**SQL WORKSHEET-6**

1. Which of the following are TCL commands?

A. Commit B. Select C. Rollback D. Savepoint

**Answer**. A. Commit, C. Rollback and D. Savepoint

2. Which of the following are DDL commands?

A. Create B. Select C. Drop D. Alter

**Answer**. A. Create, C. Drop and D. Alter

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

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

A. SELECT NULL FROM SALES;

B. SELECT NAME FROM SALES;

C. SELECT \* FROM SALES WHEN PRICE = NULL;

D. SELECT # FROM SALES;

**Answer**. SELECT NAME FROM SALES;

4. DCL provides commands to perform actions like-

A. Change the structure of Tables

B. Insert, Update or Delete Records and Values

C. Authorizing Access and other control over Database

D. None of the above

**Answer**. C. Authorizing Access and other control over Database

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

A. Dates B. Column Alias C. String D. All of the mentioned

**Answer**. D. All of the above

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

A. ROLLBACK B. COMMIT C. TRUNCATE D. DELETE

**Answer**. B. COMMIT

7. A subquery in an SQL Select statement is enclosed in:

A. Parenthesis - (...).

B. brackets - [...].

C. CAPITAL LETTERS.

D. braces - {...}.

**Answer**. A. Parenthesis-(….)

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

A. FILE B. REPORT C. TABLE D. FORM

**Answer**. C. TABLE

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

A. Data types B. Primary keys C. Default values D. All of the mentioned

**Answer**. D. All of the above

10. If you don’t specify ASC and DESC after a SQL ORDER BY clause, the following is used by\_\_\_?

A. ASC B. DESC C. There is no default value D. None of the mentioned

**Answer**. A. ASC

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

11. What is denormalization?

**Answer**. Denormalization is a process of combining data from multiple tables into one so that it can be quired quickly. Denormalization is used to increase the performance of the database infrastructure. Denormalization is used where joins are expensive and frequent query is executed on the tables.

12. What is a database cursor?

**Answer**. A database cursor is a pointer that allows us to traverse over records in a database. The pointer can be moved. Cursors allow us to facilitate subsequent processing such as retrieval, addition or removal of database records.

13. What are the different types of the queries?

**Answer**. A query is an instruction passed to the DBMS to either retrieve some data or modify the data. There are 3 types of queries in SQL:

1. Basic Query

Some of the basic SQL queries are:

Create table

Select from table

Insert into table

Update table

Delete table

Alter table

Drop table

2. Complex query

When a user wants to use different complex syntax combined with basic sql queries, complex queries are used. Example- you wish to find the second highest salary of Employee.

3. Sub queries – Subqueries are nothing but nested queries.

Different subqueries are-

Simple subquery also called single row subquery in which we use aggregate functions.

Multi-row subquery are subqueries in which we use multi row operators like in, any.

Co-related subquery where we use different tables to get result.

14. Define constraint?

**Answer**. Constraints are set of rules enforced on the data columns in a table. Constraints restrict the type of data that can be entered into the table. If there is any violation between the constraint and the data action, the action is aborted. Some of the constraints of SQL are: not null, primary key, unique, foreign key etc.

15. What is auto increment?

**Answer**. Auto increment allows a unique number to be generated automatically for a new record inserted into the table. Generally, auto increment is used for the primary key constraint as it becomes easy to fulfil the requirement of primary key with auto increment.

**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?

a) The outcome from the roll of a die

b) The outcome of flip of a coin

c) The outcome of exam

d) All of the mentioned

**Answer**. D) All of the mentioned

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

a) Discrete

b) Non-Discrete

c) Continuous

d) All of the mentioned

**Answer**. a) Discrete

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

a) pdf b) pmv c) pmf d) all of the mentioned

**Answer**. a) pdf

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

a) mode b) median c) mean d) bayesian inference

**Answer**. b) median

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

a) variance b) standard deviation c) empirical mean d) all of the mentioned

**Answer**. c) Empirical mean

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

a) variance b) standard deviation c) mode d) none of the mentioned

**Answer.** a) Variance

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

a) 0 and 10 b) 1 and 2 c) 0 and 1 d) None of the mentioned

**Answer**. c) 0 and 1

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

a) baggyer b) bootstrap c) jacknife d) none of the mentioned

**Answer**. b) bootstrap

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

a) frequency b) summarized c) raw d) none of the mentioned

**Answer**. 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?

**Answer.** A histogram is a type of bar chart that graphically displays the frequencies of a dataset. A histogram plots the frequencies on y-axis and variable on the x-axis. Histograms are preferred when we want to determine the probability distribution of a data. Histograms are better in displaying the distribution of a data.

Boxplot- A boxplot also know as box-and-whisker plot is a chart that graphically represent five descriptive values for a dataset. The values are minimum value, the first quartile, the median, the third quartile and the maximum value. Box plots are used to compare several datasets. As they are less detailed. Boxplots too can be used to tell if the distribution is symmetric or skewed, In a symmetric distribution, mean and median are nearly same.

Both histogram and boxplots allow us to visualize the central tendency, amount of variation in the data, presence of outliers in the dataset.

11. How to select metrics?

**Answer**. Metrics are the evaluating models which tells us about the performance of our model and how accurate our model was. Choosing an optimum metrics depends on the kind of problem we are trying to solve. Common problems in Machine learning are classification problem in which data is predicted in defined numbers, regression problem in which the algorithm predict continuous values.

For classification problem we have

1. Confusion matrix also called error matrix is a table layout that allows visualization of the performance of an algorithm. It is a kind of contingency table with two dimensions ‘actual’ and ‘predicted’.

2. Accuracy metric- it indicates the number of correctly classified items compared to the total number of items, there are some limitations to accuracy metric that it doesnot well with an unbalanced data. So, we should choose the right metric for the right problem.

3. Recall- Recall metric shows how many true positive the model has classified from the total number of positive values.

4. Precision metric- This metric represents the number of True Positives which are really positive compared to the total number of positively predicted values.

5. F1-score- This metric is a combination of precision and recall metrics which serves as a comprise. The best F1 score equals 1, while the worst one is 0.

Regression performance metrics

1. Mean absolute error-This regression metric indicates the average sum of absolute difference between the actual and predicted value.

2. Mean squared error- Mean Squared Error (MSE) calculates the average sum of squared difference between the actual and predicted value for the entire data points.

MSE has some advantages over MAE as it highlights large errors over small ones.

3. Root Mean Squared error- RMSE is the square root of MSE.

So above mentioned are different metrics for classification and regression problems. The right metric always depends on the problem that we are solving.

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

**Answer**. Statistical significance can be assessed using hypothesis testing.

1. Start by choosing a null hypothesis which us actually opposite of what we wish to test and alternate hypothesis.

2. We choose a statistical test and statistics used to reject the null hypothesis.

3. We choose a critical region for the statistics to lie in that is extreme enough for null hypothesis to be rejected(p-value).

4. We calculate the observed test statistics from the data and check whether it lies in the critical region.

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

14. Give an example where the median is a better measure than the mean.

**Answer.** Median is generally considered to be the best representative of the centre of the data. More data is skewed we will have greater difference between the mean and the median, greater emphasis should be put on using the median over mean. Consider a situation where we are considering average salary of an Indian. Here as the salary data will be skewed mean would not give us the accurate average, here median would work better.

15. What is the Likelihood?

**Answer.** A likelihood function measures the goodness of fit of a statistical model to a sample of data for given values of unknown parameters.

**MACHINE LEARNING WORKSHEET 6**

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 overfitting?

A) High R-squared value for train-set and High R-squared value for test-set.

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

C) High R-squared value for train-set and Low R-squared value for test-set.

D) None of the above

**Answer**. C) High R-squared value for train-set and Low R-squared value for test-set.

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

A) Decision trees are prone to outliers.

