STATISTICS WORKSHEET

1. Which of the following can be considered as random variable?
    - The outcome of flip of a coin.
2. Which of the following random variable that take on only a countable number of possibilities?
    - Discrete.
3. Which of the following function is associated with a continuous random variable?
    - Pdf.
4. The expected value or _______ of a random variable is the center of its distribution.
    - Mean.
5. Which of the following of a random variable is not a measure of spread?
    - Variance.
6. The _________ of the Chi-squared distribution is twice the degrees of freedom.
    - Variance.
7. The beta distribution is the default prior for parameters between ____________.
    - 0 and 1
8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
    - Bootstrap.

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

    - Summarized.

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

    - Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The 'whiskers' of a box plot show the least and greatest values in the data set.

11. How to select metrics?

    - We use Classification Models to predict class labels for a given input data. It is important that this choice is backed by analytical reasoning. Often, we choose *Model Accuracy* to evaluate the model. It’s a popular choice because it is very easy to understand and explain. Accuracy coincides well with the general aim of building a classification model, to predict the class of new observations accurately .

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

    - To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you’d calculate the p-value, which is the likelihood of getting the test’s observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

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

- Pick random values in the 0,1 interval - this can be done by spinning a pointer or many other ways.

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

- Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed.

15. What is the Likelihood?

- The likelihood is the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on ; it is not a probability density over the parameter . The likelihood, , should not be confused with , which is the posterior probability of given the data .

SQL WORKSHEET

1. Which of the following are TCL commands?
    - Rollback.
2. Which of the following are DDL commands?
    - Alter.
3. Which of the following is a legal expression in SQL?
    - SELECT NAME FROM SALES;
4. DCL provides commands to perform actions like-
    - Authorizing Access and other control over Database.
5. Which of the following should be enclosed in double quotes?
    - Column Alias.
6. Which of the following command makes the updates performed by the transaction permanent in the database?
    - COMMIT.
7. A subquery in an SQL Select statement is enclosed in:
    - Parenthesis - (...).
8. The result of a SQL SELECT statement is a :-
    - TABLE.
9. Which of the following do you need to consider when you make a table in a SQL?
    - All of the mentioned.
10. If you don’t specify ASC and DESC after a SQL ORDER BY clause, the following is used by___?
    - ASC.

11. What is denormalization?

- Denormalization is a technique used by database administrators to optimize the efficiency of their database infrastructure. This method allows us to add redundant data into a normalized database to alleviate issues with database queries that merge data from several tables into a single table. The denormalization concept is based on the definition of normalization that is defined as arranging a database into tables correctly for a particular purpose.

12. What is a database cursor?

- 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.

13. What are the different types of the queries?

- 3 Different types of the Queries-

  Navigational search queries.
  Informational search queries.
  Transactional search queries.

14. Define constraint?

- Constraints in SQL means we are applying certain conditions or restrictions on the database. This further means that before inserting data into the database, we are checking for some conditions. If the condition we have applied to the database holds true for the data which is to be inserted, then only the data will be inserted into the database tables.

15. What is auto increment?

- The auto increment in SQL is a feature that is applied to a field so that it can automatically generate and provide a unique value to every record that you enter into an SQL table. This field is often used as the Primary key column, where you need to provide a unique value for every record you add. However, it can also be used for the UNIQUE constraint columns.

MACHINE LEARNING

1) In which of the following you can say that the model is overfitting?
    - Low R-squared value for train-set and High R-squared value for test-set.
2) Which among the following is a disadvantage of decision trees?
    - Decision trees are prone to outliers.
3) Which of the following is an ensemble technique?
    - SVM.
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?
    - 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?
    - Model A.
6) Which of the following are the regularization technique in Linear Regression?
    - Lasso.
7) Which of the following is not an example of boosting technique?
    - Decision Tree.
8) Which of the techniques are used for regularization of Decision Trees?
    - Pruning.

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

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

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

- The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11) Differentiate between Ridge and Lasso Regression.

- Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square. Ridge regression is also referred to as L2 Regularization.

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

- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.

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

- 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.

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

- Three statistics are used in Ordinary Least Squares regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error.

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

-