Q - 2 (a) Explain Constraint-based association mining?

Constraint-based association mining is a data mining technique that aims to discover interesting relationships and dependencies between items in large datasets.

This technique uses constraints to guide the mining process and filter out uninteresting or irrelevant associations, thus allowing the discovery of more meaningful and useful patterns.

The main idea behind constraint-based association mining is to define a set of constraints that specify certain conditions that must be satisfied by the item sets or rules generated by the mining process.

The constraints can include the following which are as follows −

Knowledge type constraints − These define the type of knowledge to be mined, including association or correlation.

Data constraints − These define the set of task-relevant information such as Dimension/level constraints − These defines the desired dimensions (or attributes) of the information, or methods of the concept hierarchies, to be utilized in mining.

Interestingness constraints − These defines thresholds on numerical measures of rule interestingness, including support, confidence, and correlation.

Rule constraints : - Rule mining constraints in data mining are conditions or criteria that are used to limit the set of association rules generated by the mining process.

These constraints are often applied to the statistical measures of the rules, such as support, confidence, and lift, to ensure that only the most interesting and relevant rules are considered.

Q - 2 (a) Explain Apriori Algorithm in detail and List out Limitation of the Apriori algorithm.

Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects. It means how two or more objects are related to one another.

Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.

Components of Apriori algorithm

The given three components comprise the apriori algorithm.

Support: - Support refers to the default popularity of any product. You find the support by

Support (item 1) = (Transactions relating item 1) / (Total transactions)

Confidence: - Confidence refers to the possibility that the customers bought both item 1 and item 2 together.

Confidence = (Transactions relating both item 1 and item 2) / (Total transactions involving item 1)

Lift: - lift refers to the increase in the ratio of the sale of item 2 when you sell item 1.

Lift = (Confidence (item 1 – item 2)/ (Support (item 1)

Q) What are the error measures in Linear regression? Explain with example.

In linear regression, there are several error measures that are commonly used to evaluate the performance of a model

1. Mean Squared Error (MSE): MSE is the most commonly used error measure in linear regression. It measures the average squared difference between the predicted and actual values of the target variable.

   MSE = (1/n) * ∑(yi - ŷi)^2
   where n is the number of data points, yi is the actual value of the target variable, and ŷi is the predicted value of the target variable.

2. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE. It is also commonly used in linear regression to evaluate the performance of the model. The formula for RMS is:

   RMSE = √(MSE)
3. Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual values of the target variable.

   MAE = (1/n) * ∑|yi - ŷi|
   where n is the number of data points, yi is the actual value of the target variable, and ŷi is the predicted value of the target variable.

Example: Here's an example to illustrate how these error measures are calculated:

Suppose we have a linear regression model that is predicting the price of a house based on its size in square feet. We have the following data:

Size (sq ft)	Price ($)
1500	250000
2000	350000
2500	450000
3000	550000
3500	650000

We train the linear regression model on this data and get the following predictions:

Size (sq ft)	Price ($)	Predicted Price ($)
1500	250000	275000
2000	350000	375000
2500	450000	475000
3000	550000	575000
3500	650000	675000

Using these predicted values and the actual values of the target variable, we can calculate the error measures as follows:

MSE = (1/5) * [(250000 - 275000)^2 + (350000 - 375000)^2 + (450000 - 475000)^2 + (550000 - 575000)^2 + (650000 - 675000)^2] = 62500000

RMSE = √(62500000) = 7905.54

MAE = (1/5) * [|250000 - 275000| + |350000 - 375000| + |450000 - 475000| + |550000 - 575000| + |650000 - 675000|] = 20000

Therefore, the MSE is 62500000, the RMSE is 7905.54, and the MAE is 20000.

Q - 2 (b) What do you mean by Data Processing? Explain with suitable application.

Data processing refers to the transformation of raw data into a more useful format that can be easily analyzed, interpreted, and visualized. This process involves several steps such as data cleaning, data integration, data transformation, and data reduction.

