MACHINE LEARNING

1. c
2. a
3. d
4. a
5. b
6. d
7. d
8. b
9. a
10. a
11. d
12. a

13.How is cluster analysis calculated?

Three important factors by which clustering can be evaluated are:*(a)* Clustering tendency *(b)* Number of clusters, **k** *(c)* Clustering quality

Clustering tendency: Before evaluating the clustering performance, making sure that data set we are working has clustering tendency and does not contain uniformly distributed points is very important. If the data does not contain clustering tendency, then clusters identified by any state of the art clustering algorithms may be irrelevant. Non-uniform distribution of points in data set becomes important in clustering.

To solve this, Hopkins test, a statistical test for spatial randomness of a variable, can be used to measure the probability of data points generated by uniform data distribution.

N**ull Hypothesis (Ho) :** Data points are generated by uniform distribution (implying no meaningful clusters)

Alternate Hypothesis (Ha): Data points are generated by random data points (presence of clusters)

If H>0.5, null hypothesis can be rejected and it is very much likely that data contains clusters. If H is more close to 0, then data set doesn't have clustering tendency.

Number of Optimal Clusters, *k*

Some of the clustering algorithms like K-means, require number of clusters, *k*, as clustering parameter. Getting the optimal number of clusters is very significant in the analysis. If *k* is too high, each point will broadly start representing a cluster and if *k* is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering.

Clustering quality

Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterised by minimal intra cluster distance and maximal inter cluster distance.

There are majorly two types of measures to assess the clustering performance.

(i) *Extrinsic Measures* which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) *Intrinsic Measures* that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

14.How is cluster quality measured?

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. What is cluster analysis and its types?

Clustering is an unsupervised machine learning algorithm. It helps in clustering data points to groups. Validating the clustering algorithm is bit tricky compared to supervised machine learning algorithm as clustering process does not contain ground truth labels. If one want to do clustering with ground truth labels being present, validation methods and metrics of supervised machine learning algorithms can be used. This blog post tries to address evaluation strategies when ground truth labels are not known.

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.

Four basic types of cluster analysis used in data science. These types are Centroid Clustering, Density Clustering Distribution Clustering, and Connectivity Clustering.

WORKSHEET 1 SQL

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

11. What is data-warehouse?

A **Data Warehousing** 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.

12.What is the difference between OLTP VS OLAP?

Online Analytical Processing (OLAP) is a category of software tools that analyze data stored in a database whereas Online transaction processing (OLTP) supports transaction-oriented applications in a 3-tier architecture.

OLAP creates a single platform for all type of business analysis needs which includes planning, budgeting, forecasting, and analysis while OLTP is useful to administer day to day transactions of an organization.

OLAP is characterized by a large volume of data while OLTP is characterized by large numbers of short online transactions.

In OLAP, data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database whereas OLTP uses traditional DBMS.

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

The key characteristics of a data warehouse are as follows:

- Some data is denormalized for simplification and to improve performance
- Large amounts of historical data are used
- Queries often retrieve large amounts of data
- Both planned and ad hoc queries are common
- The data load is controlled

14.What is Star-Schema?

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. What do you mean by 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.

STATISTICS WORKSHEET-1

1. a
2. a
3. c
4. a
5. c
6. b
7. b
8. a
9. c
10. **What do you understand by the term 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 graph form, normal distribution will appear as a bell curve.

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

Assign a unique category to missing values, who knows the missing values might decipher some trend. We can remove them blatantly. Or, we can sensibly check their distribution with the target variable, and if found any pattern we'll keep those missing values and assign them a new category while removing others.

12. What is A/B testing?

A/B testing (also known as split testing) is a process of showing two variants of the same web page to different segments of website visitors at the same time and comparing which variant drives more conversions.

13. Is mean imputation of missing data acceptable practice?

Bad practice in general

14.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. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

15. What are the various branches of statistics

There are two branches of statistics:

- Descriptive statistics
- Inferential statistics