**Machine Learning**

1. The value of correlation coefficient will always be:

Ans: c) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?

Ans: d) Ridge Regularization

3. Which of the following is not a kernel in Support Vector Machines?

Ans: b) Radial Basis Function

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Ans: b) Naive Bayes

5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be?

Ans: c) old coefficient of ‘X’ ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Ans: b) Increases

7. Which of the following is not an advantage of using random forest instead of decision trees?

Ans: c) Random Forests are easy to interpret

8. Which of the following are correct about Principal Components?

Ans: c ) Principal Components are linear combinations of Linear Variables.

9. Which of the following are applications of clustering?

Ans: a) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

d) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. Which of the following is(are) hyper parameters of a decision tree?

Ans: a) max\_depth b) max\_features and d) min\_samples\_leaf

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection

Ans: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. For eg: While analyzing salaries of all the people in India, salary of Mukesh Ambani or Ratan Tata might be an outlier.

Inter Quartile Range (IQR):

IQR = Q3-Q1

Where, Q1 = 25th % ile of the data

Q2 = 50th % ile (a.k.a. median)

Q3 = 75th % ile of the data.

Upper bound = Q3 + 1.5\*Q3

Lower Bound = Q1 – 1.5\*Q1

Any data point lying above than upper bound and lower than lower bound is considered as an outlier.

12. What is the primary difference between bagging and boosting algorithms?

Ans: Bagging: Bagging is also known as bootstrap aggregating sits on top of the majority voting principle. The samples are bootstrapped each time when the model is trained. When the samples are chosen, they are used to train and validate the predictions. The samples are then replaced back into the training set. The samples are selected at random. This technique is known as bagging. To sum up, base classifiers such as decision trees are fitted on random subsets of the original training set. Subsequently, the individual predictions are aggregated (voting or averaging etc.). The final results are then used as predictions. It reduces the variance of a black box estimator. Due to this the chances of overfitting is ruled out.

Boosting: The concept of Adaptive Boost revolves around correcting previous classifier mistakes. Each classifier gets trained on the sample set and learns to predict. The misclassification errors are then fed into the next classifier in the chain and are used to correct the mistakes until the final model predicts accurate results. When a weak-classifier misclassifies a training sample, the algorithm then uses these very samples to improve the performance of the ensemble.

13. What is adjusted R2 in linear regression. How is it calculated?

Ans: Adjusted R2 and R2 both represent that how well the model fits the data points. But adjusted R2 penalizes the model for using more features. In case we increase the number of features in training data the R2 will increase but adjusted R2 will only increase if the new feature adds value to our model. Due to this reason adjusted R2 is considered as a better evaluation metric than R2. Adjusted R2 is always less than or equal to R2.

14. What is the difference between standardisation and normalisation?

Ans: In Normalization a dataset is scaled in such a way that all the data points lie between 0 and 1. Normalization is often called min-max scaling. Whereas, In Standardization a dataset is scaled in such a way that the mean of data points becomes 0 and standard deviation is 1. The transformed data may be positive as well as negative in standardization.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation

Ans: Cross validation is a technique to fit a model on data set. In cross validation the data set is divided into ‘k’ number of sets where ‘k-1’ sets are used for training and 1 set is used as validation set. And this is done for all the set one by one and the final score of model is taken as average score of all the ‘k’ number of fits. Advantage of using Cross validation is that, there is no need of separate validation data, cross validation reduces chances of overfitting and gives a more generic model. Cross validation has a disadvantage that it takes more time to fit the model over a large dataset and the model built is more complex than the basic model.

**Statistics**

1. What is central limit theorem and why is it important?

Ans: Central Limit Theorem says that with a large sample size, sample means are normally distributed. Normally distributed means that a group of numbers follows a bell-shaped curve. Most of the numbers cluster in the middle around the average, and there are fewer numbers at the extremes to the right and left. The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

2. What is sampling? How many sampling methods do you know?

Ans: Sampling methods are the ways to choose people from the population to be considered in a sample survey. Samples can be divided based on following criteria.

1) Probability samples - In such samples, each population element has a known probability or chance of being chosen for the sample.

2) Non-probability samples - In such samples, one cannot be assured of having known probability of each population element.

3. Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true

Ans: Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true

4. What do you understand by the term Normal distribution?

Ans: Normal distribution, also known as the Gaussian 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.

5. What is correlation and covariance in statistics?

Ans: Correlation is considered the best technique for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Given two random variables, it is the covariance between both divided by the product of the two standard deviations of the single variables, hence always between -1 and 1.

Covariance is a measure that indicates the extent to which two random variables change in cycle. It explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.

Ans: Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it. Example: height of students. Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables. Example: temperature and ice cream sales in the summer season. Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable. Example: data for house price prediction

7. What do you understand by sensitivity and how would you calculate it?

Ans: Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.). It can be calculated as: TN/(TN+FP).

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Ans: A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. Ex : you have a coin and you don’t know whether the coin is fair or not. First, decide null and alternate hypothesis and alpha=0.05.

H0 : that is a fair coin.

H1 : the coin is not fair

9. What is quantitative data and qualitative data?

Ans: Quantitative data are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, social, economic status, and religious preference are usually considered to be qualitative data.

10. How to calculate range and interquartile range?

Ans: The interquartile range is a measure of where the middle is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That’s why it’s preferred over many other measures of spread when reporting things like school performance or SAT scores.

The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.

11. What do you understand by bell curve distribution ?

Ans: In a bell curve, the centre contains the greatest number of a value and, therefore, it is the highest point on the arc of the line. This point is referred to the mean, but in simple terms, it is the highest number of occurrences of an element.

12. Mention one method to find outliers.

Ans: Outlier values can be identified by using boxplot or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values are:

a. To change the value and bring in within a range

b. To just remove the value.

13. What is p-value in hypothesis testing?

Ans: p-value is defined as the probability that the data would be at least as extreme as those observed, if the null hypothesis were true. The p-value reflects the strength of evidence against the null hypothesis.

14. What is the Binomial Probability Formula?

Ans: The binomial distribution consists of the probabilities of each of the possible numbers of successes in N trials for independent events that each have a probability of π. Here, the possible outcomes are two.

15. Explain ANOVA and it’s applications.

Ans: Analysis of Variance (ANOVA) is a technique which is used to compare the means of multiple samples. Whether there is a significant difference between the mean of 2 samples, can be evaluated using z-test or t-test but in case of more than 2 samples, t-test cannot be applied as it accumulates the error and it will be more difficult as the number of sample will increase (for example: for 4 samples — 12 t-test will have to be performed). The ANOVA technique enables us to perform this simultaneous test.