WORKSHEET 6 SQL

1. Which of the following are TCL commands?

C. Rollback

D. Savepoint

2. Which of the following are DDL commands?

A. Create

C. Drop

D. Alter

3. Which of the following is a legal expression in SQL?

B. SELECT NAME FROM SALES;

4. DCL provides commands to perform actions like-

C. Authorizing Access and other control over Database

5. Which of the following should be enclosed in double quotes?

B. Column Alias

6. Which of the following command makes the updates performed by the transaction permanent in the database?

B. COMMIT

7. A sub query in an SQL Select statement is enclosed in:

A. Parenthesis - (...).

8. The result of a SQL SELECT statement is a :-

C. TABLE

9. Which of the following do you need to consider when you make a table in a SQL?

D. All of the mentioned

10. If you don't specify ASC and DESC after a SQL ORDER BY clause, the following is used by___?

A. ASC

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What is denormalization?

Denormalization is a database optimization technique in which we add redundant data to one or more tables.

This can help us avoid costly joins in a relational database. Denormalization does not mean not doing normalization.

It is an optimization technique that is applied after doing normalization.

Pros of Denormalization:-

Retrieving data is faster since we do fewer joins.

Queries to retrieve can be simpler (and therefore less likely to have bugs), since we need to look at fewer tables.

Under denormalization, we decide that we're okay with some redundancy and some extra effort to update the database in order to get the efficiency advantages of fewer joins.

12. What is a database cursor?

A database cursor is an identifier associated with a group of rows. It is, in a sense, a pointer to the current row in a buffer.

Statements that return more than one row of data from the database server: A SELECT statement requires a select cursor.

Cursors are used by database programmers to process individual rows returned by database system queries.

Cursors enable manipulation of whole result sets at once. In this scenario, a cursor enables the sequential processing of rows in a result set.

A cursor is a temporary work area created in the system memory when a SQL statement is executed.

A cursor contains information on a select statement and the rows of data accessed by it.

This temporary work area is used to store the data retrieved from the database, and manipulate this data.

13. What are the different types of the queries?

Five types of SQL queries are:

1) Data Definition Language (DDL)

Data Definition Language(DDL) helps you to define the database structure or schema.

2) Data Manipulation Language (DML)

Data Manipulation Language (DML) allows you to modify the database instance by inserting, modifying, and deleting its data.

3) Data Control Language(DCL)

DCL (Data Control Language) includes commands like GRANT and REVOKE, which are useful to give "rights & permissions."

4) Transaction Control Language(TCL)

Transaction control language or TCL commands deal with the transaction within the database.

5) Data Query Language (DQL)

Data Query Language (DQL) is used to fetch the data from the database.

14. Define constraint?

SQL constraints are used to specify rules for the data in a table. Constraints are used to limit the type of data that can go into a table.

This ensures the accuracy and reliability of the data in the table.

Constraints can be column level or table level. Column level constraints apply to a column, and table level constraints apply to the whole table.

The following constraints are commonly used in SQL:

NOT NULL - Ensures that a column cannot have a NULL value

UNIQUE - Ensures that all values in a column are different

PRIMARY KEY - A combination of a NOT NULL and UNIQUE. Uniquely identifies each row in a table

FOREIGN KEY - Prevents actions that would destroy links between tables

CHECK - Ensures that the values in a column satisfies a specific condition

DEFAULT - Sets a default value for a column if no value is specified

CREATE INDEX - Used to create and retrieve data from the database very quickly

15. What is auto increment?

Auto-increment allows a unique number to be generated automatically when a new record is inserted into a table.

Often this is the primary key field that we would like to be created automatically every time a new record is inserted. Auto increment is used with the INT data type. The INT data type supports both signed and unsigned values. Unsigned data types can only contain positive numbers.

MySQL uses the AUTO_INCREMENT keyword to perform an auto-increment feature.

By default, the starting value for AUTO_INCREMENT is 1, and it will increment by 1 for each new record.

MACHINE LEARNING

In Q1 to Q5, only one option is correct, choose the correct option:

1. In which of the following you can say that the model is over fitting?

B) Low R-squared value for the train-set and High R-squared value for test-set

2. Which among the following is a disadvantage of decision trees?

B) Decision trees are highly prone to over fitting.

3. Which of the following is an ensemble technique?

C) Random Forest

4. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important.

