Assignment 5

CS 750/850 Machine Learning

- **Due**: March 9th at 11:59PM
- **Submisssion**: Turn in both a **PDF** and the **source code** on MyCourses
- **Questions**: Piazza and Office hours: *Marek*: Wed 1:30-3:00pm, *Soheil*: Mon 2-4pm, *Xihong*: Thu 1:30-3:30pm
- **Extra credit**: Especially good questions or helpful answers on Piazza regarding the assignment earn up to 5 points extra credit towards the assignment's grade.

Problem 1 [25%]

It is mentioned in Chapter 7 of ISL that a cubic regression spline with one knot at ξ can be obtained using a basis of the form x, x^2, x^3, $[x-\xi]^3_+$, where $[x-\xi]^3_+ = (x-\xi)^3$ if $x > \xi$ and equals 0 otherwise. We will now show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 [x-\xi]^3_+$$

is indeed a cubic regression spline, regardless of the values of β_0,β_1,β_2, β_3,β_4.

1. Find a cubic polynomial
$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$
such that $f(x) = f_1(x)$ for all $x \le \xi$. Express a_1,b_1,c_1,d_1 in terms of β_0,β_1,β_2,β_3,β_4.
2. Find a cubic polynomial
$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$
such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2,b_2,c_2,d_2 in terms of β_0,β_1,β_2,β_3,β_4. We have now established that $f(x)$ is a piecewise polynomial.
3. Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.

Problem 2 [25%]

Use linear, cubic, and natural regression splines investigated Chapter 7 of ISL to the `Auto` data set. Is there evidence for non-linear relationships in this data set? Create some informative plots to justify your answer.

Problem 3 [25%]

You will now derive the Bayesian connection to the lasso as discussed in Section 6.2.2. of ISL.

1. Suppose that $y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a normal distribution $\mathcal{N}(0,1)$. Write out the likelihood for the data as a function of values β.

2. Assume that the prior for $\beta : \beta_1, \ldots, \beta_p$ is that they are independent and identically distributed according to a *Laplace* distribution with mean zero and variance c. Write out the posterior for β in this setting using Bayes theorem.

3. Argue that the lasso estimate is the value of β with maximal probability under this posterior distribution. Compute log of the probability in order to make this point. *Hint*: The denominator (= the probability of data) can be ignored in computing the maximum probability.

4. Suppose that $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed according to the Laplace distribution. What are the maximum likelihood/MAP estimates of β_i under this assumption? *Hint*: See https://en.wikipedia.org/wiki/Least_absolute_deviations

Problem 4 [25%]

Based on a true story, according to: The Drunkard's Walk: How Randomness Rules Our Lives, Leonard Mlodinow

Suppose that you applied for a life insurance and underwent a physical exam. The bad news is that your application was rejected because you tested positive for HIV. The test's *sensitivity* is 99.7% and *specificity* is 98.5% [https://en.wikipedia.org/wiki/Diagnosis_of_HIV/AIDS#Accuracy_of_HIV_testing]. However, after studying the CDC website, you find that in your ethnic group (age, gender, race, . . .) only one in 10,000 people is infected. What is the probability that you actually have HIV?