knowledge or expectations. A pattern that is an exception to a more general pattern which has already been discovered can also be considered surprising. Surprising patterns are interesting because they identify failings in previous knowledge and may suggest an aspect of the data that needs further study. The difference between surprisingness and novelty is that a novel pattern is new and not contradicted by any pattern already known to the user, while a surprising pattern contradicts the user's previous knowledge or expectations.

Utility:

A pattern is of utility if its use by a person contributes to reaching a goal. Different people may have divergent goals concerning the knowledge that can be extracted from a dataset. For example, one person may be interested in finding all sales with high profit in a transaction dataset, while another may be interested in finding all transactions with large increases in gross sales. This kind of interestingness is based on user-defined utility functions in addition to the raw data.

Actionability/Applicability:

A pattern is actionable (or applicable) in some domain if it enables decision making about future actions in this domain. Actionability is sometimes associated with a pattern selection strategy. So far, no general method for measuring actionability has been devised. Existing measures depend on the applications.

The aforementioned interestingness criteria are sometimes correlated with, rather than independent of, one another. As previously described, conciseness often coincides with generality, and generality often coincides with reduced sensitivity to noise, which is a form of reliability. Also, generality conflicts with peculiarity, while the latter may coincide with novelty.

These nine criteria can be further categorized into three classifications:

1. Objective

2. Subjective

3. Semantics-based

An **objective** measure is based only on the raw data. No knowledge about the user or application is required. Most objective measures are based on theories in probability, statistics, or information theory. Conciseness, generality, reliability, peculiarity, and diversity depend only on the data and patterns, and thus can be considered objective.

A **subjective** measure takes into account both the data and the user of these data. To define a subjective measure, access to the user's domain or background knowledge about the data is required. This access can be obtained by interacting with the user during the data mining process or by explicitly representing the user's knowledge or expectations. Novelty and surprisingness depend on the user of the patterns, as well as the data and patterns themselves, and hence can be considered subjective.

A **semantic** measure considers the semantics and explanations of the patterns. Because semantic measures involve domain knowledge from the user, some researchers consider them a special type of subjective measure. Utility and actionability depend on the semantics of the data, and thus can be considered semantic. Utility-based measures, where the relevant semantics are the utilities of the patterns in the domain, are the most common type of semantic measure. To use a utility-based approach, the user must specify additional knowledge about the domain.

Researchers have proposed interestingness measures for various kinds of patterns, analyzed their theoretical properties, evaluated them empirically, and suggested strategies to select appropriate measures for particular domains and requirements. The most common patterns that can be evaluated by interestingness measures include association rules, classification rules, and summaries.

ASSOCIATION RULES/CLASSIFICATION RULES:

Objective Measures

—Based on probability (generality and reliability)

—Based on the form of the rules

—Peculiarity

—Surprisingness

—Conciseness

—Nonredundant rules

—Minimum description length

Subjective Measures

—Surprisingness

—Novelty

Semantic Measures

—Utility

—Actionability

SUMMARIES:

Objective Measures

—Diversity of summaries

—Conciseness of summaries

—Peculiarity of cells in summaries

Subjective Measures

—Surprisingness of summaries

POSSIBLE QUESTIONS

2 MARK

1. List any two common problems in data mining.
2. Draw a neat sketch on Roles of interestingness measures in the data mining process.
3. Define Diverse Data Types Issues.
4. Define subjective measure.
5. Define objective measure.
6. Define Semantic measure.

8 MARKS

1. Explain scalability in large dataset.
2. Explain major issues in data mining.
3. Explain about Interesting Measures.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore-641021

Department of Computer Science

III BSc(CS) (BATCH 2017-2020)

Data Mining (17CSU602A)

PART-A OBJECTIVE TYPE/ MULTIPLE CHOICE QUESTIONS

ONLINE EXAMINATIONS **ONE MARK QUESTIONS**

UNIT-5

S.No	Question	option1	option2	option3	option4	Answer
1	If we have found some regularities we formulate _____explaining the data	observation	theory	prediction	none	theory
2	_____data warehouse provide an accurate and consistent view of enterprise information	purchase	enterpirse	analysis	sales	enterpirse
3	The _____should have identified the initial user requirements.	technical	warehouse	buisness requirements	process	buisness requirements
4	_____is used to determine what the overall architecture of the data warehouse should be	requirements	architecture	process	technical blueprint	technical blueprint
5	Data warehouse must be architected to support_____major driving factors	3	4	2	5	3
6	_____takes data from source systems and makes it available to the data warehouse	data cleaning	data abstraction	data extraction	none	data extraction
7	clean and	transform	extract	partition	aggregate	transform
8	Make sure data	inconsiste	replicaito	consistent	none	consistent
9	The _____process is the system process that manages	process managem	load managem	system managem	query managemen	query management
10	The _____is the system component that performs all the	process manager	load manager	system manager	query manager	load manager
11	The bulk of the effort to develop a load manager should be planned within the fir _____phase	analysis	design	load	production	production
12	Meta data describes the _____.	type and format of data	structure of the contents of database	location of data	owner of data	structure of the contents of database
13	In bottom up approach _____are used by end-users	data mart	large data warehouse	central database	operational database	data mart