In this case which of the following metrics you would focus on?

C) Precision

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85.

Which of these two models is doing better job in classification?

B) Model B

In Q6 to Q9, more than one options are correct, Choose all the correct options:

6. Which of the following is the regularization technique in Linear Regression?

A) Ridge

D) Lasso

7. Which of the following is not an example of boosting technique?

B) Decision Tree

8. Which of the techniques are used for regularization of Decision Trees?

A) Pruning

C) Restricting the max depth of the tree

9. Which of the following statements is true regarding the Ad boost technique?

Q10 to Q15 are subjective answer type questions, Answer them briefly.

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

11. Differentiate between Ridge and Lasso Regression.

Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights.

Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros.

Ridge takes a step further and penalizes the model for the sum of squared value of the weights.

Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed.

Lasso Regression is generally used when we have more number of features, because it automatically does feature selection. Whereas, if we have less number of features or we don't want to loose any feature we can use Ridge Regression.

Lasso Regression: It is more similar to Ridge Regression but perform automatic variable selection.

It allows regression coefficient to be zero whereas Ridge does not.

Ridge Regression: It is used to solve multi collinearity in OLS regression models through the incorporation of shrinkage parameter.

The assumptions for the model is same as OLS model like linearity, constant variance and independence and normality not need to be assumed.

Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso go with l1.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modeling?

A variance inflation factor detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors

in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model.

The variance inflating factor (VIF) is used to prove that the regressors do not correlate among each other.

If VIF>10, there is collinearity and you cannot go for regression analysis. If it is "VIF<10", there is not co linearity and is acceptable.

13. Why do we need to scale the data before feeding it to the train the model?

Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort.

So this more significant number starts playing a more decisive role while training the model.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model.

Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

14. What are the different metrics which are used to check the goodness of fit in linear regression?

Five metrics give us some hints about the goodness-of-fit of our model.

The first two metrics, the Mean Absolute Error and the Root Mean Squared Error, have the same unit as the original data.

The RMSE is the square root of the variance of the residuals.

It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.

Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.

R-squared is simply the fraction of response variance that is captured by the model. If R-squared = 1, means the model fits the data perfectly.

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Actual/Predicted	True	False
True	1000	50
False	250	1200

-

STATISTICS WORKSHEET- 6

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following can be considered as random variable?

d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

a) Discrete

3. Which of the following function is associated with a continuous random variable?

a) pdf

4. The expected value or _____ of a random variable is the center of its distribution.

c) Mean

5. Which of the following of a random variable is not a measure of spread?

a) Variance

6. The _____ of the Chi-squared distribution is twice the degrees of freedom.

a) Variance

7. The beta distribution is the default prior for parameters between _____

c) 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

b) Bootstrap

9. Data that summarize all observations in a category are called _____ data.

b) Summarized

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What is the difference between a boxplot and histogram?

Histograms and box plots are graphical representations for the frequency of numeric data values.

They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Histograms are preferred to determine the underlying probability distribution of a data.

Box plots on the other hand are more useful when comparing between several data sets.

They are less detailed than histograms and take up less space.

Although histograms are better in displaying the distribution of data, we can use a box plot to tell if the distribution is symmetric or skewed.

11. How to select metrics?

Metrics are quantitative measurements.

Metric is consistent with the definition statistic, i.e. function of a sample.

Statistics captures the current status of the agent or group and performance metrics tell the story of how well the agent or group is performing

12. How do you assess the statistical significance of an insight?

Statistical importance of an insight can be accessed using Hypothesis Testing.

Hypothesis testing is guided by statistical analysis.

Statistical significance is calculated using a p-value, which tells the probability of result being observed,

given that a certain statement (the null hypothesis) is true.If the p-value is less than the significance level set (usually 0.05), the experimenter can assume that the null hypothesis is false and accept the

alternative hypothesis. Using a simple t-test, we can calculate a p-value and determine significance between two different groups of a dataset.

15. What is the Likelihood?

In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.

But in both frequents and Bayesian statistics, the likelihood function plays a fundamental role.

Unlike probability density functions, likelihoods aren't normalized. The area under their curves does not have to add up to 1. Likelihoods are a key part of Bayesian inference. We also use likelihoods to generate estimators; we almost always want the maximum likelihood estimator.