Statistical Learning: Homework 4

Due 30 Ordibehesht 1396 (Before the class)

18 Ordibesht 1396

Instructions: Append your Matlab (or Python) code to the end of your homework. In your solutions, you should just present your Matlab output (e.g. numbers, table, figure) or snippets of Matlab code as you deem it appropriate. Make sure to present your results (i.e., your Matlab ouput) in a clear and readable fashion. Careless or confusing presentations will be penalized.

1. (i) Given $y \in \mathbb{R}^n$, consider ridge regression with predictor matrix $X = I_{n\times n}$, i.e.,

$$\begin{aligned}\hat{\beta}^{ridge} &= \arg\min_{\beta\in\mathbb{R}^n} \|y-\beta\|_2^2+\lambda\|\beta\|_2^2 \\ &= \arg\min_{\beta\in\mathbb{R}^n} \sum_{i=1}^n (y_i-\beta_i)^2 + \lambda\sum_{i=1}^n \beta_i^2\end{aligned}$$

Show that the solution is

$$\hat{\beta}_i^{ridge} = \frac{y_i}{1+\lambda},\ i=1,\dots,n$$

(ii) For the lasso with identity predictor matrix,

$$\begin{aligned}\hat{\beta}^{lasso} &= \arg\min_{\beta\in\mathbb{R}^n} \|y-\beta\|_2^2+\lambda\|\beta\|_1 \\ &= \arg\min_{\beta\in\mathbb{R}^n} \sum_{i=1}^n (y_i-\beta_i)^2 + \lambda\sum_{i=1}^n |\beta_i|\end{aligned}$$

the solution (bonus problem) is

$$\hat{\beta}_i^{lasso} = \begin{cases} y_i+\lambda/2 & y_i < -\lambda/2 \\ 0 & |y_i|\le\lambda/2 \\ y_i-\lambda/2 & y_i > \lambda/2 \end{cases},\ i=1,\dots,n$$

For a fixed value of λ (e.g., you can take $\lambda = 1$), draw $\hat{\beta}_i^{ridge}(y_i)$ and $\hat{\beta}_i^{lasso}(y_i)$ as functions of y_i. Describe the difference between these two coefficient functions.

(iii) Suppose that $X \in \mathbb{R}^{n\times p}$ is orthogonal, i.e., $X^TX = I_{p\times p}$. Consider the ridge regression and lasso problems:

$$\hat{\beta}^{ridge} = \arg\min_{\beta\in\mathbb{R}^n} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$
$$\hat{\beta}^{lasso} = \arg\min_{\beta\in\mathbb{R}^n} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Let $X_i \in \mathbb{R}^n$ denote the ith column of X. Show that the solutions $\hat{\beta}^{ridge}$, $\hat{\beta}^{lasso}$ are given by the same function as in parts (i) and (ii), but with X_i^Ty in place of y_i (and p in place of n).
(Hint 1: if $O \in R^{n\times n}$ is an orthogonal and square matrix, recall that it preserves distances, i.e., $\|Oz\|_2 = \|z\|_2$ for any $z \in \mathbb{R}^n$.)
(Hint 2: using Hint 1 and the fact that X has orthonormal columns, show that

$$\|y - X\beta\|_2^2 = \|X^Ty - \beta\|_2^2 + c$$

where c is a constant, meaning that it doesn't depend on β.)

(iv) If $X \in \mathbb{R}^{n\times p}$ is orthogonal, what are the linear regression coefficients $\hat{\beta}^{LS}$ of y on X? Given your answers for the ridge regression and lasso coefficients in part (iii) (and the picture you drew in part (ii)), give a few sentences interpreting the ridge and lasso coefficients as a functions of the linear regression coefficients.

2. Consider the linear regression of predictors $y \in \mathbb{R}^n$ on predictors $X \in \mathbb{R}^{n\times p}$. Let $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ denote the predictors measurements, i.e., these are the rows of X. Recall that the linear regression estimator is given by

$$\hat{f}(x_i) = x_i^T\hat{\beta}^{LS} = x_i^T(X^TX)^{-1}X^Ty.$$

In this problem you will prove the formula,

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}, \tag{1}$$

where $S = X(X^TX)^{-1}X^T$, and $\hat{f}^{-i}$ is the linear regression estimator fit to all but the ith training pair x_i, y_i. This gives us big savings when computing the leave-one-out cross-validation error:

$$\frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}^{-i}(x_i))^2 = \frac{1}{n}\sum_{i=1}^n \left[\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}\right]^2 \tag{2}$$

because we don't have to actually compute $\hat{f}^{-i}$ for each i.

