Machine learning assignment-1 answers

Q1. Option B

Q2. Option D

Q3. Option D

Q4. Option A

Q5. Option B

Q6. Option D

Q7. Option A

Q8. Option B

Q9. Option A

Q10. Option a

Q11. Option D

Q12. Option A

Q13. How is cluster analysis calculated?

Ans: - 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.

Q14. How is cluster quality measured?

Ans:- 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. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.

Q15. What is cluster analysis and its types?

Ans:- 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. The most common applications of cluster analysis in a business setting is to segment customers or activities.

In this post we will explore four basic types of cluster analysis used in data science. These types are

- Centroid Clustering
- Density Clustering
- Distribution Clustering
- Connectivity Clustering.

SQL worksheet 1 answers

Q1. Option A and D

Q2. Option A and C

Q3. Option B

Q4. Option D

Q5. Option A

Q6. Option C

Q7. Option B

Q8. Option B

Q9. Option D

Q10. Option C

Q11. What is data-warehouse?

Ans: - 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. The data warehouse is the core of the BI system which is built for data analysis and reporting.

Q12. What is the difference between OLTP vs OLAP?

Ans: -

BASIS OF COMPARISON	OLTP	OLAP
Basic	It is an online transactional system and manages database modification.	It is an online data retrieving and data analysis system.

Focus	Insert, Update, Delete information from the database.	Extract data for analyzing that helps in decision making.
Data	OLTP and its transactions are the original source of data.	Different OLTPs database becomes the source of data for OLAP.
Transaction	OLTP has short transactions.	OLAP has long transactions.
Time	The processing time of a transaction is comparatively less in OLTP.	The processing time of a transaction is comparatively more in OLAP.

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

Ans: - There are three prominent data warehouse characteristics:

- **Integrated**: The way data is extracted and transformed is uniform, regardless of the original source.
- **Time-variant**: Data is organized via time-periods (weekly, monthly, annually, etc.).
- **Non-volatile**: A data warehouse is not updated in real-time. It is periodically updated via the uploading of data, protecting it from the influence of momentary change.

Q14. What is star-schema?

Ans: - **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.

Q15. What do you mean by SETL?

Ans: - Short for Set Theory as a Language (or Set Language), SETL is a high-level programming language that's based on the mathematical theory of sets.

Statistics worksheet 1 answers

Q1. Option A

Q2. Option A

Q3. Option B

Q4. Option D

Q5. Option C

Q6. Option B

Q7. Option B

Q8. Option A

Q9. Option C

Q10. What do you understand by the term normal distribution?

Ans: - 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.

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

Ans: - Missing data reduces the representativeness of the sample and can therefore distort inferences about the population. Generally speaking, there are three main approaches to handle missing data:

- Imputation—where values are filled in the place of missing data,
- omission—where samples with invalid data are discarded from further analysis
- analysis—by directly applying methods unaffected by the missing values.

Q12. What is A/B testing?

Ans: - 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.

Q13. Is mean imputation of missing data acceptable practice?

Ans: - It is a bad practice in general. Mean imputation preserves the mean of the observed data which leads to an underestimate of the standard deviation. This distorts relationships between variables by "pulling" estimates of the correlation toward zero.

Q14. What is linear regression in statistics?

Ans: - In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression. For more than one, the process is called multiple linear regression.

Q15. What are the various branches of statistics?

Ans: - If we consider the branches of statistics, there are two branches in it. They are

- **Descriptive statistics**
- **Inferential statistics**

Descriptive-statistics:

It organizes raw data into meaningful information. An house hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full. Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this

particular branch “ descriptive-statistics". That is, it focuses on collecting, summarizing and presenting set of data.

Inferential-statistics:

It analyses sample data to draw conclusion about population. It analyses sample data to draw conclusion about population. Marketing research team of a company wants to know how far the people need a particular product manufactured by the company. There is one hundred thousand population in a particular city. It is bit difficult to go and ask all one hundred thousand people, due to time consumption and other factors. Hence, it takes a sample of 1000 people to draw conclusion for the whole population. That is making general statement from the study of particular cases or any treatment of data, which leads to prediction or inference concerning a larger group of data.