Machine Learning

1. B) 4
2. D) 1,2 and 4
3. D) Formulating clustering problem
4. A) Euclidean Distance
5. D) K-Mean 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-Mean clustering algorithm
11. D) All of the above.
12. A) Labeled Data

14. The quality of a clustering result depends on

- the similarity measure used
- implementation of the similarity measure

The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Requirements of Clusterin

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape

- Minimal requirements for domain knowledge to determine input parameters
- Ability to deal with noise and outliers
- Insensitivity to the order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

15. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data analysis, and a common technique for statistical data analysis

Types of Clustering –

Hierarchical clustering

K Mean Clustering

Density Based

Grid Based

SQL

1. A) Create and D) Alter
2. A) Update, B) Delete and C) Select
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. B) Alter Table A Alter Column D int
10. C) Alter Table A Add Primary key B

11. Data Warehouse –

    Data warehousing is the electronic storage of a large amount of information by a business or organization. Data warehousing is a vital component of business intelligence that employs analytical techniques on business data. Data warehousing is used to provide greater insight into the performance of a company by comparing data consolidated from multiple heterogeneous sources.

12. difference between OLTP VS OLAP? -

OLTP and OLAP both are the online processing systems. OLTP is a transactional processing while OLAP is an analytical processing system. OLTP is a system that manages transaction-oriented applications on the internet for example, ATM. OLAP is an online system that reports to multidimensional analytical queries like financial reporting, forecasting, etc.

The basic difference between OLTP and OLAP is that OLTP is an online database modifying system, whereas, OLAP is an online database query answering system.

13. Characteristics of Data warehouse –

1. **Subject Oriented –** That means the data warehousing process is proposed to handle with a specific theme which is more defined.
2. **Integrated –** It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases
3. **Time-Variant –** In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process.
4. **Non-Volatile –** As the name defines the data resided in data warehouse is permanent. It also means that

data is not erased or deleted when new data is inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantity on logical business

14. Star Schema –

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.

It 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.

Statistics

1. A) True

2. A) Central Limit Theorem

3. B) Modeling 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 is a type of continuous probability distribution for a real-valued random variable.

The simplest case of a normal distribution is known as the standard normal distribution. This is a special case when mean is 0 and standard deviation is 1.

Data which is Normally distributed forms a bell shape curve and follows the rule of standard deviation i.e. 97% of the data will lie in 3 s.d. function.

11. There are various type of missing data which requires different techniques for imputation such as –

Missing completely random

Missing at random

Not missing at random.

And we use various imputation techniques such as –

Mean/Median imputation – when our data set is continuous and is making a bell shape curve i.e. forming a Normal Distribution curve then I impute it with the mean so that there won't be any outliers.

In-case of categorical data I like to replace the missing values with mode values again keeping in mind of the outliers.

There are some other imputations too such as Backward and forward propagation, Multivariate Imputation and Random Forest Imputation.

12. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

13. Mean Imputation is considered as a bad practice by many people as it leads to two major problems.

- It does not preserve the relationship among variables.

  If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate. It will still bias your standard error.

- It leads to underestimate of Standard errors.

  It is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low, Because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.

14. In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression, for more than one, the process is called multiple linear regression.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters.