What is a simple introduction to boosting in machine learning?

Authored by

stats writer

April 22, 2024

Boosting is a powerful machine learning algorithm that combines multiple weak learners to create a strong learner. It works by iteratively training weak learners on different subsets of the data and then combining their predictions to produce a final prediction. This approach allows for improved accuracy and performance compared to using a single strong learner. Boosting is widely used in various fields of machine learning, such as classification, regression, and clustering. It is a popular technique due to its ability to handle complex data and produce accurate results.

A Simple Introduction to Boosting in Machine Learning

Most supervised machine learning algorithms are based on using a single predictive model like linear regression, logistic regression, ridge regression, etc.

Methods like bagging and random forests, however, build many different models based on repeated bootstrapped samples of the original dataset. Predictions on new data are made by taking the average of the predictions made by the individual models.

These methods tend to offer an improvement in prediction accuracy over methods that only use a single predictive model because they use the following process:

First, build individual models that have high variance and low bias (e.g. deeply grown decision trees).Next, take the average of the predictions made by individual

models in order to reduce the variance.

Another method that tends to offer even further improvement in predictive accuracy is known as boosting.

What is Boosting?

Boosting is a method that can be used with any type of model, but it is most often used with decision trees.

The idea behind boosting is simple:

1. First, build a weak model.

A "weak" model is one whose error rate is only slightly better than random guessing.In practice, this is typically a decision tree with only one or two splits.

2. Next, build another weak model based on the residuals of the previous model.

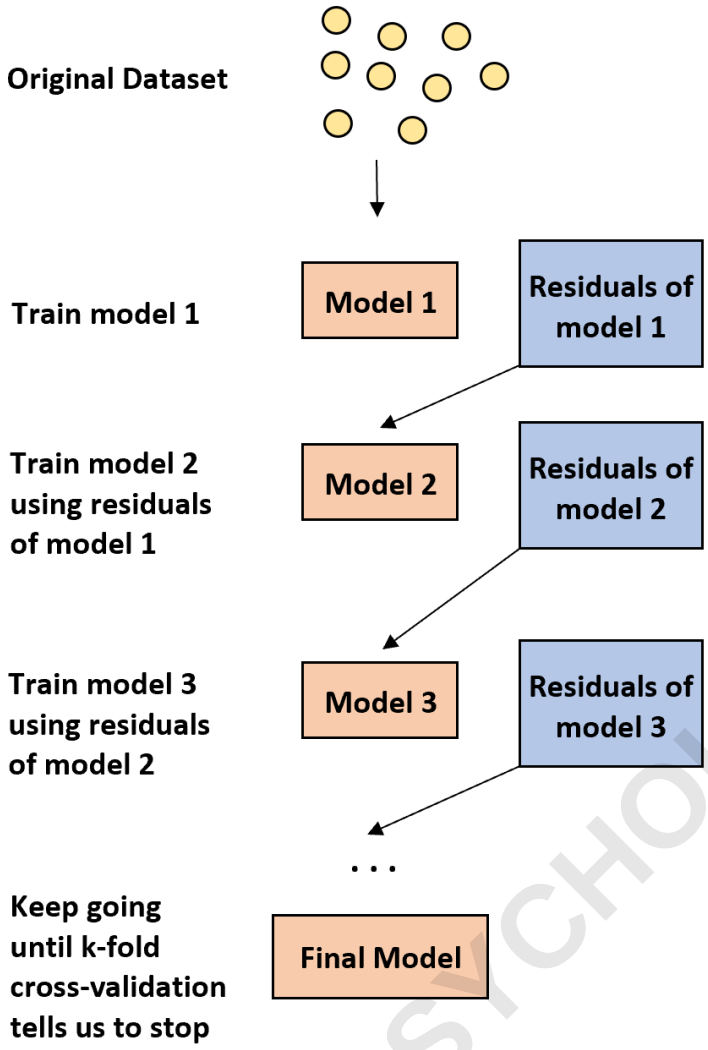
In practice, we use the residuals from the previous model (i.e. the errors in our predictions) to fit a new model that slightly improves upon the overall error rate.

3. Continue this process until k-fold cross-validation

tells us to stop.

In practice, we use k-fold cross-validation to identify when we should stop growing the boosted model.

By using this method, we can start with a weak model and keep "boosting" the performance of it by sequentially building new trees that improve upon the performance of the previous tree until we end up with a final model that has high predictive accuracy.



Why Does Boosting Work?

In many industries, boosted models are used as the go-to models in production because they tend to outperform all other models.

The reason boosted models work so well comes down to understanding a simple idea:

1. First, boosted models build a weak decision tree that has low predictive accuracy. This decision tree is said to have low variance and high bias.

2. As boosted models go through the process of sequentially improving previous decision trees, the overall model is able to slowly reduce the bias at each step without increasing the variance by much.

3. The final fitted model tends to have sufficiently low bias *and* low variance, which leads to a model that is able to produce low test error rates on new data.

Pros & Cons of Boosting

The obvious benefit of boosting is that it's able to produce models that have high predictive accuracy compared to almost all other types of models.

One potential drawback is that a fitted boosted model is very difficult to interpret. While it may offer tremendous ability to predict the response values of new data, it's difficult to explain the exact process that it uses to do so.

In practice, most data scientists and machine learning

practitioners build boosted models because they want to be able to predict the response values of new data accurately. Thus, the fact that boosted models are hard to interpret usually isn't an issue.

Boosting in Practice

In practice there are actually many types of algorithms that are used for boosting, including:

XGBoostAdaBoostCatBoostLightGBM

Depending on the size of your dataset and the processing power of your machine, one of these methods may be preferable to the other.