(i) Recall that $S = X(X^TX)^{-1}X^T$, and let $Z \in \mathbb{R}^{(n-1)\times p}$ denote the predictor matrix X but its ith row removed. Argue that

$$S_{ii} = x_i^T(X^TX)^{-1}x_i \quad \text{and} \quad \hat{f}^{-i}(x_i) = x_i^T(S^TS)^{-1}Z^Ty_{-i},$$

where $y_{-i} \in \mathbb{R}^{n-1}$ denotes the observation vector y but with its ith component removed. Argue also that

$$X^TX = Z^TZ + x_ix_i^T \text{ and } X^Ty = Z^Ty_{-i} + x_iy_i.$$

(ii) For a matrix $A \in \mathbb{R}^{k\times k}$ and vectors $u, v \in \mathbb{R}^k$, the Sherman-Morrison update formula gives the inverse of $A + uv^T$ in terms of the inverse A:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}.$$

Using this formula, prove that $\hat{f}^{-i}(x_i)$ can be expressed as

$$\hat{f}^{-i}(x_i) = \frac{\hat{f}(x_i) - S_{ii}y_i}{1 - S_{ii}}.$$

Rearrange this to conclude the results in (1). (Hint: use your results from (i), and the Sherman-Morrison formula to express $(Z^TZ)^{-1} = (X^TX - x_ix_i^T)^{-1}$ in terms of $(X^TX)^{-1}$.)

(iii) Prove the result (1) when $\hat{f}$ is the ridge regression estimator, $\hat{f}(x_i) = x_i^T\hat{\beta}^{ridge}$ at any arbitrary tuning parameter value $\lambda \geq 0$.

3. Given a response vector $y \in \mathbb{R}^n$, predictor matrix $X \in \mathbb{R}^{n\times p}$, and tuning parameter $\lambda \geq 0$, recall the ridge regression estimate,

$$\hat{\beta}^{ridge} = \arg\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

(i) Show that $\hat{\beta}^{ridge}$ is simply the vector of linear regression coefficients from regressing the response $\tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{n+p}$ onto the predictor matrix $\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix} \in \mathbb{R}^{(n+p)\times p}$, where here $0 \in \mathbb{R}^p$, and $I \in \mathbb{R}^{p\times p}$ is the identity matrix.

(ii) Show that the matrix $\tilde{X}$ always has full column-rank, i.e., its columns are always linearly independent, regardless of the columns of X. Hence argue that the ridge regression estimate is always unique, for any matrix of predictors X.

(iii) Write out an explicit formula for $\hat{\beta}^{ridge}$ involving X, y, λ. Conclude that for any $a \in \mathbb{R}^p$, the estimate $a^T\hat{\beta}^{ridge}$ is a linear function of y.

(iv) Now consider the estimation of $a^T\beta^*$, with β^* be true coefficient vector. Based on what we've seen in class, ridge regression can have a lower MSE than linear regression. But we have seen that the linear regression estimate is BLUE, (see Gauss–Markov Theorem). Given that it is indeed linear (part (iii)), what does this imply about the ridge regression estimate, $a^T\hat{\beta}^{ridge}$?

(v) Let X have singular value decomposition $X = UDV^T$, where $U \in \mathbb{R}^{n\times r}$, $D \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{p\times r}$, U, V have orthonormal columns, and D is diagonal with elements $d_1 \geq \ldots d_r \geq 0$. Rewrite your formula for the ridge regression solution $\hat{\beta}^{ridge}$ from (iii) by replacing X with UDV^T, and simplifying the expression as much as possible.

(vi) Assume that

$$y = X\beta^* + \epsilon, \quad \text{with } E[\epsilon] = 0,\ Cov(\epsilon) = \sigma^2 I,$$

and let $a \in \mathbb{R}^p$. Prove that $a^T\hat{\beta}^{ridge}$ is indeed a biased estimate of $a^T\beta^*$, for any $\lambda > 0$.

4. Consider the spam data, which is described in the chapter 1 of ESL book. Repeat the analysis of Table 3.3 on page 63 on the spam data. While we didn't cover the PLS algorithm in the class, it is described in the ESL book. The added points are given to students for writing this function. Also note that you should write your own codes for LS, PCR, and Ridge. About Lasso, you can use available functions in the Matlab or everywhere you find. Choose the penalty terms in Ridge and Lasso using 10-fold cross-validation method.