ASSIGNMENT – 3
MACHINE LEARNING
SET-1

In Q1 to Q11, only one option is correct, choose the correct option:

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Error **(ANSWER)**

B) Maximum Likelihood

C) Logarithmic Loss

D) Both A and B

2. Which of the following statement is true about outliers in linear regression?

A) Linear regression is sensitive to outliers **(ANSWER)**

B) linear regression is not sensitive to outliers

C) Can't say

D) none of these

3. A line falls from left to right if a slope is _____?

A) Positive

B) Negative **(ANSWER)**

C) Zero

D) Undefined

4. Which of the following will have symmetric relation between dependent variable and independent variable?

A) Regression

B) Correlation **(ANSWER)**

C) Both of them

D) None of these

5. Which of the following is the reason for over fitting condition?

A) High bias and high variance

B) Low bias and low variance

C) Low bias and high variance **(ANSWER)**

D) none of these

6. If output involves label then that model is called as:

A) Descriptive model

B) Predictive modal **(ANSWER)**

C) Reinforcement learning

D) All of the above

7. Lasso and Ridge regression techniques belong to _____?

A) Cross validation

B) Removing outliers

C) SMOTE

D) Regularization **(ANSWER)**

8. To overcome with imbalance dataset which technique can be used?

A) Cross validation

B) Regularization

C) Kernel

D) SMOTE **(ANSWER)**

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

A) TPR and FPR **(ANSWER)**

B) Sensitivity and precision

C) Sensitivity and Specificity

D) Recall and precision

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

A) True

B) False **(ANSWER)**

11. Pick the feature extraction from below:

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data **(ANSWER)**

C) Removing stop words

D) Forward selection

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

A) We don't have to choose the learning rate.

B) It becomes slow when number of features is very large.

C) We need to iterate.

D) It does not make use of dependent variable. **(ANSWER)**

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization.

ANSWER:- In Machine Learning Regularization is a set of techniques used to ensure that a machine learning model can generalize to new data within the same data set. Techniques like regularization lead to a lack of noise (errors) which falls outside the expected range of pattern. Its helps to model reduce the complexity in it and makes it easier to detect relevant edge cases within a classification task. Regularization adeptly steering clear of both underfitting and overfitting.

Overfitting as described above, occurs when a model is too complex and learns errors or noise in the training data. On the Other hand, Underfitting occurs when a model is too simple to capture underlying data patterns. The term Regularization provides means to finds the optimal balance between these two extremes.

14. Which particular algorithms are used for regularization?

ANSWER:- Regularization techniques balance fitting training data with keeping the model simple, which can lead to better performance on new data. So, there are some algorithms used

for Regularization for its better performance are L1 regularization (LASSO) and L2 regularization (RIDGE)

1.) LASSO – In LASSO Regularization it modifies overfitted or underfitted models by adding a penalty equivalent to the sum of the absolute values of the coefficients. LASSO regression also performs coefficients minimization, but instead of squaring the magnitudes of the coefficient, it takes the actual values of the coefficients. This means that the sum of the coefficients can also be 0 because there are negative coefficients. Consider the cost function for the lasso regression.

2.) RIDGE – RIDGE Regression, adjusts models with overfitting or underfitting by adding a penalty equivalent to the sum of the squares of the magnitudes of the coefficients. This implies that we minimize the mathematical function representing our machine-learning model and calculate the coefficients. We multiply and add the size of the coefficients. Ridge Regression performs regularization machine learning by reducing the coefficients present.

15. Explain the term error present in linear regression equation.

ANSWER:- An Error term present in the Linear Regression Equation is **$y= b_0 + b_x + E.$** So, the E in the equation denotes the Error term which tells the lack of perfect goodness of fit. An Error or Residuals variable that appears in a statistical model when it doesn't fully represent the actual relationship between the independent (features) and the Dependent(labels) variables. An error term essentially means that the model is not completely accurate and results in differing results during real-world applications.

The error, it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results.

STATISTICS– WORKSHEET

SET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

a) True **(ANSWER)**

b) False

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem **(ANSWER)**

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data **(ANSWER)**

c) Modeling contingency tables

d) All of the mentioned

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned **(ANSWER)**

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson **(ANSWER)**

d) All of the mentioned

6. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False **(ANSWER)**

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis **(ANSWER)**

c) Causal

d) None of the mentioned

8. Normalized data are centered at_____ and have units equal to standard deviations of the original data.

a) 0 **(ANSWER)**

b) 5

c) 1

d) 10

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship **(ANSWER)**

d) None of the mentioned

WORKSHEET

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

ANSWER- The term Normal Distribution or Gaussian Distribution is the probability distribution that is symmetric about mean, showing that data near the mean are most frequent in occurrence than data far from the mean. In Normal distribution, mean(average), median (mid-point), and mode (most frequent observation) are equal. These values represent the peak or highest point. The normal distribution is the proper term for a probability bell curve. The mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal Distribution are symmetrical, but not all symmetrical distributions are normal.

11. How do you handle missing data? What imputation techniques do you recommend?

ANSWER- Data imputation involves filling in missing values with estimates, allowing that data to be analyzed and handled by using standard techniques. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.so the most prevalent methods is recommend:-

- Mean imputation- Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

- Substitution- Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.
- Regression imputation- The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

12. What is A/B testing?

ANSWER- A/B testing is also known as Split Testing or Bucket Testing. It's a methodology for comparing two versions of a webpage or app against each other to determine which one performs better.

13. Is mean imputation of missing data acceptable practice?

ANSWER- Mean imputation is common method for replacing missing data with the mean, median ,or mode of a column. Mean imputation can be acceptable in some cases, such as when the data is numeric and not skewed, and when there are a small number of missing observations:

- Numeric and not skewed: Mean imputation is preferred if the data is numeric and not skewed.
- Small number of missing observations: When there is a small number of missing observations, data scientists can calculate the mean of the existing observations and insert them in place of the missing observations.

14. What is linear regression in statistics?

ANSWER- Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical practice of calculating a straight line that specifies a mathematical relationship between two variables. It is a statistical method used in data science and machine learning for predictive analysis.

15. What are the various branches of statistics?

ANSWER- The two main branches of statistics are descriptive statistics and inferential statistics.

1. DESCRIPTIVE STATISTICS- In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data. Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars or scatter plots.

2. INFERENTIAL STATISTICS- In Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarised we use Inferential Statistics to describe the Meaning of the collected Data. Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions. Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.