baggingBoostingRFs

November 4, 2019

1 Bagging, Boosting, and Random Forests

Decision trees have many advantages. They are easy to interpret. To make a prediction, you only need to follow a set of rules. Predictions are also data-driven, not having to follow a more structured linear pattern like the past model we studied. Regression trees also have two big disadvantages: (i) they typically have poor performance compared to other regression models, and (ii) suffer from high variance.

1.1 Goal of ensemble

The goal of an ensemble model is to combine many weak predictors, and in doing so, build a model that has lower variance and bias. This phenomena occurs in simple statistics. Given a sample of data $Y_1, Y_2, \cdots, Y_n$ the average $\bar{Y}$ has a Normal distribution with variance σ^2/n. The average lowers the variance.
A single regression tree tends to have high variance. Combining many different regression trees, in an ensemble, attempts to lower the variance.

1.2 Bagging for a continuous target

1.2.1 Bootstrap

1.2.2 Averaging

1.3 Bagging for a categorical target

1.3.1 Majority Vote

1.4 Measuring out of sample error

1.5 Random forests