STAT 435
SPRING QUARTER 2022

Homework # 4
Due Via Online Submission to Canvas: Tuesday, May 24 at 11:30AM

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, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

On this assignment, some of the problems may involve random number generation. Be sure to set a random seed (using the command `set.seed()`) before you begin.

1. Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of n points using the following formula:

$$\hat{g} = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 dx \right),$$

   where $g^{(m)}$ represents the mth derivative of g (and $g^{(0)} = g$). Provide example sketches of $\hat{g}$ in each of the following scenarios.

   (a) $\lambda = \infty, m = 0$.
   (b) $\lambda = \infty, m = 1$.
   (c) $\lambda = \infty, m = 2$.
   (d) $\lambda = \infty, m = 3$.
   (e) $\lambda = 0, m = 3$.

2. Suppose we fit a curve with basis functions $b_1(X) = I(0 \leq X \leq 2) - (X + 1)I(1 \leq X \leq 2)$, $b_2(X) = (2X - 2)I(3 \leq X \leq 4) - I(4 < X \leq 5)$. We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

   and obtain coefficient estimates $\hat{\beta}_0 = 2, \hat{\beta}_1 = 3, \hat{\beta}_2 = -2$. Sketch the estimated curve between $X = -2$ and $X = 6$. Note the intercepts, slopes, and other relevant information.

3. Prove that any function of the form

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - \psi)_+^3$$

is a cubic spline with a knot at ψ.

4. For this problem, we will use the `Wage` data set that is part of the `ISLR` package. Split the data into a training set and a test set, and then fit

   (a) polynomial

   (b) step function

   (c) piecewise polynomial

   (d) cubic spline

   (e) smoothing spline

   models to predict `Wage` using `Age` on the training set. Make some plots, and comment on your results. Which approach yields the best results on the test set?

5. Use the `Auto` data set to predict a car's `mpg`. (You should remove the `name` variable before you begin.)

   (a) First, try using a regression tree. You should grow a big tree, and then consider pruning the tree. How accurately does your regression tree predict a car's gas mileage? Make some figures, and comment on your results.

   (b) Fit a bagged regression tree model to predict a car's `mpg`. How accurately does this model predict gas mileage? What tuning parameter value(s) did you use in fitting this model?

   (c) Fit a random forest model to predict a car's `mpg`. How accurately does this model predict gas mileage? What tuning parameter value(s) did you use in fitting this model?

   (d) Fit a generalized additive model (GAM) model to predict a car's `mpg`. How accurately does your GAM model predict a car's gas mileage? Make some figures to help visualize the fitted functions in your GAM model, and comment on your results.

   (e) Considering both accuracy and interpretability of the fitted model, which of the models in (a)–(d) do you prefer? Justify your answer.