Homework 3

MSA 8150 - Machine Learning for Analytics (Spring 2019)
Instructor: Alireza Aghasi
Due date: March 27th, 1PM

March 10, 2019

Please revise the homework guidelines reviewed during the first session. Specifically note that:

- Start working on the homework early
- Late homework **is not accepted** and will receive zero credit.
- Each student must write up and turn in their own solutions
- **(IMPORTANT)** If you solve a question together with another colleague, each need to write up your own solution and **need to list the name of people who you discussed the problem with on the first page of the material turned in**
- The homework should be totally manageable given the material lectured in the class. The long questions are to help clarifying the problem, the majority of the solutions are short

Q1. As you already saw in homework 1, using some basic calculus we can show that if we want to fit the simple model

$$y = \beta + \beta x,$$

to some data points $(x_1, y_1), \cdots, (x_n, y_n)$, the optimal choice of β is (**you do not need to show this**)

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(1+x_i)y_i}{\sum_{i=1}^{n}(1+x_i)^2}. \tag{1}$$

(a) Now consider a Ridge regression problem which requires minimizing

$$RSS_{Ridge} = \sum_{i=1}^{n}(y_i - \beta - \beta x_i)^2 + \lambda\beta^2.$$

Show that in this case the optimal selection of β is

$$\hat{\beta}_R = \frac{\sum_{i=1}^{n}(1+x_i)y_i}{\lambda + \sum_{i=1}^{n}(1+x_i)^2}. \tag{2}$$

(b) If the true regression function is in the form of $f(x) = \beta + \beta x$ and we measure the noisy observations $y = \beta + \beta x + \epsilon$, where ϵ is a random variable with $\mathbb{E}\epsilon = 0$, $var(\epsilon) = \sigma^2$, show that

$$var(\hat{\beta}_R) = \frac{\sum_{i=1}^{n}(x_i+1)^2}{\left(\lambda + \sum_{i=1}^{n}(x_i+1)^2\right)^2}\sigma^2.$$

(c) Going through basic steps (**you do not need to show this**), we can show that for the optimal value of β in (1), $var(\hat{\beta})$ can be calculated as

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i+1)^2},$$

show (mathematically or discuss technically) that for all $\lambda > 0$:

$$var(\hat{\beta}_R) \leq var(\hat{\beta}).$$

Q2. In this question we analyze the data file `Fertility.csv`, which is attached to the homework folder. In this dataset Fertility, the first column, is the response variable, and the other variables are potential predictors. We will use several different statistical modeling techniques. The data set contains 47 rows (samples), split the data into training and test sets. Set the first 30 rows to training samples and the rows 31 through 47 as the test samples.

(a) Fit a linear model to the training set, and report the test error (MSE) obtained.

(b) Fit a Ridge regression model to the training set, with λ chosen by cross-validation. Report the test error obtained.

(c) Fit a LASSO model to the training set, with λ chosen by cross-validation. Report the test error obtained.

(d) Fit a PCR model to the training set, with the number of principal components (M) chosen by cross-validation. Make sure you set the scale option to `TRUE`. Report the test error obtained, along with the value of M selected.

(e) Compare the results of (a), (b), (c) and (d). Which one seems to outperform the others for this specific setup?

Q3. This question performs two simple nonlinear modelings via polynomials and splines, followed by an exercise in GAMs. The data set is `Auto.csv`, which is available in the homework folder.

(a) After reading the data file, plot mpg (response) in terms of horsepower, and fit mpg to a polynomial of order 6 in terms of horsepower. Plot the 95% confidence intervals for your fit.

(b) Fit mpg (response) to a natural spline with 6 degrees of freedom in terms of horsepower. Plot the 95% confidence intervals for your fit. Compared to the result in (a), which fit has a narrower confidence interval around the boundaries?

(c) Split the data into training and test sets. The first 350 samples form the training set and the samples 351 to 392 form the test samples. Fit a linear model that models mpg in terms of `horsepower, acceleration` and `year` and evaluate the test error. Also fit a GAM that models mpg in terms of f_1(`horsepower`), f_2(`acceleration`) and f_3(`year`), where f_1 and f_2 are natural splines with 4 degrees of freedom and f_3(`year`) = `year` (i.e., the feature is passed as it is). All features are treated as numerical entries. Compare the test error between the linear model and GAMs. Which one produces a smaller test error?