ASSIGNMENT - 7

Submitted by
Abhay Rawat

MACHINE LEARNING

1. B
2. D
3. D
4. A
5. C
6. D
7. D
8. B
9. D
10. A
11. D
12. A

13. The hierarchical cluster analysis follows three basic steps:
    1 - calculate the distances
    2 - link the clusters
    3 - choose a solution by selecting the right number of clusters. First, we have to select the variables upon which we base our clusters.

14. To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

15. Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. Types of clustering are -
1) Centroid Clustering
2) Density Clustering
3) Distribution Clustering
4) Connectivity Clustering

SQL

1. A,D
2. A,B,C
3. B
4. B
5. A
6. C
7. B
8. B
9. B
10. C

11. Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

12. OLAP - Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.

OLTP - Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization. The primary objective is data processing and not data analysis

13. Characteristics of data warehouse are:
    1) Subject-Oriented
    2) Integrated
    3) Non-Volatile
    4) Time Variant

14. Star Schema in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets.

15. ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.

Statistics

1. A
2. A
3. B
4. D
5. C
6. A
7. B
8. A
9. C
10. A normal distribution is a type of continuous probability distribution for a real-valued random variable.
11. Missing data can be handled by using Imputation techniques such as :
    - Simple Imputation
    - Hot Deck Imputation
    - Cold Deck Imputation
    - Regression Imputation
    - Mean Imputation
    - Mode Imputation

12. A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a web-page or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Yes, it is done at times when we have numeric data. It is performed by calculating the mean of the feature and then substituting null values with it.

14. 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.

15. Branches of statistics are :
    - Descriptive Statistics
    - Inferential Statistics