However, the processing of data largely depends on the following –

- The volume of data that need to be processed

- The complexity of data processing operations
- Capacity and inbuilt technology of respective computer system
- Technical skills
- Time constraints

Example:

1. E-commerce: In e-commerce, data processing is used to analyze customer behavior, purchase history, and other data related to sales and marketing. By processing this data, companies can gain insights into customer preferences, buying patterns, and other factors that influence purchasing decisions. For example, an e-commerce company might use data processing to analyze customer reviews and feedback to identify areas for improvement in their product offerings.
2. Healthcare: In healthcare, data processing is used to analyze patient data, including medical history, treatment plans, and test results.For example, data processing can be used to analyze patient data to identify risk factors for certain diseases or to predict which treatments are most effective for specific patients.

Q - 3 (a) What is data cleaning? Discuss various ways of handling missing values during data cleaning.

explained

Q - 3 (b) Define the term “Data Mining”. With the help of a suitable diagram explain the process of knowledge discovery from databases.

Explained-

Q - 3 (a) Discuss the importance of Association Rule Mining.

Association rule mining is a data mining technique that helps to identify interesting patterns and relationships in large datasets. The goal of association rule mining is to find frequent patterns or associations among a set of items in a transactional dataset.

Here are some of the reasons why association rule mining is important

1. Market basket analysis: Association rule mining is widely used in market basket analysis to identify frequently occurring combinations of products that are often purchased together. This information can be used to develop targeted marketing strategies, optimize product placement, and increase sales.
2. Fraud detection: Association rule mining can be used to identify patterns of fraudulent behavior, such as credit card fraud, insurance fraud, and money laundering. By analyzing large datasets, association rule mining can help to detect anomalies and identify suspicious transactions.

3. Healthcare: Association rule mining can be used to analyze patient data and identify patterns of disease and treatment. This information can be used to develop personalized treatment plans, improve patient outcomes, and identify potential health risks.
4. Supply chain management: Association rule mining can be used to analyze supply chain data and identify patterns of demand, inventory, and delivery.

Q - 3 (b) How is Data Mining different from OLAP? Explain Briefly.

Data Mining	OLAP
Data mining refers to the field of computer science, which deals with the extraction of data, trends and patterns from huge sets of data.	OLAP is a technology of immediate access to data with the help of multidimensional structures.
It deals with the data summary.	It deals with detailed transaction-level data.
It is discovery-driven.	It is query driven.
It is used for future data prediction.	It is used for analyzing past data.
It has huge numbers of dimensions.	It has a limited number of dimensions.
Bottom-up approach.	Top-down approach.
It is an emerging field.	It is widely used.

Q - 4 Attempt any one/two.

(i) Explain OLAP operations in detail with suitable examples.

OLAP (Online Analytical Processing) is a technology used for analyzing and querying large multidimensional databases, also known as data cubes. OLAP operations enable users to perform complex analysis on large datasets with multiple dimensions, such as time, geography, and product categories.

1. Slice: The slice operation allows users to select a subset of data from the cube by fixing one or more dimensions to a specific value. For example, a user may want to see sales data for a specific time period, such as the first quarter of the year, or for a specific product category, such as electronics.

2.Dice: The dice operation allows users to select a subset of data from the cube by fixing some dimensions and selecting a range of values for other dimensions.

3.Roll-up: The roll-up operation allows users to aggregate data across one or more dimensions. For example, a user may want to see sales data for a specific region, such as North America, and aggregate it across different product categories or time periods.

4.Drill-down: The drill-down operation allows users to view data at a more detailed level by expanding one or more dimensions. For example, a user may want to see sales data for a specific region and then drill down to see sales data for individual stores or cities within that region.

5. Drill-down: The drill-down operation allows users to view data at a more detailed level by expanding one or more dimensions. For example, a user may want to see sales data for a specific region and then drill down to see sales data for individual stores or cities within that region.

(ii) Explain Datawarehouse Architecture in detail.

Explained-

Section 2

Q-1 Answer the Following: (Any five)

(i) What is classification?
Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets

(ii) What is accuracy of a classifier?
The accuracy of a classifier is given as the percentage of total correct predictions divided by the total number of instances.

(iii) What are core points and border points?

(iii) What is regression?
Regression involves the process of fitting a curve or a straight line on various data points. It is done in such a way that the distances between the curve and the data points come out to be the minimum.

(iv) Define clustering.
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups

(vi) Enlist types of regression.

(vii) Suppose you want to know whether a person can have heart disease or not in the future. Which data mining technology you can use?

backpropagation

Q-2(a) What is Decision Tree? Explain how classification is done using decision tree induction.

