Machine Learning

1. Between -1 and 1
2. Recursive feature elimination
3. Linear
4. Logistic Regression
5. same as old coefficient of 'X'
6. Increases
7. Random Forests are easy to interpret
8. All of the above

9. i) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

ii) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.

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

10. max_depth, n_estimators and min_samples_leaf.

11. Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations. Outliers impact the prediction in a dataset and often these needs to be treated before predicting a model. One of the ways to deal Outliers in model building in machine learning is by importing zscore.

Interquartile range or IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 3 equal parts- Q1, Q2, Q3.

Q1 represents the 25th percentile of the data.

Q2 represents the 50th percentile of the data.

Q3 represents the 75th percentile of the data.

12. Bagging and Boosting are two types of Ensemble Learning. These two decrease the variance of a single estimate as they combine several estimates from different models.

Bagging attempts to tackle the over-fitting issue.

Boosting tries to reduce bias.

Each model in bagging is trained parallelly and indeendently where in a final prediction is created from the prediction of every models.

Boosting is an interative process which trains all the models together and gets a certain prediction, a second model is then built ries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

13. The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

14. Normalization is a part of data processing and cleansing techniques. The main goal of normalization is to make the data homogenous over all records and fields. It helps in creating a linkage between the entry data which in turn helps in cleaning and improving data quality. Normalization of data is a type of Feature scaling and is only required when the data distribution is unknown or the data doesn't have Gaussian Distribution.

Standardization is the process of placing dissimilar features on the same scale. Standardized data in other words can be defined as rescaling the attributes in such a way that their mean is 0 and standard deviation becomes 1. It is usually applied when the data has a bell curve i.e. it has gaussian distribution.

15. Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models.

Advantage: It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

Disadvantage: Cross Validation is computationally very expensive in terms of processing power required.

Statistics

1. Central Limit Theorem is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean. The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution. This is very useful there is a large amount of data.

2. A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

There are two types of sampling:

A) Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
B) Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

3. A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.
4. 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. In graphical form, the normal distribution appears as a "bell curve".
5. Both covariance and correlation measure the relationship and the dependency between two variables. Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables. Correlation values are standardized. Covariance values are not standardized

6. Univariate: This type of data consists of only one variable, example: Height.

Bivariate: This type of data involves two different variables, example: Rain and crops harvesting.

Multivariate: When the data involves three or more variables

7. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input. Sensitivity Index = 100 - Percentage of Change required in the current value of the variable to alter the decision. Therefore, here, it is = 100 - 10% = 90%
8. Hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis. Hypothesis testing is formulated in terms of two hypotheses: H0: the null hypothesis; • H1: the alternate hypothesis.
9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

10. Q1 is the cut in the first half of the rank-ordered data set

    Q2 is the median value of the set

Q3 is the cut in the second half of the rank-ordered data set.

Interquartile range = Upper Quartile – Lower Quartile

Q1Q1 = (1/4)[(n + 1)]th term)

Q3Q3 = (3/4)[(n + 1)]th term)

n = number of data points.

11. A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.
12. Zscore is one of the method to find outliers.
13. The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis.
14. The binomial distribution formula is used in statistics to find the probability of the specific outcome-success or failure in a discrete distribution.
15. Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.