STATISTICS WORKSHEET- 6

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. a) variance
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. b) summarized
10. 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 allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.

    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. There are 4 steps in selecting the metrics: Articulate the goals, list the actions that matters, defining the metrics, evaluating the metrics.

12. Steps in Testing for Statistical Significance

    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.

13. The examples of data that does not have a Gaussian distribution, nor log-normal are the amount of time that a car battery lasts or the amount of time until an earthquake occurs.
14. Income is an example where the median is a better measure than the mean.
15. 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.

MACHINE LEARNING - 6

1. A) High R-squared value for train-set and High R-squared value for test-set.
2. B) Decision trees are highly prone to overfitting.
3. C) Random Forest
4. A) Accuracy
5. B) Model B
6. A) Ridge   A) Ridge
7. B) Decision Tree   D) Xgboost
8. A) Pruning
9. D) None of the above
10. 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. Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros.
12. The Variance Inflation Factor measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. Less than 10 is suitable value of a VIF for a feature to be included in a regression modelling.
13. 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. There are three metrics that are commonly used for evaluating the performance of a regression model, they are: Mean Squared Error (MSE), Root Mean Squared Error (RMSE),Mean Absolute Error (MAE).
15. Recall/Sensitivity = TP/TP+FN = 1000/ (1000+250) = 0.80
    Specificity = TN/(TN+FP) = 1200/ (1200+50) = 0.96
    Precision = TP/(TP+FP) = 1000/ (1000+50) = 0.95
    Accuracy = (TP+TN)/(TP+TN+FP+FN) = (1000+1200)/(1000+1200+50+250) = 0.88