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-b) standard deviation

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

Ans-b) standard deviation

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 indicate the whole frequency distribution of a variable, whereas the boxplot summarises its most prominent features**. These features include median and spread as well as the extent and nature of departures from symmetry, and the possible presence of observations having extreme values (outliers).

11. How to select metrics?

1. Ans-Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.
2. Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.
3. Ranking. The model will predict an order of items.

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

Ans-1) State the Research Hypothesis
2) State the Null Hypothesis
3) Select a probability of error level (alpha level)
4) Select and compute the test for statistical significance
5) Interpret the results

1) State the Research Hypothesis

A research hypothesis states the expected relationship between two variables. It may be stated in general terms, or it may include dimensions of direction and magnitude.

2) State the Null Hypothesis

A null hypothesis usually states that there is no relationship between the two variables.

3) TYPE I AND TYPE II ERRORS

Even in the best research project, there is always a possibility (hopefully a small one) that the researcher will make a mistake regarding the relationship between the two variables. There are two possible mistakes or errors.

4) The Chi Square Test

For nominal and ordinal data, Chi Square is used as a test for statistical significance. For example, we hypothesize that there is a relationship between the type of training program attended and the job placement success of trainees.

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

Ans-Exponential distributions do not have a log-normal distribution or a Gaussian distribution.
In fact, any type of data that is categorical will not have these distributions as well.
Example: Duration of a phone car, time until the next earthquake, etc.

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

Ans-

If the median is greater than the mean on a set of test scores,
describe the situation.

Shawna,

The official answer is that the data are "skewed to the left", with a long tail of low scores pulling the mean down more than the median. There is one definition of skewness (Pearson's) by which this is the case by definition.

15. What is the Likelihood?

Ans-Likelihood function is a fundamental concept in statistical inference. It **indicates how likely a particular population is to produce an observed sample**. Let P(X; T) be the distribution of a random vector X, where T is the vector of parameters of the distribution.

MACHINE LEARNING

ASSIGNMENT - 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- 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?

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-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?

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?

A) Pruning

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

D) None of the above

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

Ans-Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model. Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

11. Differentiate between Ridge and Lasso Regression.

Ans-Yes…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 variable(s) out of given n variables while performing lasso regression.

Another type of regularization method is ElasticNet, it is hybrid of lasso and ridge regression both. It is trained with L1 and L2 prior as regularizer. A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation.

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

Ans-12. 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. As a rule of thumb, a VIF of three or below is not a cause for concern. As VIF increases, the less reliable your regression results are going to be.

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

Ans- The reason to scale data before Regularizing is that **regularizing penalizes the model for large coefficients**. The magnitude of coefficients depends on 2 major factors: Strength of relation between input(x) and output variable(y).

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

Ans- There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are:

- Mean Squared Error (MSE).
- Root Mean Squared Error (RMSE).
- Mean Absolute Error (MAE)

Mean Squared Error

Mean Squared Error, or MSE for short, is a popular error metric for regression problems. It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here "*least squares*" refers to minimizing the mean squared error between predictions and expected values.

Root Mean Squared Error

The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.

Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.

For example, if your target variable has the units “*dollars*,” then the RMSE error score will also have the unit “*dollars*” and not “*squared dollars*” like the MSE.

Mean Absolute Error

Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.
Unlike the RMSE, the changes in MAE are linear and therefore intuitive.

That is, MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error.

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

Ans-Accuracy-90050,

Sensitivity-100050

Precision- -99050