MACHINE LEARNING

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

   A) Least Square Error
2. Which of the following statement is true about outliers in linear regression?

   A) Linear regression is sensitive to outliers
3. A line falls from left to right if a slope is ______?

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

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

   C) Low bias and high variance
6. If output involves label, then that model is called as:

   B) Predictive modal
7. Lasso and Ridge regression techniques belong to _________?

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

   A) Cross validation

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

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

    A) True

11. Pick the feature extraction from below:

    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?

    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?

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it. Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique. This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model. It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "In regularization technique, we reduce the magnitude of the features by keeping the same number of features.

14. Which particular algorithms are used for regularization?

Understanding the use of Regularization algorithms like LASSO, Ridge, and Elastic-Net regression.

Working of Ridge, LASSO, and Elastic-Net Regression

The working of all these algorithms is quite similar to that of Linear Regression, it’s just the loss function that keeps on changing!

$$Loss = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (w_i x_i + c))^2$$

Ridge Regression- Ridge regression is a method for analyzing data that suffer from multi-collinearity.

$$Loss = \sum_{i=1}^{n} (y_i - (w_i x_i + c))^2 + \lambda \sum_{i=1}^{n} w_i^2$$

Ridge regression adds a penalty ***(L2 penalty)*** to the loss function that is equivalent to the square of the magnitude of the coefficients.
The regularization parameter *(λ)* regularizes the coefficients such that if the coefficients take large values, the loss function is penalized.
λ → 0, the penalty term has no effect, and the estimates produced by ridge regression will be equal to least-squares i.e. the loss function resembles the loss function of the Linear Regression algorithm. Hence, a lower value of λ will resemble a model close to the Linear regression model.

λ → ∞, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will **approach zero** (coefficients are close to zero, but not zero).

LASSO Regression -LASSO is a regression analysis method that performs both feature selection and regularization in order to enhance the prediction accuracy of the model.

$$Loss = \sum_{i=1}^{n} (y_i - (w_i x_i + c))^2 + \lambda \sum_{i=1}^{n} |w_i|$$

LASSO regression adds a penalty ***(L1 penalty)*** to the loss function that is equivalent to the magnitude of the coefficients.In LASSO regression, the penalty has the effect of forcing some of the coefficient estimates to be **exactly equal to zero** when the regularization parameter λ is sufficiently large.

Elastic-Net Regression -Elastic-Net is a regularized regression method that linearly combines the L1 and L2 penalties of the LASSO and Ridge methods respectively.

$$Loss = \sum_{i=0}^{n} (y_i - (w_i x_i + c))^2 + \lambda_1 \sum_{i=0}^{n} |w_i| + \lambda_2 \sum_{i=0}^{n} w_i^2$$

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

Within a linear regression model tracking a stock's price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed.

STATISTICS WORKSHEET

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

a) True

2. Which of1 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

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

b) Modeling bounded count

4. Point out the correct statement.

d) All of the mentioned

5. ______ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

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

a) 0

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

c) Outliers cannot conform to the regression relationship

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

A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.

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

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data. For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, overfitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.

12. What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a methodology for comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

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?

Simple linear regression is used to estimate the relationship between **two quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

15. What are the various branches of statistics

Statistics have majorly categorised into two types:

1. Descriptive statistics

2. Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.