B) Decision trees are highly prone to overfitting.

C) Decision trees are not easy to interpret

D) None of the above.

**Answer**. B) decision trees are highly prone to overfitting

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression C) Random Forest D) Decision tree

**Answer.** 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?

A) Accuracy B) Sensitivity C) Precision D) None of the above.

**Answer**. B) Sensitivity

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?

A) Model A B) Model B C) both are performing equal D) Data Insufficient

**Answer**. B) Model B

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

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

A) Ridge B) R-squared C) MSE D) Lasso

**Answer**. A) Ridge, D) Lasso

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

A) Adaboost B) Decision Tree C) Random Forest D) Xgboost.

**Answer**. A) Adaboost, D) Xg Boost

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

A) Pruning B) L2 regularization C) Restricting the max depth of the tree D) All of the above

**Answer**. D) All of the above

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

A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

C) It is example of bagging technique

D) None of the above

**Answer**. B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

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?

**Answer.** The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

11. Differentiate between Ridge and Lasso Regression.

**Answer**. Ridge regression is a regularized linear regression model that minimizes the impact of irrelevant features on the trained model. Ridge regression cannot be used for features reduction as it reduces the complexity of the model but does not reduce the number of variables since it only minimizes the coefficient and never reduces it to zero.

Lasso regression on the other hand sets coefficient of the irrelevant features to zero. We might end up with fewer features as it helps in feature reduction.

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

**Answer.** Variance Inflation Factor (VIF) is used to detect the strength of correlation between the independent variables of a multiple regression. It is predicted by taking a variable and regressing it against every other variable*.* VIF score of an independent variable represents how well the variable is explained by other independent variables. A VIF around 1 indicates no correlation between independent variable and other variables. So, when the two independent features are not correlated, they can be included in the regression modelling. VIF exceeding 5 or 10 indicates high multicollinearity between the independent variable and the others. So, we have to drop one of the two correlated features.

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

**Answer.** Feature scaling is a part of pre-processing which is applied to independent features of the dataset. Feature scaling helps to normalize the data within a particular range which makes it easy for a machine learning model to understand the variables with similar range. Independent variables may have different units (e.g., feet, kilometres, and hours) that, in turn, may mean the variables have different scales. Differences in the scales across input features may increase the difficulty of the problem being modelled. Feature scaling scales the input data which makes machine learning models to perform better. We can either use Normalization or standardization techniques to scale the data.

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

**Answer.** Goodness of fit of a linear regression model attempts to tell how well a model fits a given set of data and how well it will predict the future set of observations. A well fit regression model results in predicted values close to the observed data values. Different metrics to check the goodness of fit are:

1. R-squared or R2- R-square is the square of the correlation between the response values and the predicted response values. R-square can take on any value between 0 and 1, with a value closer to 1 indicating a better fit. For example, an R2 value of 0.8234 means that the fit explains 82.34% of the total variation in the data about the average.

2. MAE or mean absolute error which is the absolute difference between the actual values and predicted values. Closer the value of MAE to 0 better the model.

3. MSE- Mean squared error is calculated by taking the average of the square of the difference between the original and predicted values of the data.

4. RMSE- Root mean squared error is the square root of MSE.

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

|  |  |  |
| --- | --- | --- |
| Actual/predicted | Actual True | Actual False |
| Predicted True | 1000 TP | 50 FP |
| Predicted False | 250 FN | 120 TN |

**Answer.** Sensitivity=TP/(TP+FN) = 1000/ (1000+250) =1000/1250

Specificity=TN/(TN+FP) = 120/ (120+50) = 120/ 170

Precision = TP/ (TP + FP) = 1000/ (1000+50) = 1000/1050

Recall= TP/ (TP+FN) = 1000 / (1000 + 250) = 1000/1250

Accuracy= TP + TN / (TP + TN + FP + FN) = 1000+120 / (1000+120+250+50) = 1120/1420