Assignment- 1

MACHINE LEARNING

Answers

1. A
2. D
3. D
4. A
5. B
6. D
7. A
8. B
9. D
10. A
11. D
12. A

13. How is cluster analysis calculated?

A. Grouping of data objects in a way that the objects which are most similar are grouped together and others are grouped in different groups

Cluster analysis is used to identify the structure in the data

1. Calculate the distances.
2. Link the clusters.
3. Choose a solution by selecting the right number of clusters.

14. How is cluster quality measured?

A. We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.

If ground truth is available, it can be used by **extrinsic methods**, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of cluster labels. Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

15. What is cluster analysis and its types?

A. Grouping of data objects in a way that the objects which are most similar are grouped together and others are grouped in different groups

Types of Cluster Analysis

1.Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify.

The algorithm will start by randomly selecting centroids to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iterations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters.

2.Density Clustering

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

3.Distribution Clustering

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid the algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions the algorithm optimizes the characteristics of the distributions to best represent the data.

These maps look a lot like targets at an archery range. In the event that a data point hits the bull's eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bull's eye represents lessening percentage or certainty.

Distribution clustering is a great technique to assign outliers to clusters, whereas density clustering will not assign an outlier to a cluster.

4. Connectivity Clustering

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.

Assignment – 2

SQL

Answers

1. A, D
2. A, B, C
3. B
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. The data warehouse is the core of the BI system which is built for data analysis and reporting.

The data stored in the warehouse is uploaded from different sections of operations. The data may pass through an operational data store and may require data cleansing for additional operations to ensure data quality before it is used in the DW for reporting.

Extract, transform, load (ETL) and extract, load, transform (ELT) are the two main approaches used to build a data warehouse system.

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

Example for OLAP:

A company might compare their mobile phone sales in September with sales in October, then compare those results with another location which may be stored in a sperate database.

Amazon analyses purchases by its customers to come up with a personalized homepage with products which likely interest to their customer.

Example for OLTP:

- Online banking
- Online airline ticket booking
- Sending a text message
- Order entry
- Add a book to shopping cart

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

- 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

Subject-oriented –

A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations.

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.

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 (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems.

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. It evaluates the analysis within the technologies of warehouse. Two types of data operations done in the data warehouse are:

- Data Loading
- Data Access

14. What is Star-Schema?

A. A star schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model. The star schema is the explicit data warehouse schema. It is known as

star schema because the entity-relationship diagram of this schemas simulates a star, with points, diverge from a central table. The centre of the schema consists of a large fact table, and the points of the star are the dimension tables.

Dimension Tables

A dimension is an architecture usually composed of one or more hierarchies that categorize data. If a dimension has not got hierarchies and levels, it is called a flat dimension or list. The primary keys of each of the dimension's table are part of the composite primary keys of the fact table. Dimensional attributes help to define the dimensional value. They are generally descriptive, textual values. Dimensional tables are usually small in size than fact table.

Fact tables store data about sales while dimension tables data about the geographic region (markets, cities), clients, products, times, channels.

Characteristics of Star Schema

- The star schema is intensely suitable for data warehouse database design because of the following features:
- It creates a DE-normalized database that can quickly provide query responses.
- It provides a flexible design that can be changed easily or added to throughout the development cycle, and as the database grows.
- It provides a parallel in design to how end-users typically think of and use the data.
- It reduces the complexity of metadata for both developers and end-users.

15. What do you mean by SETL?

A. SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets. It was originally developed by Jacob T. Schwartz.

SETL provides two basic aggregate data types: unordered sets, and sequences (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 and can have arbitrary domain and range types. Primitive operations in SETL include set membership, union, intersection, and power set construction, among others.

SETL provides quantified Boolean expressions constructed using the universal and existential quantifiers of first-order predicate logic.

SETL provides several iterators to produce a variety of loops over aggregate data structures.

Assignment-3

Statistics

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

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

A. A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same.

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

Most important and time taking task while cleaning the data is handling the missing data there are many techniques used to handle it.

- The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion.
- Replacing the missing values with Mean, Median or Mode depending on the type of numeric data will be a better imputation technique.

The many imputation techniques can be divided into two subgroups: single imputation or multiple imputation.

- In single imputation, a single imputation value for each of the missing observations is generated. The imputed value is treated as the true value, ignoring the fact that no imputation method can provide the exact value. Therefore, single imputation does not reflect the uncertainty of the missing values.
- In multiple imputation, many imputed values for each of the missing observations are generated. This means many complete datasets with different imputed values are created. The analysis is performed on each of these datasets and the results are polled. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations.

12. What is A/B testing?

A. 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.

In Simple words we will test out best possible way of completing a task and check which possible way is performing the best and choose that according to the hypothesis created around the result.

The hypothesis

We are considering a Website Layout

A hypothesis is a formal statement describing the relationship you want to test. A hypothesis must be a simple, clear and testable statement that contrasts a control sample (e.g., Layout A) with a treatment sample (e.g., Layout B).

To form a hypothesis, we re-phrase “does an SMS system improve repayment” into two statements, a null hypothesis and an alternative hypothesis:

Null hypothesis (H0): The null hypothesis usually states that there is no difference between treatment and control groups. (To put this another way, we’re saying our treatment outcome will be statistically similar to our control outcome)

Alternative hypothesis (H1): The alternative hypothesis states that there is a difference between treatment and control groups. (In other words, the treatment outcome will be statistically different to the control outcome).

Thus, the hypothesis is tested and conclusions are made.

13.Is mean imputation of missing data acceptable practice?

A. Mean imputation is one of the widely followed techniques in filling the missing data.

But there are few drawbacks in doing the same

- Mean imputation reduces a variance of the data
- Mean and mode ignore feature correlations

Following are the alternative imputation algorithms

- MICE
- KNN
- MissForest
- Fuzzy K-means Clustering

Following are the python libraries

- Fancyimpute
- Impyute
- missingpy

14. What is linear regression in statistics?

A. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

- Which variables in particular are significant predictors of the outcome variable, and in what way do they impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = mx+c,

y = estimated dependent variable score, c = constant, m = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis

- First, the regression might be used to identify the strength of the effect that the independent variable has on a dependent variable.
- Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.
- Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates.

15.What are the various branches of statistics?

There are basically two types of statistics

1.Descriptive Statistics

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

2.Inferential Statistics

Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.