PS1

Alex Przybycin

1/24/2021

##Question 1

These two rows of models vary in terms of complexity, bias, and variance. The linear regression models in the top row are less complex than the bottom with greater variance and higher bias. The bottom row of models is less complex, exhibiting less variance and lower bias than the top row. In this the model fit is tighter fitting to data points.

##Question 2

These curves have the shape they do because of the number of observations in each set. The blue line for the training error rate has less observations than the orange line for testing. Jumpiness in the curves is a result of the training set having a small sample size.

##Question 3

An inflexible model would be better for a large sample size and small number of predictors as there would higher bias.

##Question 4 Predictors large, small sample size, In the case of a large number of predictors and a small sample size a flexible model would be better in order to reduce the possibility of over fitting.

##Question 5 When relationship between the predictors and response is highly non-linear, the flexible model would be a better fit. A flexible model will have more degrees of freedom. (p. 32)

#Question 6 Trying to minimize MSE in attempt to find a perfect fit will lead to over generalization/over fitting. If the predicted responses are close to true responses then the MSE will be small. There is no guarantee that the a low training set MSE will result in a low testing set MSE.

##Question 7 Create & plot 2 bootstrapped samples manually without using boot(), drop NA’s Borrow code from 1/19 class

anes <- read\_csv(“data/anes\_pilot\_2016.csv”)

anes\_small <- drop\_na(anes) %>% dplyr::select(ftobama, fttrump)

anes\_smallftobama, anes\_smallfttrump <- replace(anes\_smallfttrump == 88, NA) summary(anes\_small)

##Question 8

The distributions in question 7 do look like what sampling with replacement. The summary statistics vary for variables in each trial because the training data set is not identical in each.

##Question 9

##Question 10 How are bootstrapping and cross-validation approaches to resamping different?

Bootstrapping is resampling with replacement, and same as original sample size. Cross-validation is resampling without replacement. They are similar in that they reuse existing data. Differences in computational costs, bootstrapping is less computationally expensive.

Collaborated with: Sabina Harnett, Egemen Pamucku, Max Kramer, Jade Benson, Alex Levi