STAT 435
SPRING QUARTER 2018

Homework # 7
Due Friday, May 25, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

1. For this problem, you will analyze a data set of your choice, *not* taken from the `ISLR` package. Choose a data set that has $n \gg p$, since you will apply methods from Chapter 7 to this data. You will also need to have $p > 1$. Throughout this problem, make sure to label your axes appropriately, and to include legends when needed.

   (a) Describe the data in words. Where did you get it from, and what is the data about? You will perform supervised learning on this data, so you must identify a response, Y, and features, $X_1, \ldots, X_p$. What are the values of n and p? Describe the response and the features (e.g. what are they measuring; are they quantitative or qualitative?).

   (b) Fit a generalized additive model, $Y = f_1(X_1) + \ldots + f_p(X_p) + \epsilon$. Use cross-validation to choose the level of complexity. For $j = 1, \ldots, p$, make a scatterplot of X_j against Y, and plot $\hat{f}_j(X_j)$. Comment on your results and on the choices you made in fitting this model.

   (c) Now fit a linear model, $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$. For $j = 1, \ldots, p$, display the linear fit $(X_j \hat{\beta}_j)$ on top of a scatterplot of X_j against Y.

   (d) Estimate the test error of the generalized additive model and the test error of the linear model. Comment on your results. Which approach gives a better fit to the data?

2. In this problem, we'll play around with regression splines.

   (a) Generate data as follows:

```
set.seed(7)
x <- 1:1000
y <- sin((1:1000)/100)*4+rnorm(100)
```

Consider the model

$$Y = f(X) + \epsilon.$$

What is the form of $f(X)$ for this simulation setting? What is the value of $\text{Var}(\epsilon)$? What is the value of $E(Y - f(X))^2$?

(b) Fit regression splines for various numbers of knots to this simulated data, in order to get spline fits ranging from very wiggly to very smooth. Make a plot of your results, showing the raw data, the true function $f(X)$, and the spline fits. Be sure to include a legend containing relevant information, and to label the axes appropriately.

(c) Based on visual inspection, how many knots seem to give the "best" fit? Explain your answer.

(d) Now perform cross-validation in order to select the optimal number of knots. What is the "best" number of knots? Make a plot displaying the raw data, the true function $f(X)$, and the spline fit $\hat{f}(X)$ that uses the number of knots selected by cross-validation. Be sure to include a legend and to label the axes appropriately. Comment on your results.

(e) Provide an estimate of the test error, $E(Y - \hat{f}(X))^2$, associated with the spline $\hat{f}(\cdot)$ from (d). How does this relate to your answer in (a)?

(f) Now fit a linear model of the form

$$Y = \beta_0 + \beta_1 X + \epsilon$$

to the data instead. Plot the raw data and the fitted model and the true function $f(\cdot)$. Provide an estimate of the test error associated with the fitted model. Comment on your results.