Machine Learning assignment-1

Question no.:	Answer(options):
1	B
2	D
3	D
4	A
5	B
6	D
7	A
8	B
9	D
10	A
11	D
12	A

Q13. How is cluster analysis calculated?

Ans: In cluster analysis, we divide the data into homogeneous and distinct groups. It is a non-supervised learning method in which we draw references from datasets consisting of input data without labeled responses. We try to find some meaningful insight from the data structure and groupings them accordingly.

Firstly we have to choose the centers equal to the number of desired clusters. Then measure the distance of all points from center, then assigning these points to center with the highest similarity. Calculate new group centers by taking mean of all points in a cluster. Repeat the above steps until there is no change in cluster assignment or maximum number of iteration.

Q14. How is cluster quality measured?

Ans:

Q15. What are cluster analysis and its types?

Ans15. Cluster analysis is a type of unsupervised learning method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Different types of cluster analysis are:-

1) **Density-Based Methods:** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.
2) **Hierarchical Based Methods:** The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
    a) **Agglomerative** (*bottom up approach*)
    b) **Divisive** (*top down approach*)
3) **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K-means*
4) **Grid-based Methods:** In this method the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects

Statistics Worksheet-1

Question no.:	Answer(options):
1	a
2	a
3	b
4	d
5	c
6	b
7	b
8	a
9	c

Q10. What do you understand by the term Normal Distribution?

Ans: Normal Distribution is a function that represents the distribution of many random variables as a symmetrical bell-shaped graph. The total area under the curve should be equal to 1. The half of the values is to the right of the centre and exactly half of the values are to the left of the centre. The Normal Distribution is defined by the probability density function for a continuous random variable in a system.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans: Missing Data can be handling in two ways- Deletion and imputation.

In deletion we have three following ways:

a) List-wise deletion removes all data for an observation that has one or more missing values. Particularly if the missing data is limited to a small number of observations
b) Pair-wise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. A strength to this technique is that it increases power in your analysis
c) Dropping variables if the data is missing for more than 60% observations but only if that variable is insignificant.

Some of the imputation methods are:

a) Inter-quartile range (IQR): The first quartile (Q1) is 25^{th} percentile of the data set, second quartile (Q2) is 50^{th} percentile (median) of the data set, third quartile (Q3) is 75^{th} of the data set. IQR is the difference of Q3 & Q1 i.e. IQR=Q3-Q1.
   Thus a data value x is outlier if either $x \leq Q1-1.5*IQR$ or $x \geq Q3+1.5*IQR$.
b) Z-score: The z-score of a particular data set shows how many standard deviation the data value lies above and below the mean.

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

Q12. What is A/B testing?

Ans. An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

Q13. Is mean imputation of missing data acceptable practice?

Ans. In general, it is bad practice to impute missing data with mean.

Q14. What is linear regression in statistics?

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.

A linear regression line has an equation of the form

Y = a + bX

Where, ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

Q15. What are the various branches of statistics?

Ans. The two main branches of statistics are descriptive statistics and inferential statistics. Descriptive statistics is used to say something about a set of information that has been collected only. Inferential statistics is used to make predictions or comparisons about larger group (a population) using information gathered about a small part of that population.

SQL_Worksheet-1

Question no.:	Answer(options):
1	a&d
2	a,b&c
3	b
4	b
5	a
6	c
7	b
8	b
9	b
10	a

Q11. What is data-warehouse?

Ans. A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources.

Q12. What is the difference between OLTP VS OLAP?

Ans. The difference between OLTP and OLAP are as follows:-

1) The point that distinguishes OLTP and OLAP is that OLTP is an online transaction system whereas; OLAP is an online data retrieval and analysis system.
2) Online transactional data becomes the source of data for OLTP. However, the different OLTPs database becomes the source of data for OLAP.
3) OLTP's main operations are insert, update and delete whereas; OLAP's main operation is to extract multidimensional data for analysis.
4) OLTP has short but frequent transactions whereas; OLAP has long and less frequent transaction.
5) Processing time for the OLAP's transaction is more as compared to OLTP.
6) OLAPs queries are more complex with respect OLTPs.
7) The tables in OLTP database must be normalized whereas; the tables in OLAP database may not be normalized.
8) As OLTPs frequently executes transactions in database, in case any transaction fails in middle it may harm data's integrity and hence it must take care of data integrity. While in OLAP the transaction is less frequent hence, it does not bother much about data integrity.

Q13. What are the various characteristics of data-warehouse?

Ans: The characteristics of data-warehousing are:-

1) Subject oriented: A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations. Data warehousing process is proposed to handle with a specific theme which is more defined.
2) Integrated: Integration means founding a shared entity to scale the all similar data from the different databases. A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database.
3) Time variant: It founds various time limits which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective.
4) Non-volatile: the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. Data is read-only and refreshed at particular intervals. This is beneficial in analyzing historical data and in comprehension the functionality.

Q14. What is Star-Schema??

The star schema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables referencing any number of dimension tables. The star schema separates business process data into facts, which hold the measurable, quantitative data about a business, and dimensions which are descriptive attributes related to fact data.

Q15. What do you mean by SETL?

Ans.