MACHINE LEARNING ASSIGNMENT - 1

1. A) 2

2. D) 1,2, and 4

3. D) Formulating the clustering problem

4. A) Euclidean distance

5. B) Divisive clustering

6. D) All answers are correct

7. A) Divide the data points into groups

8. B) Unsupervised learning

9. D) All of the above

10. A) K-Means clustering algorithm

11. D) All of the above

12. A) Labeled data

13. Cluster Analysis

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. Cluster quality

To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

Silhouette Method: This technique measures the separability between clusters. First, an average distance is found between each point and all other points in a cluster. Then it measures the distance between each point and each point in other clusters.

15. Cluster Analysis

Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg.,

products, respondents, or other entities) based on a set of user selected characteristics or attributes.

Types :

1. Partitioning Clustering Method.

2. Hierarchical Clustering Methods.

3. Density-Based Clustering Method.

4. Grid-Based Clustering Method.

5. Model-Based Clustering Methods.

6. Constraint-Based Clustering Method.

SQL WORKSHEET - 1

1. A), D) Create , ALTER

2. A), B) Update, Delete

3. B) Structured Query Language

4. B) Data Definition Language

5. A) Data Manipulation Language

6. C) Create table A (B int, C float)

7. B) Alter Table A add column D float

8. B) Alter Table A drop column D

9. C) Alter Table A D float int

10. C) Alter Table A add Primary key B

11. Data warehouse

A data warehouse is a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprisewide data analysis and reporting for predefined business needs.

The five components of a data warehouse are:

Production data sourcesData extraction and conversionData warehouse database management systemData warehouse administrationBusiness intelligence (BI) tools

A data warehouse contains data arranged into abstracted subject areas with time-variant versions of the same records, with an appropriate level of data grain or detail to make it useful across two or more different types of analyses most often deployed with tendencies to third normal form. A data mart contains similarly time-variant and subject-oriented data, but with relationships implying dimensional use of data wherein facts are distinctly separate from dimension data, thus making them more appropriate for single categories of analysis.

12. Difference Between OLTP and OLAP

DIFFERENCE	OLTP	OLAP
Data source	Operational data. OLTP systems are the original data sources.	Consolidation data. OLAP data comes from the OLTP databases.
Use	Responsible for controlling and running basic business tasks.	Responsible for planning, xproblem-solving and supporting business decisions.
Queries	Queries are standard and straightforward	Responsible for planning, problem-solving and supporting business decisions.
Speed of processing	Fast speed.	Complex queries can take a long time to process.
Backup and	Frequent	No regular

recovery	complete backups along with incremental backups.	backups. Instead, OLTP data is reloaded as a recovery method.
Process	Online transactional system.	Online analysis and data retrieving process.
Method used	Uses traditional DBMS.	Uses a data warehouse.
Quality of data	Detailed organisation of data.	Disorganised data.
Nature of audience	Market-oriented process.	Customer-oriented process.
Database design	Application-oriented design.	Subject-oriented design.
Types of users	Clerks, online shoppers, etc., use OLTP.	Data knowledge workers like managers and CEOs use OLAP
Productivity	Enhances the productivity of the user.	Enhances the productivity of business analysts.
Updates	The user starts the data updates, which are short and fast.	Regular refreshing of data with long, scheduled batch jobs.

13. Characteristics of Data Warehouse

Subject-Oriented

A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.

INTEGRATED

A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.

Time-variant

Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.

Non-volatile

The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed. It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval. Non-Volatile defines that once entered into the warehouse, and data should not change.

14. Star schema

Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is SETa necessary cause of the snowflake schema. It is also efficient for handling basic queries.

It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star’s points.

Model of Star Schema :

In Star Schema, Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data. Sales price, sale quantity, distant, speed, weight, and weight measurements are few

examples of fact data in star schema.

Often, A Star Schema having multiple dimensions is termed as Centipede Schema. It is easy to handle a star schema which have dimensions of few attributes.

15. SETL

SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets. It was originally developed by (Jack) Jacob T. Schwartz at the New York University (NYU) Courant Institute of Mathematical Sciences in the late 1960s.

SETL provides two basic aggregate data types: unordered sets, and sequences (the latter also called tuples). The elements of sets and tuples can be of any arbitrary type, including sets and tuples themselves. Maps are provided as sets of pairs (i.e., tuples of length 2) and can have arbitrary domain and range types. Primitive operations in SETL include set membership, union, intersection, and power set construction, among others.

STATISTICS WORKSHEET - 1

1. A) True

2. A) Central Limit Theorem

3. B) Modelling bounded count data

4. D) All of the mentioned

5. C) Poisson

6. B) False

7. B) Hypothesis

8. A) 0

9. c) Outliers cannot conform to the regression relationship

10. Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

In graphical form, the normal distribution appears as a "bell curve".

11. Mishandling Data

This occurs when sensitive information is copied, shared, accessed, stolen or otherwise used by an employee who isn't authorised to do so.

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data. The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low.

Some methods to prevent mishandling data are :

Implement identity and access management. ...

Establish need-to-know access. ...

Set up behavior alerts and analytics. ...

Educate your teams. ...

Build clear processes around data access.

12. A/B Testing

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

13. Mean imputation of missing data

It is a bad practice in general.

Mean imputation preserves the mean of the observed dataLeads to an underestimate of the standard deviation

Distorts relationships between variables by "pulling" estimates of the correlation toward zero

14. Linear Regression

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

15. Various Branches of Statistics

There are three real branches of statistics:

1. data collection

2. descriptive statistics

3. Inferential statistics.