Thomas Courtney

Machine Learning, I

Professor Wilck

2/7/2021

Through the Moneyball data exercise associated with module four, I used three distinct data modeling procedures to validate model accuracy within my linear model. The Moneyball dataset combines aspects and statistics from a baseball team throughout a season to predict the number of wins.

Before determining the best values in my linear model, I first needed to read in the data and remove NA's. I did this by assigning them the mean value of the data variable. I did this throughout both the training data and the test data.

Next, I cut my training data into two sets: training and validation. I used validation before testing it on my model. My seconding training set is 70% of the original training dataset, and 30% of it is now validation.

From there, I ran three procedures to determine new model validation and to use as my linear predictive model. I used lasso, ridge and boot. Each have distinct benefits. Lasso is a modification of linear regression, and the model is penalized for the sum of absolute values of weights. Ridge regression takes this further and penalizes the model of the sum of squared values of the weights. Bootstrapping is a resampling method that estimates properties by measuring those properties when sampling across from an approximating distribution.

After working with the Moneyball dataset and training and validating both tests, I chose ridge regression. I chose this as it gave me the most logical mean (around 80 wins) while having the smallest range of possible values.