Decision Tree is a supervised learning method used in data mining for classification and regression methods. A decision tree is a structure that includes a root node, branches, and leaf nodes.

Classification using decision tree induction is a popular machine learning technique that involves building a decision tree model based on a given dataset. The decision tree model represents a tree-like structure, where each internal node represents a test on a specific attribute, each branch represents an outcome of the test, and each leaf node represents a class label.

1.Data preparation: The first step in building a decision tree model is to prepare the dataset by cleaning, pre-processing, and transforming the data as necessary.

2.Attribute selection: The next step is to select the best attribute for each node in the decision tree. The most commonly used attribute selection methods are information gain and gain ratio. These methods measure the amount of information gained by splitting the dataset on a particular attribute.

3.Tree construction: The tree construction process starts with the root node, which represents the entire dataset. The attribute selection method is used to select the best attribute for the root node.

4.Pruning: Once the decision tree is built, it may be necessary to prune the tree to prevent overfitting. Pruning involves removing branches that do not improve the accuracy of the model on the test data.

5.Classification: Once the decision tree model is built and pruned, it can be used for classification. To classify a new instance, the decision tree is traversed from the root node to a leaf node, based on the attribute values of the instance. The class label of the leaf node is then assigned to the instance.

Q-2(b) Explain how rule-based classification works.

Rule-based classification in data mining is a technique in which class decisions are taken based on various “if...then… else” rules.

Let us consider a rule R1,

```
R1: IF age = youth AND student = yes
  THEN buy_computer = yes
```

Points to remember –

- The IF part of the rule is called **rule antecedent** or **precondition**.

- The THEN part of the rule is called **rule consequent**.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.

Note − We can also write rule R1 as follows −

R1: (age = youth) ^ (student = yes))(buys computer = yes)

Q-2(b) What is backpropagation? Explain with suitable diagram.

Backpropagation is an algorithm that backpropagates the errors from the output nodes to the input nodes. Therefore, it is simply referred to as the backward propagation of errors.

Backpropagation is “backpropagation of errors” and is very useful for training neural networks. It’s fast, easy to implement, and simple. Backpropagation does not require any parameters to be set, except the number of inputs. Backpropagation is a flexible method because no prior knowledge of the network is required.

Backpropagation Algorithm:

Step 1: Inputs X, arrive through the preconnected path.

Step 2: The input is modeled using true weights W. Weights are usually chosen randomly.

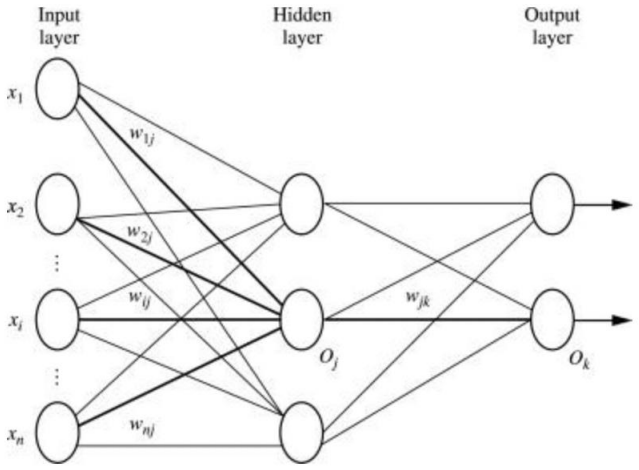
Step 3: Calculate the output of each neuron from the input layer to the hidden layer to the output layer.

Step 4: Calculate the error in the outputs

Backpropagation Error= Actual Output – Desired Output

Step 5: From the output layer, go back to the hidden layer to adjust the weights to reduce the error.

Step 6: Repeat the process until the desired output is achieved.



Q-3(a) Explain ensemble Learner.

• Construct a set of base classifiers learned from the training data

• Predict class label of test records by combining the predictions made by multiple classifiers (e.g., by taking majority vote)

Bagging and Boosting are the two different types of ensemble method

1. Bagging: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.

→Bootstrap Aggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods. Bagging is a special case of the model averaging approach.

