1) What is the Main Goal of Data Mining?

Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

2) What are the Two Types of 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.

3) What are the Data Mining Functionalities?

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters
- Classification
- Prediction
- Outlier Analysis
- Evolution Analysis

4) What are the Features of WEKA?

Weka features include machine learning, data mining, preprocessing, classification, regression, clustering, association rules, attribute selection, experiments, workflow and visualization.

5) Navigate the options available in the WEKA.

- Preprocess
- Classify
- Cluster
- Associate
- Select Attributes
- Visualize

6) What is ARFF file format?

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

7) Define Support and Confidence.

Support represents the popularity of that product of all the product transactions. Confidence can be interpreted as the likelihood of purchasing both the products A and B.

8) What are the frequent patterns?

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set, is a frequent itemset.

9) Where we are using Apriori Algorithm in Real time scenario?

Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a supermarket. It helps the customers buy their items with ease, and enhances the sales performance of the departmental store.

10) Explain Association rule with a suitable example.

A classic example of association rule mining refers to a relationship between diapers and beers. The example, which seems to be fictional, claims that men who go to a store to buy diapers are also likely to buy beer.

11) What is Apriori Property?

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database

and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

12) How can we further improve the efficiency of Apriori-based mining?

Based on the inherent defects of Apriori algorithm, some related improvements are carried out: 1) using new database mapping way to avoid scanning the database repeatedly; 2) further pruning frequent itemsets and candidate itemsets in order to improve joining efficiency;

13) Define Classification.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

14) Define Prediction.

The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent variables.

15) What are the classification techniques in data mining?

- Logistic Regression
- Naïve Bayes
- Stochastic Gradient Descent
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Support Vector Machine

16) What are the advantages of different classification algorithms

Mining Based Methods are cost effective and efficient. Helps in identifying criminal suspects. Helps in predicting risk of diseases.

17) What is the Kappa Statistic

"The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves.

18) Which classification algorithm is best for prediction and analysis?

- Time Series Model. The time series model comprises a sequence of data points captured, using time as the input parameter. ...
- Random Forest. Random Forest is perhaps the most popular classification algorithm, capable of both classification and regression.

19) What are the classification algorithms in data mining?

Six classification algorithms—Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression

20) Which data mining algorithm provides best accuracy for classification?

The performances of the algorithms were compared according to accuracy, root mean squared error, ROC area, F-measure, precision, and recall criteria, and the logistic regression classification algorithm was found to be the best algorithm.

21) What are the best classification algorithms?

Top 5 Data Mining Algorithms for Classification

- Decision Trees.
- Logistic Regression.
- Naive Bayes Classification.
- k-nearest neighbors.
- Support Vector Machine.

22) Which are the best classifier algorithms for disease predictions?

Comparative analysis of classification techniques has shown that decision tree classifiers are simple and accurate [9]. Naïve Bayes was found to be the best algorithm, followed by neural networks and decision trees

23) Explain Decision Tree.

A decision tree is a diagram or chart that people use to determine a course of action or show a statistical probability. Each branch of the decision tree represents a possible decision, outcome, or reaction.

24) What is the Cross-Validation?

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent data set. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

25) What is the Naïve-Bayes Classification?

It is a **classification** technique based on **Bayes**' Theorem with an assumption of independence among predictors. In simple terms, a **Naive Bayes classifier** assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

26) What are the methods in classification methods in data mining?

There are many **techniques** for solving **classification** problems: **classification** trees, logistic regression, discriminant analysis, neural networks, boosted trees, random forests, deep learning **methods**, nearest neighbors, support vector machines.

27) Briefly explain the K-nearest neighbor algorithm

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (**K**) **closest** to the query, then votes for the most frequent label (in the case of classification) or averages the labels

28) How do you choose an algorithm for a classification problem?

- Size of the training data. It is usually recommended to gather a good amount of data to get reliable predictions.
- Accuracy and/or Interpretability of the output.
- Speed or Training time.
- Linearity.
- Number of features.

29) What are the most useful algorithms used for data mining

K Nearest Neighbors Algorithm, Naïve Bayes Algorithm, SVM Algorithm, ANN Algorithm, 48 Decision Trees, Support Vector Machines, and SenseClusters.

30) What is the difference between classification and prediction?

Classification is the process of identifying the category or class label of the new observation to which it belongs. Predication is the process of identifying the missing or unavailable numerical data for a new observation.

31) What is clustering with example?

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

32) What are the examples of clustering?

- Identifying Fake News. Fake news is not a new phenomenon, but it is one that is becoming prolific.
- Spam filter

- Marketing and Sales
- Classifying network traffic
- Identifying fraudulent or criminal activity
- Document analysis
- Fantasy Football and Sports

33) Why Clustering is used in data mining

Clustering in Data Mining helps in the classification of animals and plants are done using similar functions or genes in the field of biology. It helps in gaining insight into the structure of the species. Areas are identified using the clustering in data mining.

34) Why we use K means clustering

The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

35) What are the advantages and disadvantages of K means clustering

Advantages

Relatively simple to implement.
Scales to large data sets.
Guarantees convergence.

Disadvantages

Clustering data of varying sizes and density.
Clustering outliers.

36) Which is the best clustering algorithm?

- K-means Clustering Algorithm. ...
- Mean-Shift Clustering Algorithm. ...
- DBSCAN – Density-Based Spatial Clustering of Applications with Noise.

37) Explain the Hierarchical Clustering

HCA is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom.

38) State the other Clustering Techniques.

