Optimization and Machine Learning, Spring 2020
Homework 1

(Due Wednesday, Mar. 18 at 11:59pm (CST))

March 14, 2020

1. Suppose that we have N training samples, in which each sample is composed of p input variables and one continuous/binary response.
   (a) Please define the input and output variables, and show a linear relationship between them. (5 points)
   (b) Please define a data matrix and corresponding response vector, and find your i-th ($i = 1, ..., N$) sample with its response. (5 points)
   (c) Please use the least squares to estimate the parameters of the linear model in (a) based on the dataset in (b), and explain in which case the solution is unique. (10 points)
   (d) Is there any way to get an unique closed-form solution once the least squares fails? If yes, please show how do you obtain the solution. (5 points)
   (e) How can you select the best model in (d) based only on your training data. (5 points)

2. Given the input variables $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$, the Expected Prediction Error (EPE) is defined by
$$\text{EPE}(\hat{f}) = \mathbb{E}[L(Y, \hat{f}(X))], \tag{1}$$
where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, \hat{f}(X))$ is a loss function measuring the difference between the estimated $\hat{f}(X)$ and observed Y.
   (a) Given the squared error loss $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$, please derive the regression function $\hat{f}(x) = \mathbb{E}(Y|X = x)$ by minimizing $\text{EPE}(\hat{f})$ w.r.t. $\hat{f}$. (5 points)
   (b) Please explain why the nearest neighbors is an approximation to the regression function in (a). (5 points)
   (c) Please explain how the least squares approximates the regression function in (a). (5 points)
   (d) Please discuss the difference between the nearest neighbors and the least squares based on your results in (b) and (c). (5 points)

3. Given a set of observation pairs $(x_1, y_1) \cdots (x_N, y_N)$. By assuming the linear model is a reasonable approximation, we consider fitting the model via least squares approaches, in which we choose coefficients β to minimize the residual sum of squares (RSS),
$$\hat{\beta}_0,\ \hat{\beta} = \underset{\beta_0,\ \beta}{\text{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta x_i)^2.$$
   (a) Show that
$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}\bar{x}, \end{aligned} \tag{2}$$
   where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ are the sample means. (3 points)
   (b) Using (2), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$. (2 points)

4. Given a set of training data $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$ from which to estimate the parameters $\boldsymbol{\beta}$, where each $\boldsymbol{x}_i = [x_{i1}, \cdots, x_{ip}]^T$ denotes a vector of feature measurements for the ith sample. Consider a linear regression problem in which we want to "weight" different training examples differently. Specifically, suppose we aim at minimizing
$$\text{RSS}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2. \tag{3}$$

(a) Show that $\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T \boldsymbol{W} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})$ for an appropriate diagonal matrix $\boldsymbol{W}$, and where $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^T$ and $\boldsymbol{y} = [y_1, \cdots, y_N]^T$. State clearly what $\boldsymbol{W}$ is. (1 points)

(b) By finding the derivative $\nabla_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta})$ and setting that to zero, write the normal equations to this weighted setting and give the value of $\boldsymbol{\beta}$ that minimizes $\text{RSS}(\boldsymbol{\beta})$ in closed form as a function of $\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{y}$. (2 points)

(c) Suppose the y_i's were observed with differing variances. To be specific, suppose that

$$p(y_i | \boldsymbol{x}_i; \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{2\sigma_i^2} \right), \tag{4}$$

i.e., y_i has mean $\boldsymbol{x}_i^T \boldsymbol{\beta}$ and variance σ_i^2, where the σ_i's are fixed, known, constants). Show that finding the maximum likelihood estimate of β is equivalent to solving a weight linear regression problem. State clearly what the w_i's are in terms of the σ_i's. (4 points)

5. To perform variable selection, three classical approaches were introduced in class, including variable subset selection, forward stepwise selection and backward stepwise selection.

(a) To deepen your understanding of these approaches, please make a table to describe their key procedures as well as the pros and cons. (6 points)

(b) Suppose we perform these three approaches on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \cdots, p$ predictors. **Explain** your answers:

- Which of the three models with k predictors has the smallest training RSS? (1 points)
- Which of the three models with k predictors has the smallest test RSS? (1 points)

(Note that: Solutions with the correct answer but without adequate explanation will not earn credit.)

6. Refer to [1, Ex. 3.5]. Consider the ridge regression problem

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{5}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage. Show that problem (5) is equivalent to the problem

$$\hat{\beta}^{\text{c}} = \underset{\beta^{\text{c}}}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0^{\text{c}} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j) \beta_j^{\text{c}})^2 + \lambda \sum_{j=1}^{p} (\beta_j^{\text{c}})^2 \right\}. \tag{6}$$

Give the correspondence between β^{c} and the original β in (5). Characterize the solution to this modified criterion. Moreover, show that a similar result holds for the least absolute shrinkage and selection operator (LASSO). (10 points)

7. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the LASSO may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or LASSO model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting. (2 points)

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$. (4 points)

(c) Write out the LASSO optimization problem in this setting. (2 points)

(d) Argue that in this setting, the LASSO coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions. (2 points)

8. Refer to [1, Ex. 3.30]. Consider the elastic-net optimization problem:

$$\min_{\beta} \ \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \left[\alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1 \right]. \tag{7}$$

Show how one can turn this into a LASSO problem, using an augmented version of $\boldsymbol{X}$ and $\boldsymbol{y}$. (10 points)

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.