STATISTICS WORKSHEET- 6

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 a) variance

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 for difficult statistics?

ANS b) bootstrap

9. Data that summarize all observations in a category are called _____ data.
ANS b) summarized

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

ANS 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?

ANS For selecting metrics

Step 1: Articulate Your Goals. This is obvious, but you should always start by defining your goals for your product. - Before you even begin to sift through the various metrics and statistics available to you, it is essential that your company's governing objectives have been clearly established.

Step 2: List the Actions That Matter- Once a clear, overarching objective has been established, most marketing companies look to major metrics to determine their success—factors such as the generation of sales and leads.

Step 3: Define Your Metrics- By taking advantage of the analytical tools provided by various online platforms, a marketing agency could easily find that for one client, video content is the chief driver of engagement, while for another, lengthy, informative blogs are the most effective type of content.

Step 4: Evaluate your Metrics-. The metrics and statistics that drive value for your clients can change over time, especially as new technologies emerge and target demographics shift. Regularly evaluating your methods and adapting, when necessary, may cause you to throw away some of your work. But this is by no means a waste.

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

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

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

ANS Many random variables have distributions that are *asymptotically* Gaussian but may be significantly non-Gaussian for small numbers. For example, the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large

For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My images are about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps "from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

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

ANS It is best to use the median when the distribution is either skewed or there are outliers present.

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?

ANS The likelihood function is central to the process of estimating the unknown parameters. Older and less sophisticated methods include the method of moments,

and the method of minimum chi-square for count data. These estimators are not always efficient, and their sampling distributions are often mathematically intractable.

Likelihood Let X1, X2, ..., Xn have a joint density function f(X1, X2, ..., Xn|θ).

Given X1 = x1, X2 = x2, ..., Xn = xn is observed,

the function of θ defined by:

$L(\theta) = L(\theta|x1, x2, ..., xn) = f(x1, x2, ..., xn|\theta)$ (1)

is the likelihood function.

• The likelihood function is not a probability density function.

• It is an important component of both frequentist and Bayesian analyses

• It measures the support provided by the data for each possible value of the parameter.

SQL ASSIGNMENT

1. Which of the following are TCL commands?

ANS A. Commit C. Rollback D. Save point

2. Which of the following are DDL commands?

ANS A. Create C. Drop D. Alter

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

ANS; B. SELECT NAME FROM SALES; C. SELECT * FROM SALES WHEN PRICE = NULL;

4. DCL provides commands to perform actions like

ANS C. Authorizing Access and other control over Database

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

ANS B. Column Alias

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

ANS B. COMMIT

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

ANS A. Parenthesis - (...).

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

ANS C. TABLE

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

ANS B. Primary keys

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

ANS A. ASC

11. What is denormalization?

ANS Denormalization is the process of adding precomputed redundant data to an otherwise normalized relational database to improve read performance of the database. Normalizing a database involves removing redundancy so only a single copy exists of each piece of information.

12. What is a database cursor?

ANS 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.

You must use a cursor in the following cases:

- Statements that return more than one row of data from the database server:
  - A SELECT statement requires a select cursor.
  - An EXECUTE FUNCTION statement requires a function cursor.
- An INSERT statement that sends more than one row of data to the database server requires an insert cursor.

13. What are the different types of the queries?

ANS A query is a question, regularly communicated formally. A database query can be either a select question or an action query.

Types of SQL Queries

- Select Query- The select query is the least difficult kind of inquiry. It very well may be utilized to choose and show information from possibly one table or a progression of them relying upon what is required.

- Action Query-At the point when the activity question is called, the database experiences a particular activity relying upon what was indicated in the query itself. This can incorporate such things as making new tables, erasing lines from existing ones and refreshing records or making totally new ones. Action queries are extremely famous in information the board since they take into account numerous records to be changed at one time rather than just single records like in a select query.

  Four types of action queries are:

1. **Append Query** – takes the set consequences of a query and "adds" (or includes) them to a current table.
2. **Delete Query** – erases all records in a hidden table from the set results of a query.
3. **Make Table Query** – as the name proposes, it makes a table dependent on the set consequences of a query
4. **Update Query** – takes into account at least one field in your table to be refreshed.

- Aggregate Query- A unique kind of query is known as an aggregate query. It can chip away at different queries, (for example, choice, activity or parameter) simply like the parameter query does, yet as opposed to passing a parameter to another query it aggregates up to the things by chosen by the various groups.

14. Define constraint?

ANS 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. If there is any violation between the constraint and the data action, the action is aborted.

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

15. 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 ASSIGNMENT

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

ANS A) High 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?

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 A) Accuracy

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 A) Model A

6. Which of the following are the regularization technique in Linear Regression??
ANS A) Ridge D) Lasso

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

ANS B) Decision Tree C) Random Forest

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

ANS A) Pruning

9. 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

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

ANS 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.

ANS Lasso Regression - This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method**.** Lasso is short for Least Absolute Shrinkage and Selection Operator, which is used both for regularization and model selection. If a model uses the L1 regularization technique, then it is called lasso regression.

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m}|\hat{\beta}_j|.$$

Ridge 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.

$$L_{hridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m} w_j \hat{\beta}_j^2.$$

Ridge regression is also referred to as **L2 Regularization**.

Conclusion

We have seen an implementation of ridge and lasso regression models and the theoretical and mathematical concepts behind these techniques. Some of the key takeaways from this tutorial include:

1. The cost function for both ridge and lasso regression are similar. However, ridge regression takes the square of the coefficients and lasso takes the magnitude.
2. Lasso regression can be used for automatic feature selection, as the geometry of its constrained region allows coefficient values to inert to zero.
3. An alpha value of zero in either ridge or lasso model will have results similar to the regression model.
4. The larger the alpha value, the more aggressive the penalization.

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

ANS A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity. As a rule of thumb, a VIF of three or below is not a cause for concern.

13. 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.

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

ANS **R-squared**- The difference between SST and SSE is the improvement in prediction from the regression model, compared to the mean model. Dividing that difference by SST gives R-squared. It is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the model.

R-squared has the useful property that its scale is intuitive. It ranges from zero to one. Zero indicates that the proposed model does not improve prediction over the mean model. One indicates perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

Adjusted R-squared-Adjusted R-squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile.

Adjusted R-squared should always be used with models with more than one predictor variable. It is interpreted as the proportion of total variance that is explained by the model.

The F-test-The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not. An equivalent null hypothesis is that R-squared equals zero.

A significant F-test indicates that the observed R-squared is reliable and is not a spurious result of oddities in the data set. Thus the F-test determines whether the proposed relationship between the response variable and the set of predictors is statistically reliable. It can be useful when the research objective is either prediction or explanation.

RMSE-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. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance. It has the useful property of being in the same units as the response variable.

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