

Decision tree:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

Decision trees are composed of three main parts—decision nodes (denoting choice), chance nodes (denoting probability), and end nodes (denoting outcomes). Decision trees can be used to deal with complex datasets, and can be pruned if necessary to avoid overfitting

K means:

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

How do you explain k-means?

K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart. Similarity of two points is determined by the distance between them. There are many methods to measure the distance.

Linear regression:

linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

linear regression means that we can predict a dependent variable from an independent one, so whenever we need to know from the beginning each time we add information.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional sub-space.

To interpret each principal components, examine the magnitude and direction of the coefficients for the original variables. The larger the absolute value of the coefficient, the more important the corresponding variable is in calculating the component.

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

The Mean Squared Error

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

Mean squared error tells us whether or not a regression line is an accurate model for predicting data points in a particular data set. A low mean squared error value indicates an accurate regression model, or, goodness of fit.

R-Squared (R^2 or the coefficient of determination)

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

What is r^2 score? The r^2 score varies between 0 and 100%. It is closely related to the MSE (see below), but not the same. Wikipedia defines r^2 as. ” ...the proportion of the variance in the dependent variable that is predictable from the independent variable(s).”