MACHINE LEARNING

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

Ans :- A) Least Square Error

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

Ans :- A) Linear regression is sensitive to outliers

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

Ans :- B) Negative

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

Ans :- B) Correlation

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

Ans :- C) Low bias and high variance

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

Ans :- B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

Ans :- D) Regularization

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

Ans :- A) Cross validation

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

Ans :- A) TPR and FPR

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

Ans :- B) False

11. Pick the feature extraction from below:

Ans :- A) Construction bag of words from a email
B) Apply PCA to project high dimensional data
C) Removing stop words

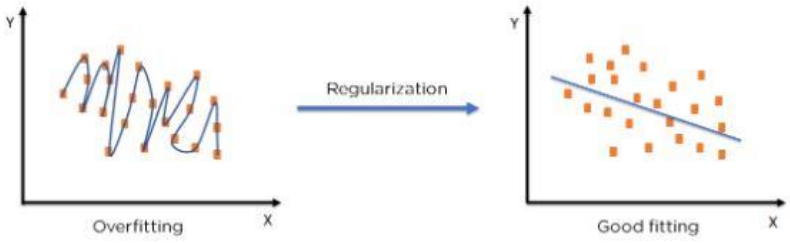
12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Ans :- 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.

13. Explain the term regularization?

Ans :- In the context of machine learning, regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting.

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.



Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

14. Which particular algorithms are used for regularization?

Ans :- There are three main regularization techniques, namely:

1) Ridge Regression (L2 Regularization) :-
A regression model that uses L2 regularization technique is called Ridge regression.

2) Lasso Regression (L1 Regularization) :-
A regression model which uses L1 Regularization technique is called LASSO (Least Absolute Shrinkage and Selection Operator) regression.

3) Dropout :-
Dropout is a regularization technique used in neural networks. It prevents complex co-adaptations from other neurons.

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

Ans :- An error term represents the margin of error within a statistical model; 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.

An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example,

Linear Regression: $Y = a + bX + e$

Where:
Y = the variable that you are trying to predict (dependent variable)
X = the variable that you are trying to predict Y (independent variable)
a = the intercept
b = the slope
e = the regression residual error

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
Ans :- a) True

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?
Ans :- a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
Ans :- b) Modeling bounded count data

4. Point out the correct statement.
Ans :- d) All of the mentioned

5. _____ random variables are used to model rates.
Ans :- c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
Ans :- b) False

7. Which of the following testing is concerned with making decisions using data?
Ans :- b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
Ans :- a) 0

9. Which of the following statement is incorrect with respect to outliers?
Ans :- c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Ans :- The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.
Normal distribution is used in investing to represent asset class returns and their distribution patterns. Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans :- Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. 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.

The following are some of the most prevalent imputation techniques:
1. Mean imputation
2. Substitution
3. Hot deck imputation

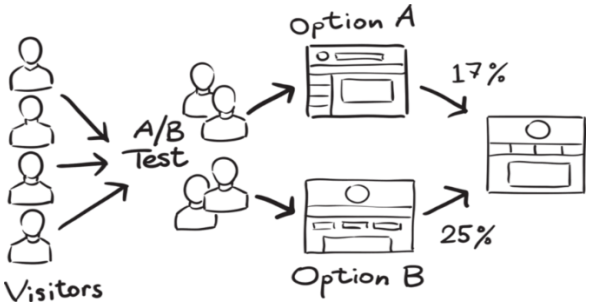
4. Cold deck imputation
5. Regression imputation
6. Stochastic regression imputation
7. Interpolation and extrapolation
8. Single or Multiple Imputation

12. What is A/B testing?

Ans :- A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



13. Is mean imputation of missing data acceptable practice?

Ans :- The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans :- Linear regression analysis is used to predict the value of a variable based on the value of another variable.

The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data.

15. What are the various branches of statistics?

Ans :- **Statistics:** Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are **two main branches** of statistics
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:
Inferential statistics used to make inference and describe about the population. These stats are more useful when it’s not easy or possible to examine each member of the population.

Descriptive Statistics:
Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.