STAT 542, Homework 4

March 23, 2018
Due date: Apr 6 (Fri), 5 pm to Compass

Requirements: This homework consists of four problems. You should submit your **report and `R` code** as separate files to Compass2g. The report should be in PDF/html/MS Word format, with no more than 10 pages, including plots. Your report should focus on describing the analysis strategy, demonstrating and interpreting the main result. No excessive `R` code should be included in the report. Font size should be 12pt and plots need to be clearly labeled. If you use R markdown and need a template, there is an example source file provided at our course website. This homework worth 100 points total. Late submission penalty is 5 points for each day (round up) of delay.

Question 1: [20 points] In this question, we want to construct the natural cubic spline basis from the cubic spline basis. The truncated power series basis representation for cubic splines with K interior knots $\xi_1, \xi_2, \ldots, \xi_K$ is given by

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$$

To construct the natural cubic spline, we set the following constrains:

$$f(X) \text{ is linear for } X \leq \xi_1 \text{ and } X \geq \xi_K$$

Utilize the constrains (the corresponding derivatives f' and f''), show the following:

1). β_2 and β_3 are both 0

2). $\sum_{k=1}^{K} \theta_k = 0$

3). $\sum_{k=1}^{K} \theta_k \xi_k = 0$

With these results established, show that the power series representation can be rewrote as

$$f(X) = \beta_0 + \beta_1 X + \sum_{k=1}^{K-2} \alpha_k (d_k(X) - d_{K-1}(X)),$$

where

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k} \quad \text{and} \quad \alpha_k = \theta_k(\xi_K - \xi_k).$$

Your report should contain a rigorous proof with clear logic.

Question 2: [20 points] Implement this natural cubic spline method by coding it yourself. Apply your model on the birthrate data (available on the course website). You should consider different choices for the number of knots, and select the best model. Comment on the advantages and disadvantages of natural cubic spline, compared with basis spline.

Question 3: [35 points] Recall in HW2, we did a simulation study to estimate the degrees of freedom. Lets do a similar experiment for random forests. The purpose is to understand the effect of different tuning parameters of random forests. Generate X from independent standard normal distribution with $n = 200$ and $p = 20$ and fix these X values (with `set.seed(1)`). For the true model, use $f(X) = 1 + 0.5 \sum_{j=1}^{4} X_j$ and standard normal errors. Use the `randomForest` package for this question.

a) [20 points] Use the default values of `ntree`, tune different values of `mtry` and `nodesize` (make your own choices). Estimate and compare the degrees of freedom for these models. Summarize your results and comment on the effect of these two parameters.

b) [15 points] Fix `mtry` and `nodesize` as the default value, and tune `ntree`. Estimate the variance of this estimator based on the X values that you originally generated, i.e. $\frac{1}{n}\sum_i E_{\widehat{f}}\big(\widehat{f}(x_i) - E[\widehat{f}(x_i)]\big)^2$. Summarize your results and comment on the effect of `ntree`.

Question 4: [25 points] Write your own code to fit a kernel density estimator for a one dimensional problem. Let's consider just the Gaussian kernel function, and your function should include a tuning parameter for the bandwidth. Perform the following:

a) [10 points] You should generate a one dimensional toy density example, then fit the data using your code. The underlying true density function should have bounded support and two modes. The sample size should be at least 300. Consider tuning the bandwidth on a grid with several distinct values (at least 10). Comment on the boundary effect of this estimator.

b) [15 points] Fix the X values. For each of the bandwidth values you considered, estimate the mean integrated squared error (MISE) using the quantity $\frac{1}{n}\sum_i E_{\widehat{f}}\big(\widehat{f}(x_i) - f(x_i)\big)^2$. To do this, you should consider the same strategy used in the estimation of the degrees of freedom. Comment on the effect of bandwidth on the bias-variance trade-off and the MISE.