- Connectivity-based Clustering (Hierarchical clustering)
- Centroids-based Clustering (Partitioning methods)
- Distribution-based Clustering.
- Density-based Clustering (Model-based methods)
- Fuzzy Clustering.
- Constraint-based (Supervised Clustering)

39) What are the two types of hierarchical clustering?

- Agglomerative clustering: It's also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. ...
- Divisive hierarchical clustering: It's also known as DIANA (Divise Analysis) and it works in a top-down manner.

40) What are the disadvantages of agglomerative hierarchical clustering?

One drawback is that groups with close pairs can merge sooner than is optimal, even if those groups have overall dissimilarity. Complete Linkage: calculates similarity of the farthest away pair.

41) What is the use of hierarchical clustering?

Hierarchical clustering is the most popular and widely used method to analyze social network data. In this method, nodes are compared with one another based on their similarity. Larger groups are built by joining groups of nodes based on their similarity.

42) What is the difference between K means and hierarchical clustering?

Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.

43) What is the mean, median and mode

The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest. The mode is the number that occurs most often in a data set.

44) How do you find the mode in r?

To find the mode, or modal value, it is best to put the numbers in order. Then count how many of each number. A number that appears most often is the mode.

45) Define analysis of covariance

Analysis of Covariance (ANCOVA) is the inclusion of a continuous variable in addition to the variables of interest (i.e., the dependent and independent variable) as means for control.

46) What is analysis of covariance used for?

Analysis of covariance is used to test the main and interaction effects of categorical variables on a continuous dependent variable, controlling for the effects of selected other continuous variables, which co-vary with the dependent. The control variables are called the "covariates."

47) How do you analyze covariance?

The Analysis of covariance (ANCOVA) is done by using linear regression. This means that Analysis of covariance (ANCOVA) assumes that the relationship between the independent variable and the dependent variable must be linear in nature.

48) Why should we use R?

R is a programming language for statistical computing and graphics that you can use to clean, analyze, and graph your data. It is widely used by researchers from diverse disciplines to estimate and display results and by teachers of statistics and research methods.

49) Define regression and its types

Regression is a technique used to model and analyze the relationships between variables and often times how they contribute and are related to producing a particular outcome together.

The two basic types of regression are simple linear regression and multiple linear regression.

50) What is regression in statistics with example?

Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. ... For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

51) What is meant by linear regression?

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

52) What is multiple linear regression explain with example

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.

53) What is the difference between linear regression and multiple regression?

Linear regression is one of the most common techniques of regression analysis. Multiple regression is a broader class of regressions that encompasses linear and nonlinear regressions with multiple explanatory variables.

54) What is difference between linear and logistic regression

The essential difference between these two is that

Logistic regression is used when the dependent variable is binary in nature.

Incontrast, linear regression is used when the dependent variable is continuous and nature of t he regression line is linear.

55) What is time series analysis with example?

a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

56) How do you analyze time series data in R?

The first thing that you will want to do to analyse your time series data will be to read it into R, and to plot the time series. You can read data into R using the scan() function, which assumes that your data for successive time points is in a simple text file with one column.

57) What is the purpose of time series analysis?

Time series analysis can be useful to see how a given asset, security, or economic variable changes over time. It can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period.

58) What is difference between linear and nonlinear?

While a linear equation has one basic form, nonlinear equations can take many different forms. Thetas represent the parameters and X represents the predictor in the nonlinear functions. Unlike linear regression, these functions can have more than one parameter per predictor variable.

59) What is decision tree and example?

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. ... An example of a decision tree can be explained using above binary tree.

60) How do you create a decision tree in R?

Step 1: Import the data.

Step 2: Clean the dataset.

Step 3: Create train/test set.

Step 4: Build the model.

Step 5: Make prediction.

Step 6: Measure performance.

Step 7: Tune the hyper-parameters.

61) What is the Rnorm function in R?

rnorm is the R function that simulates random variates having a specified normal distribution. As with pnorm , qnorm , and dnorm , optional arguments specify the mean and standard deviation of the distribution.

62) What is the Pnorm and Qnorm in R

pnorm: cumulative density function of the normal distribution. qnorm: quantile function of the normal distribution.

63) What is the normal distribution with example?

For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

64) Where is normal distribution used?

Normal distribution, also called Gaussian distribution, the most common distribution function for independent, randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation.

65) What is Dbinom

The function dbinom returns this probability. There are three required arguments: the value(s) for which to compute the probability (j), the number of trials (n), and the success probability for each trial (p).

66) What is the difference between Pbinom and Dbinom

dbinom is a probability mass function of binomial distribution, while pbinom is a cumulative distribution function of this distribution.

67) Define chi square test and its application

The Chi Square test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. The Chi square test is used to compare a group with a value, or to compare two or more groups, always using categorical data.

68) When should we use chi square test

The Chi Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.

69) What are the limitations of chi square?

Chi-square, like any analysis has its limitations. One of the limitations is that all participants measured must be independent, meaning that an individual cannot fit in more than one category. If a participant can fit into two categories a chi-square analysis is not appropriate.

70) What is the difference between t test and F TEST?

T-test is used to test if two samples have the same mean. The assumptions are that they are samples from normal distribution. F-test is used to test if two samples have the same variance.

71) What is the difference between chi square and t test?

A t-test tests a null hypothesis about two means; most often, it tests the hypothesis that two means are equal, or that the difference between them is zero. A chi-square test tests a null hypothesis about the relationship between two variables.

72) Where do we use chi square test

The Chi Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.