STATS 415 Homework 8

Due Thursday March 22, 2018

Please include your name, uniqname, and lab section (number or time or GSI). A point will be taken off homework without the section info. Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm on the due date.

1. Suppose a curve $\hat{g}$ is fitted to data by computing

$$\hat{g} = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_t (g^{(m)}(t))^2 dt \right)$$

where $g^{(m)}$ represents the m-th derivative of g and $g^{(0)} = g$. Provide an example sketch of $\hat{g}$ in the following scenarios:

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

2. This question uses the following variables from the `Boston` data: `dis` (the weighted mean of distances to five Boston employment centers), `nox` (nitrogen oxides concentration in parts per 10 million), and `indus` (proportion of non-retail business acres). We are interested in predicting the air quality variable `nox`.

   (a) Set the random seed to 34567 and randomly split the data into 80% training and 20% test data. The test data is not used until the last question.
   (b) Fit a smooth nonlinear function on the training data to predict `nox` from `dis` (one predictor). Do it three different ways: polynomial regression, natural spline, and smoothing spline. For each method, choose the relevant "degrees of freedom" parameter by cross-validation, and report and comment on the regression output.

(c) For each of the methods in the previous question, make three plots: the fitted curve with your optimally selected degrees of freedom (df), one with less df, one with more df. Comment on your results.

(d) Fit a GAM on the training data to predict `nox` from `dis` and `indus`.Use what you learned in the previous question to select the best nonlinear function to model `dis`, and use the same type of function for `indus`. Plot the results and explain your findings.

(e) Report the test MSEs for all the methods you have implemented, with relevant parameters suggested by cross-validation. Comment on your results.

Please limit your solution to at most 6 pages.