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?

Ans: - d) All of the mentioned

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

Ans: - a) Discrete

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

Ans: - a) pdf

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

Ans: - c) mean

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

Ans: - c) empirical mean

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

Ans: - a) variance

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

Ans: - c) 0 and 1

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

Ans: - b) bootstrap

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

Ans: - 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?

Ans: - 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.

Both histograms and box plots are used to explore and present the data in an easy and understandable manner. 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.

11. How to select metrics?

Ans: - Step 1 Why is the measurement required?

Step 2 What needs to be measured?

Step 3 What is the precision of measurement required?

Step 4 How will it be measured?

Step 5 What use will the measurement be put to? By whom?

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

Ans: - Statistical significance can be accessed using hypothesis testing: – Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B) – Then, we choose a suitable statistical test and statistics used to reject the null hypothesis – Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value) – We calculate the observed test statistics from the data and check whether it lies in the critical region Common tests:

– One sample Z test – Two-sample Z test – One sample t-test – paired t-test – Two sample pooled equal variances t-test – Two sample unpooled unequal variances t-test and unequal sample sizes (Welch’s t-test) – Chi-squared test for variances – Chi-squared test for goodness of fit – Anova (for instance: are the two regression models equals? F-test) – Regression F-test (i.e: is at least one of the predictors useful in predicting the response?)

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Ans: - life table is example of exponential distribution, wind speed is Weibull distribution, surgery patient's stay in hospital is gamma distribution

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

Ans: - The median is usually preferred to other measures of central tendency when your data set is skewed (i.e., forms a skewed distribution) or you are dealing with ordinal data.

For example, suppose you want to calculate the amount of money an average citizen in a town earns here the data may be baised as some people might be earning way too much or way too less so here median would be a good measure to calculate central tendancy than mean.

15.What is the Likelihood?

Ans: - 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. It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values.

WORKSHEET 6 SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1.Which of the following are TCL commands?

Ans: - A. Commit

C. Rollback

D. Savepoint

2.Which of the following are DDL commands?

Ans: - A. Create

C. Drop

D. Alter

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

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

Ans: - B. SELECT NAME FROM SALES;

2.DCL provides commands to perform actions like

Ans: - C. Authorizing Access and other control over Database

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

Ans: - B. Column Alias

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

Ans: - B. COMMIT

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

Ans: - A. Parenthesis - (...).

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

Ans: - C. TABLE

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

Ans: - D. All of the mentioned

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

Ans: - A. ASC

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

1.What is denormalization?

Ans: - Denormalization is a strategy that database managers use to increase the performance of a database infrastructure. It involves adding redundant data to a normalized database to reduce certain types of problems with database queries that combine data from various tables into a single table. The definition of denormalization is dependent on the definition of normalization, which is defined as the process of organizing a database into tables correctly to promote a given use.

2.What is a database cursor?

Ans: - A database cursor can be thought of as a pointer to a specific row within a query result. The pointer can be moved from one row to the next. Depending on the type of cursor, you may be even able to move it to the previous row.

Think of it this way: a SQL result is like a bag, you get to hold a whole bunch of rows at once, but not any of them individually; whereas, a cursor is like a pair of tweezers. With it, you can reach into the bag and grab a row, and then move onto the next.

3.What are the different types of the queries?

Ans: - Here are five types of widely used SQL queries.

Data Definition Language (DDL)

Data Manipulation Language (DML)

Data Control Language(DCL)

Transaction Control Language(TCL)

Data Query Language (DQL)

4.Define constraint?

Ans: - Constraints are the rules enforced on the data columns of a table. These 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 database.

Constraints could be either on a column level or a table level. The column level constraints are applied only to one column, whereas the table level constraints are applied to the whole table.

5.What is auto increment?

Ans: - 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.

MACHINE LEARNING 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?

Ans: - 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?

Ans: - B) Decision trees are highly prone to overfitting.

3.Which of the following is an ensemble technique?

Ans: - 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?

Ans: - 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?

Ans: - B) Model B

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

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

Ans: - A) Ridge

D) Lasso

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

Ans: - B) Decision Tree

C) Random Forest

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

Ans: - A) Pruning

C) Restricting the max depth of the tree

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

Ans: - 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

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

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

Ans: - 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. The adjusted R-squared can be negative, but it’s usually not. It is always lower than the R-squared.

2.Differentiate between Ridge and Lasso Regression.

Ans: - Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso go with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it’s the sum of the absolute values of the coefficients. It’s a shrinkage towards zero using an absolute value (l1 penalty) rather than a sum of squares(l2 penalty).

As we know that ridge regression can’t zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficients of collinear variables. Here it helps to select the variables out of given n variables while performing lasso regression.

3.What is VIF? What is the suitable value of a VIF for a feature to be included in a regressionmodelling?

Ans: - Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above. Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression, like x and x2.

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

Ans: - To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

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

Ans: - These (R Squared, Adjusted R Squared, F Statistics , RMSE / MSE / MAE) are some metrics which are used to check the goodness of fit in linear regression.

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

Ans: - Sensitivity = TPR = TP / (TP + FN) = 0.8000

Specificity = SPC = TN / (FP + TN) = 0.9600

Precision = PPV = TP / (TP + FP) = 0.9524

Recall = TPR = TP / (TP + FN) = 0.8000

Accuracy = ACC = (TP + TN) / (P + N) = 0.8800