14	_____are used to load the information from the operational database	Statistical Techniques	Visualization Techniques	Windowing mechanism	Replication Techniques	Replication Techniques
15	_____responses end-users queries in a very short space of time.	Client / Server Technique	Time sharing system	Multiprocessing computer system	Real time system	Multiprocessing computer system
16	Expert system contain _____.	spatial data	knowledge of specialists	knowledge of business logic	transaction records	knowledge of specialists
17	OLAP store their data in _____.	table format	object oriented format	a special multi-dimensional format	points in multi-dimensional space	a special multi-dimensional format
18	OLAP tools	do not learn but create new knowledge	do learn but cannot create new knowledge	do learn and also can create new knowledg	do not learn and cannot create new	do not learn and cannot create new knowledge
19	Data mining algorithm should not have a complexity that is higher than_____.	logn	$_n2$	nlogn	n	nlogn
20	Association rules are defined on _____.	single attribute	n attributes	binary attributes	none of the above	binary attributes
21	A perceptrons consists of _____.	two layered networks	single layered networks	three layered networks	multiple layered networks	three layered networks
22	Back propagation method gives _____.	answers and idea as to how they arrived at the answers	answers and no clear idea as to how they arriv	only ideas as to how to obtain answers	answers and poor ideas as to how th	answers and no clear idea as to how they arrived at the answers
23	A Kohonen's self-organizing map is a collection of _____.	networks	neurons	processors	Protons	neurons
24	Genetic algorithms can be viewed as a kind of _____strategy.	self learning	meta learning	knowledge discovering	concept learning	meta learning
25	Voroni diagram divide a sample space into _____.	different groups	different divisions	different sub networks	different regions	different regions
26	Neural networks are somewhat better at _____.	problem solving tasks	classification tasks	knowledge engineering	none of the above	classification tasks

27	SQL retrieves _____.	hidden knowledge	hidden rules	shallow knowledge	deep knowledge	shallow knowledge
28	_____can be found by pattern recognition algorithms	Hidden knowledge	Deep knowledge	Encrypted information	Fine grained segmentation	Hidden knowledge
29	The reporting stage combines _____ _.	analysis of the results & application of the result to new data	the results & application of the result to new dat	analysis of the results & application of the result	a. analysis of the results & applicatio	analysis of the results & applicatio
30	The delivery process is staged in order to _____.	reduce the execution time	to minimize error	minimize risk	measure benefits	minimize risk
31	Delivery process is designed to deliver _____.	an enterprise data warehouse	a point solution	maximum information	quality solution	an enterprise data warehouse
32	Delivery process ensures _____.	to reduce the overall delivery time-slice	benefits are delivered incrementally	to reduce the overall delivery time-slice & be	to reduce investment	to reduce the overall delivery time-
33	The technical blueprint phase must deliver an overall architecture that satisfies the _____requirements.	short term	long term	Current	Past	long term
34	---------- _____is part of day-to-day management of the data warehouse.	Creating / deleting summaries	Extracting data	Cleaning data	Populating data	Creating / deleting summaries
35	The data extracted from the source systems is loaded into a _____.	data warehouse	data mart	temporary data store	operational database	temporary data store
36	The purpose of business case is to identify the projected _____.	out put structure	business process	business benefits	business process risks	business benefits
37	In order to deliver an early release of part of a data warehouse we should _____.	focus on the business requirements	focus on long term requirements	technical blueprint phases	focus on the business requirements	focus on the business requirements &

38	The technical blueprint must deliver _____that satisfies the long-term requirements.	overall system architectu re	overall benefits	overall expenditu re	overall process strategy	overall system architecture
39	_____is not a web based tool	Acrobat	Arbor Essbase web	Informati on advantage web OLAP	Brio technology	Acrobat
40	A high speed decision support from multidimensional database is _____software	Pilot	Applicati on	System	DBMS	Pilot
41	_____is a server component of brio technology	Brio quickview	Brio web	Brio warehous e	Brio technology	Brio quickview
42	_____is data about data	Metadata	Informati on	Tuples	Datawareh oue	Metadata
43	_____table is a large control table in a dimensional design that has a multipart key	Fact	Hierarchi cal	One dimension al	Relational	Fact
44	_____operation system software that can be used to host a datawarehouse	UNIX	Linux	Android	Java	UNIX
45	_____database product specifically optimized for datawarehouse	Red bric warehous e	Sybase	Oracle	DBMS	Red bric warehouse
46	Datawarehouse areas of application in central government secto	Agricultur e	Census	Town planning	Child growth	Agriculture
47	GISTNIC is General Information Services _____Informatics Centre	Terminal of national	Tamil nadu	Transacti on	Terminolog y	Terminal of national
48	Essential commodity prices data for Tamilnadu was compiled using SAS tool	GISTNIC	EOU	EPZ	CEPZ	GISTNIC
49	World bank developed a live database in 1995 called _____	LDB	OLAP	OLTP	GISTNIC	LDB