STATISTICS WORKSHEET- 6

Q1. D

Q2. A

Q3. A

Q4. C

Q5. C

Q6. B

Q7. C

Q8. B

Q9. B

Q10. Histogram divides the numeric data into uniform intervals and displays the number of data values falling within each bin. It shows the number of values within an interval but not the actual values. Box plot is a good way to summarize large amounts of dataIt is easier to read minimum value, median, outliers, quantiles, and maximum value.

Q11. As per require, we need to understand what kind of problem we are trying to solve.Here is a list of some problems in ML:

- Classification algorithm wil predict data type from defined data array. Say like it may respond with yes/no/not sure
- Regression algorithm will predict some values, for ex. Tomarrow's weather forecasting
- Categorical model will predict an order of items.For Ex. A group of student and ranking them depending on their height from the tallest to the shortest.

Q12. Hypothesis testing is guided by statistical analysis. Statistical significance is calculated using a p-value, which tells you the probability of your result being observed, given that a certain statement (the null hypothesis) is true. If this p-value is less than the significance level set (usually 0.05), the experimenter can assume that the null hypothesis is false and accept the alternative hypothesis. Using a simple t-test, you can calculate a p-value and determine significance between two different groups of a dataset.

Q13. Any type of categorical data won't have a Gaussian Distribution or lognormal distribution. Exponential distributions - Ex. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

Q14. Let us say that there are nine students in a class with the following scores on a test: 2, 4, 5, 7, 8, 10, 12, 13, 83. In this case the average score (or the **mean**) is the sum of all the scores divided by nine. This works out to 144/9 = 16. Note that even though 16 is the arithmetic average, it is distorted by the unusually high score of 83 compared to other scores. Almost all of the students' scores are below the average. Therefore, in this case the mean is not a good representative of the central tendency of this sample.

The median, on the other hand, is the value which is such that half the scores are above it and half the scores below. So in this example, the median is 8. There are four scores below and four above the value 8. So 8 represents the mid point or the central tendency of the sample.

Q15. The term Likelihood refers to the process of determining the best data distribution given a specific situation in the data. For Ex. Suppose you have an unbiased coin. If you flip the coin, the probability of getting head and a tail is equal, which is 0.5. When calculating the probability of coin getting heads, you assume that P(head) = 0.5. However, when calculating the likelihood, you are trying to find if the model parameter (p = 0.5) is correctly specified or not.

MACHINE LEARNING

Q1. C

Q2. B

Q3. C

Q4. B

Q5. B

Q6. A & D

Q7. B & C

Q8. A & C

Q9. A & C

Q10.

As we see R2 always increases with an increase in the no. of independent variables. Thus, it doesn't give a better picture and so we need Adjusted R2 value. Asjusted R2 is equal t oR2. Thus adjusted R2 will always be less than or equal to R2 and it penalize the excess of independent variables which do not affect the dependent variable.

Q11.

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.

Q12.

Variance inflation factor (VIF) is **a measure of the amount of multicollinearity in a set of multiple regression variables**. The suitable value of feature is 5.

Q13.

Scaling of the data **makes it easy for a model to learn and understand the problem**. Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set.

Q14.

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)**

Q15.

1)Sensitivity or Recall = TP/(TP+FN)

=1000/(1000+250) =1000/1250 =0.8

2)Specificity = TN/(TN+FP)

=1200/(1200+50) =1200/1250 =0.96

3)Precision = TP/(TP+FP)

=1000/(1000+50) =1000/1050 =0.95

4)Accuracy = (TP+TN)/(TP+FP+FN+TN)

=(1000+1200)/(1000+50+250+1200) =2200/2500 =0.88