2. Boosting: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

→Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added

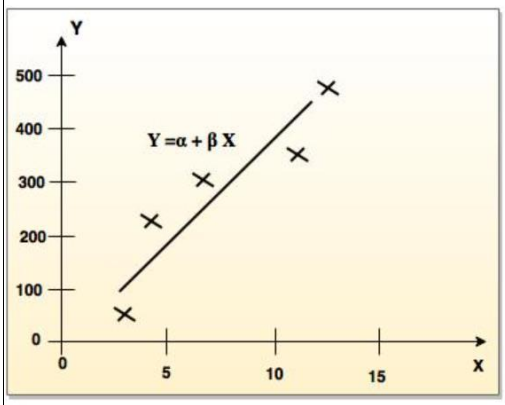
Q-3(b) Explain linear and non-linear regression in detail

Regression refers to a type of supervised machine learning technique that is used to predict any continuous-valued attribute. Regression helps any business organization to analyze the target variable and predictor variable relationships. It is a most significant tool to analyze the data that can be used for financial forecasting and time series modeling.

- It is simplest form of regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data.
- Linear regression attempts to find the mathematical relationship between variables.
- If outcome is straight line then it is considered as linear model and if it is curved line, then it is a non linear model.
- The relationship between dependent variable is given by straight line and it has only one independent variable.

  Y= α+BX

- Model **'Y'**, is a linear function of **'X'**.
- The value of 'Y' increases or decreases in linear manner according to which the value of 'X' also changes.



Linear Regression

2.Non-Linear Regression

The term “nonlinear” refers to the parameters in the model, as opposed to the independent variables.

Unlimited possibilities exist for describing the deterministic part of the model. Such flexibility provides a good ground on which to make statistical inferences.

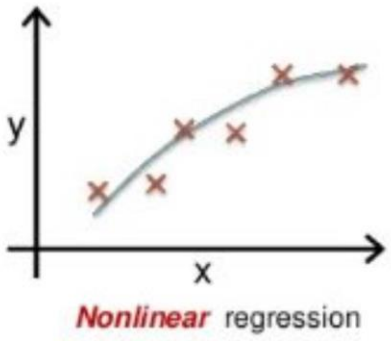
The goal of the model is to minimize the sum of the squares as least as possible using iterative numeric procedures.

A simple nonlinear regression model is expressed as follows:

Where:

$Y = f(X,\beta) + \varepsilon$

- **X** is a vector of P predictors
- **β** is a vector of k parameters
- **F** (-) is the known regression function
- **ε** is the error term



Q-3(a) Explain Grid-Based Clustering

- Grid-Based Clustering: Explore multi-resolution grid data structure in clustering
  - Partition the data space into a finite number of cells to form a grid structure
  - Find clusters (dense regions) from the cells in the grid structure
- Features and challenges of a typical grid-based algorithm
  - Efficiency and scalability: # of cells << # of data points
  - Uniformity: Uniform, hard to handle highly irregular data distributions
  - Locality: Limited by predefined cell sizes, borders, and the density threshold
  - Curse of dimensionality: Hard to cluster high-dimensional data
- Methods to be introduced
  - **STING** (a Statistical Information Grid Approach) (Wang, Yang and Muntz, VLDB'97)
  - **CLIQUE** (Agrawal, Gehrke, Gunopulos, and Raghavan, SIGMOD'98)
    - Both grid-based and subspace clustering

Q-3(b) Define "clustering"? Mention any two applications of clustering.

Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Application

- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.

Q-4 Attempt any one/two.

(i) What is r2 in regression? Explain what is SSE, SSR, and SST in regression and how to find out the values of each term.

- It is one of the criteria to find out whether we are having good regression model.
- We can say the higher it is the better the model in terms of capturing the error prediction.
- Higher the value of R^2 ~ good regression model
- $\mathbf{R^{2}=\frac{SSR}{SST}}$
- $$\frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
- SSR = $\sum(y_i - \hat{y}_i)^2$
- SST = $\sum(y_i - \bar{y})^2$

- The **sum of squares total**, denoted **SST**, is the squared differences between the observed *dependent variable* and its **mean**.

- **sum of squares due to regression**, or **SSR**. It is the sum of the differences between the *predicted* value and the **mean** of the *dependent variable*.
- **sum of squares error**, or **SSE**. The error is the difference between the *observed* value and the *predicted* value.

(ii) DBSCAN.