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. D) SMOTE

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

Ans. C) Sensitivity and Specificity

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. B) Apply PCA to project high dimensional data

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.

13. **Explain the term regularization?**

Ans. 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.

In general, regularization means to make things regular or acceptable. 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.

14. **Which particular algorithms are used for regularization?**

Ans. There are two main algorithms which is used for regularization :

Ridge Regression(L2) -- Ridge Regression is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.

- It shrinks the coefficients towards 0. therefore, it is mostly used to prevent multicollinearity.
- It reduces the model complexity by coefficient shrinkage.
- It uses L2 regularization technique.

Lasso Regression(L1) -- Lasso regression is another regularization technique to reduce the complexity of the model. It reduce the coefficient to zero and its quite similar to ridge.

lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

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

Ans. Linear regression will try to make the best fit line, where the eq**. (y= a + b x + e)** could be settled**.**

Where,

x= independent variable

Y= dependent variable

b=coefficient value

a=intercept

e= Error

when we work on the machine learning, we have to give the data to the machine for any type of prediction, and we give the data to the machine in the form of x train, x test, y train and y test.

Here y test is the output given by the machine.

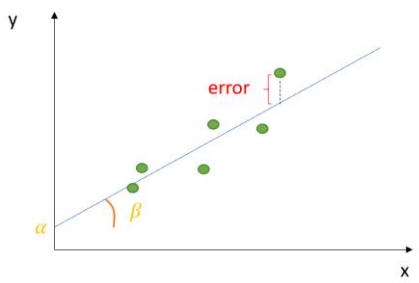
In this case, we have two types of data:

1st one is the data given by machine called machine driven result(predicted ans)

2nd one is the Actual data

The difference between actual data and machine driven data is called an error which is denoted by e.

If the error is less, then we can say our model is working well.



STATISTICS

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. a) Modeling event/time data.

4. Point out the correct statement

Ans. d) All of the mentioned

5. _____ random variables are used to model rates.

Ans. d) All of the mentioned

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 cantered 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. Outliers cannot conform to the regression relationship

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

Ans. Normal Distribution is also called as Gaussian distribution or bell curve. In this distribution data are normally distributed.

1) In normal distribution mean, median and mode all are equal.
2) The curve is symmetric at the center.
3) Half of the data is present on the right and other half is present on the left side from the center.
4) The total area under the curve is 1. (std= plus minus 1)

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

Nowadays, Missing data is quite common in a well-designed and controlled study, and that can lead to having a major impact on the conclusions that can be drawn from the data.

When dealing with missing data, we can use two primary methods to solve the error:

1)Imputation

2) Removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

There are few Imputation techniques which can be recommend for handling missing data:

1 The time series methods

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid.

2 Multiple imputation

Multiple imputation is considered a good approach for data sets with a large amount of missing data.

3 k neighbors.

In this method, data scientists choose a distance measure for k neighbors, and the average is used to impute an estimate.

4. Mean, Median and Mode

This is one of the most common methods of imputing values when dealing with missing data.

12. What is A/B testing?

Ans. A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment.

A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product. Large social media sites like linked in, Fb and insta use A/B testing to make user experiences more successful and as a way to streamline their services.

13. Is mean imputation of missing data acceptable practice?

Ans. No, it is not acceptable practice.

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.

Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Linear regression will try to make the best fit line, where the eq**. (y= a + b x)** could be settled**.**

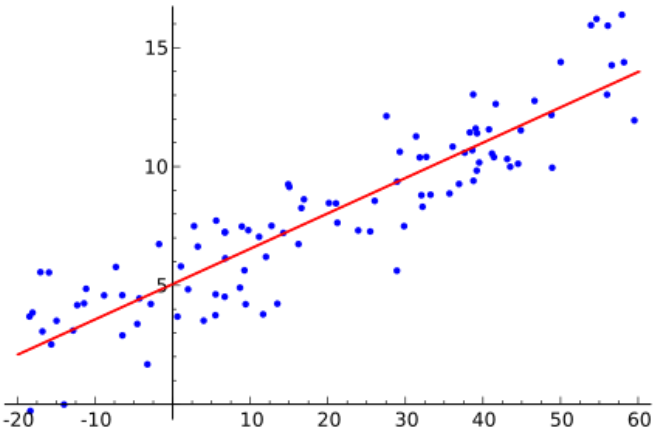
Where,

x= independent variable

Y= dependent variable

b=coefficient value

a=intercept



15.What are the various branches of statistics?

Ans. Statistics has two various branches:

1) Descriptive statistics
2) Inferential statistics

1) Descriptive are divided in 2 parts:

- Central Tendency

A. Mean (i.e Average of all the data)
B. Median (i.e Central position of data)
C. Mode (i.e Frequency of that data which is occurring most of the time)

- Dispersion of data

A. Range (max-min)
B. Standard deviation
C. Variance
D. Skew

2) Inferential statistics has contain some tests like:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis