

Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. Join them; it only takes a minute:

Sign up

Here's how it works:

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

Gradient Boosting Tree vs Random Forest

More jobs means more choice



Get started

Gradient tree boosting as proposed by Friedman uses decision trees as base learners. I'm wondering if we should make the base decision tree as complex as possible (fully grown) or simpler? Is there any explanation for the choice?

Random Forest is another ensemble method using decision trees as base learners. Based on my understanding, we generally use the almost fully grown decision trees in each iteration. Am I right?

machine-learning data-mining random-forest cart gbm

edited Sep 24 '15 at 17:13



Sycorax
22.8k 5 60 98

asked Sep 20 '15 at 20:44



FihopZz
388 1 4 8

2 Answers

$error = bias + variance$

- Boosting is based on **weak** learners (high bias, low variance). In terms of decision trees, weak learners are shallow trees, sometimes even as small as decision stumps (trees with two leaves). Boosting reduces error mainly by reducing bias (and also to some extent variance, by aggregating the output from many models).
- On the other hand, Random Forest uses as you said **fully grown decision trees** (low bias, high variance). It tackles the error reduction task in the opposite way: by reducing variance. The trees are made uncorrelated to maximize the decrease in variance, but the algorithm cannot reduce bias (which is slightly higher than the bias of an individual tree in the forest). Hence the need for large, unpruned trees, so that the bias is initially as low as possible.

Please note that unlike Boosting (which is sequential), RF grows trees in **parallel**. The term **iterative** that you used in thus inappropriate.

edited Feb 15 '16 at 9:09


answered Sep 24 '15 at 17:09



Antoine
2,353 2 12 37

Build smarter apps with cognitive API's and machine learning.

Try Azure free



This question is addressed in this very nice post. Please take a look at it and the references therein. <http://fastml.com/what-is-better-gradient-boosted-trees-or-random-forest/>

Notice in the article that the speaks about calibration, and links to another (nice) blog post about it. Still, I find that the paper [Obtaining Calibrated Probabilities from Boosting](#) gives you a better understanding of what calibration in the context of boosted classifiers is, and what are standard methods to perform it.

And finally one aspect missing (a bit more theoretical). Both RF and GBM are ensemble methods, meaning you build a classifier out of a big number of smaller classifiers. Now the fundamental difference lies on the method used:

1. RF uses decision trees, which are very prone to overfitting. In order to achieve higher accuracy, RF decides to create a large number of them based on **bagging**. The basic idea is to resample the data over and over and for each sample train a new classifier. Different classifiers overfit the data in a different way, and through voting those differences are averaged out.
2. GBM is a boosting method, which builds on **weak classifiers**. The idea is to add a classifier at a time, so that the next classifier is trained to improve the already trained ensemble. Notice that for RF each iteration the classifier is trained independently from the rest.

edited Apr 13 at 12:44



Community ♦

1

answered Feb 13 '16 at 14:46



jpmuc

7,345

15

35