MACHINE_LEARNING (V) s

In the case of decision tree's, they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behavior of this model that makes it prone to learning every point extremely well — to the point of perfect classification — ie: overfitting.

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.

The regularization parameter in machine learning is λ :

It imposes a higher penalty on the variable having higher values, and hence, it controls the strength of the penalty term. This tuning parameter controls the bias-variance trade-off. λ can take values 0 to infinity If λ = 0, then means there is no difference between a model with and without regularization.

The explained sum of squares, defined as the sum of squared deviations of the predicted values from the observed mean of y, is Using in this, and simplifying to obtain, gives the result that TSS = ESS + RSS if and only if.

The residual standard error (RSE) is another statistical term used to describe the difference in standard deviations of observed values versus predicted values as shown by points in a regression analysis. It is a goodness-of-fitmeasure that can be used to analyze how well a set of data points fit with the actual model.

The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees that do not contain that data point in their respective bootstrap sample. This way, the Random Forest model is constantly being validated while being trained.

K- Fold Cross Validation for Parameter Tuning In this article I will explain about K- fold cross-validation, which is mainly used for hyperparameter tuning. Cross-validation is a technique to evaluate predictive models by dividing the original sample into a training set to train the model, and a test set to evaluate it.

Hyperparameters are the knobs or settings that can be tuned before running a training job to control the behaviour of an ML algorithm. They can have a big impact on model training as it relates to training time, infrastructure resource requirements (and as a result cost), model convergence and model accuracy.

The learning rate can see as step size, η. As such, gradient descent is taking successive steps in the direction of the minimum. If the step size η is too large, it can (plausibly) "jump over" the minima we are trying to reach, Ie. we overshoot. This can lead to osculation's around the minimum or in some cases to outright divergence.

Yes, it might work, but logistic regression is more suitable for classification task and we want to prove that logistic regression yields better results than linear regression. Let’s see how logistic regression classifies our dataset. Logistic regression model, a sigmoid curve that fit the training dataset

The main difference between these two algorithms is that Gradient boosting has a fixed base estimator i.e., Decision Trees whereas in AdaBoost we can change the base estimator according to our needs. What is Boosting technique? Gradient Boosting Algorithm

Finding the right balance between the bias and variance of the model is called the Bias-Variance trade-off. There is inverse relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

The linear, polynomial and RBF or Gaussian kernel are simply different in case of making the hyperplane decision boundary between the classes. The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space with view to making it linear dataset.

The residual standard error (RSE) is another statistical term used to describe the difference in standard deviations of observed values versus predicted values as shown by points in a regression analysis. It is a goodness-of-fit measure that can be used to analyse how well a set of data points fit with the actual model.