

Simulation of Bias-Variance Tradeoff with Tree-Based Models

P9120 Final Project

Ravi Brenner

Introduction

The bias-variance tradeoff is a common framework for understanding the error of a prediction model in statistics and machine learning. Once a model is trained on some training data and evaluated on some test data, the mean square error (MSE) can be calculated, which is composed of the bias and the variance in the following way:

$$MSE(\hat{f}(x)) = \text{bias}^2(f(x), \hat{f}(x)) + \text{var}(\hat{f})$$

The goal of most modeling pipelines is to minimize this MSE in some way, either by reducing the variance for simpler models, or reducing the bias for more complicated models. This is necessary because bias and variance have a fundamentally inverse relationship. Take an extremely simple model as an example, like a linear regression with only an intercept term. This model will make the same prediction $\hat{f}(x)$ for all test data points x . Thus the *variance* of the predictions will be low. At the same time, the difference between the predictions and the true values will be quite high, so we say the *bias* is high. On the other end of the spectrum, think of a very complex model like a neural network. The network can be trained to exactly predict every one of the training data points, so the variance of the predictions will be high. Meanwhile the difference between the predictions and the true values will be relatively low, so the bias will be low.

Many examples of the bias-variance tradeoff treat this as mainly a theoretical construct, which it is. While the MSE can be mathematically decomposed into bias and variance terms, there are also error terms that are not accounted for. Furthermore, most textbook examples rely on simulated data in the regression setting only. Here, I will demonstrate the bias-variance tradeoff using real datasets, and apply it to both regression and classification. I will use various types of tree-based and ensemble models to do so, since these are relatively easy to understand but provide enough flexibility in their settings to demonstrate the changes in bias and variance. Finally, I will explore how well the MSE can be estimated by cross validation and out-of-bag error.

Methods

All analysis was conducted in R software using the `tidymodels` framework. Full code is available at

Results

Regression

Bias Variance Tradeoff

Training approximations (CV and OOB error)

Classification

Bias Variance Tradeoff

Training approximations (CV and OOB error)

Consensus Voting vs. Probability Averaging

Discussion