STA4241 Homework 5, Fall 2021

Please turn in your own work, though you may discuss these problems with your classmates, professor, and TA. The assignment is due on Friday, November 19th at 11:59pm.

(1) Assume that we have p covariates $\boldsymbol{X}$ and we want to fit a linear model without an intercept. Prove that the effective degrees of freedom for linear regression estimated using least squares is equal to p. Remember that effective degrees of freedom is calculated by taking the trace of the matrix $\boldsymbol{S}$, such that our predictions can be written as

$$\widehat{Y} = \boldsymbol{S}Y$$

(2) Suppose that we estimate a function $f(\cdot)$ using the following

$$\widehat{f} = \arg\min_f \left(\sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int \left[f^{(m)}(x) \right]^2 dx \right)$$

where $f^{(m)}(x)$ is the mth derivative of f. In each of the following examples below, explain what the estimated function will look like. If possible, state exactly what values f will take.

(a) $\lambda = \infty, m = 0$

(b) $\lambda = \infty, m = 1$

(c) $\lambda = \infty, m = 2$

(d) $\lambda = 0, m = 3$

(3) Read in Problem3train.csv and Problem3testing.csv off of the course website. Our goal is to predict Y given the covariates, X. Any models should be fit and tuned using the training data only.

(a) Fit a single regression tree to the data. Use cross-validation to determine how deep the tree should be. Plot your final tree after cross-validation and pruning. What is the prediction error on the testing data set?

(b) Now use bagging to predict Y with 1000 trees. What is the prediction error on the testing data set when you use bagging instead of a single tree?

(c) Now use a generalized additive model (GAM) to predict Y using smoothing splines with 4 effective degrees of freedom for each covariate. What is the prediction error on the testing data set?