Machine Learning for Economic Analysis
Problem Set 6

Jonas Lieber[*]

Due: 11:59pm Wed, March 6, 2024

Problem 1. *Coordinate Descent for Regularized Regression*

In ps6.csv, you are given $(Y_1, X_1), \ldots, (Y_n, X_n)$, stacked together vertically so that every row is one observation. We denote the stacked variables by Y and X. Consider some $\alpha \in [0, 1]$ and $\lambda \geq 0$. The purpose of this problem is to minimize the function

$$\begin{aligned} f(y, X, \beta, \lambda, \alpha) &= \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i'\beta)^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha)\frac{1}{2} \|\beta\|_2^2 \right) \\ &= \frac{1}{2n} \|Y - \beta_0 1_{n\times 1} - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha)\frac{1}{2} \|\beta\|_2^2 \right) \end{aligned}$$

1. *Getting rid of the Ridge penalty by "data-augmentation"*

   (a) *For any two vector $z_1 \in \mathbb{R}^{m_1}$, $z_2 \in \mathbb{R}^{m_2}$, we define the vector $z = (z_1, z_2) \in \mathbb{R}^{m_1+m_2}$. Show that*

   $$\|z\|_2^2 = \|z_1\|_2^2 + \|z_2\|_2^2$$

   (b) *Show that for $\tilde{y} = (y^t, 0^t_{p\times 1})^t$, $\tilde{I} = (1^t_{n\times 1}, 0^t_{p\times 1})^t$ and $\tilde{X} = (X^t, \sqrt{n(1-\alpha)\lambda} I_p)^t$ we have*

   $$f(y, X, \beta, \lambda, \alpha) = \frac{1}{2n} \left\| \tilde{y} - \tilde{I}\beta_0 - \tilde{X}\beta \right\|_2^2 + \lambda\alpha \|\beta\|_1,$$

   where I_p is the identity matrix. In the rest of the question, we will consider only the LASSO problem and denote the outcome by y (instead of $\tilde{y}$), the regressors/features by X (instead of $\tilde{X}$) and the number of rows by $nrow(X)$ which could be either n (if $\alpha = 0$) or $n + p$ (if $\alpha > 0$).

   (c) *Why is this useful?*

2. *Coordinate Descent Update*

   (a) *Suppose that we are given a candidate value $\tilde{\beta}$. For some $j \in \{1, \ldots, p\}$, we consider the problem*

   $$\min_{\beta} f(y, X, \beta, \alpha) \quad s.t. \quad \beta_l = \tilde{\beta}_l \text{ for all } l \neq j. \qquad (1)$$

   Describe this program in words.

[*] Department of Economics, Yale University. jonas.lieber@yale.edu

(b) *Implement a function that takes $y, X, \lambda, \alpha, \tilde{\beta}, j$ where X has been standardized, i.e. for each column j, the j-th column of X, denoted by x_j, satisfies*

$$\overline{x_j} := \frac{1}{nrow(X)} \sum_{i=1}^{nrow(X)} X_{i,j} = 0$$

and

$$\overline{x_j^2} = \frac{1}{nrow(X)} \sum_{i=1}^{nrow(X)} X_{i,j}^2 = 1.$$

and calculates

$$\beta_j^* = S\left(\frac{1}{n}x_j^t\left(y - \left(\tilde{\beta}_0 + \sum_{\substack{l=1\\l\neq j}} \tilde{\beta}_l x_l\right)\right), \alpha\lambda\right),$$

where S is the soft-thresholding function, i.e. for $a, b \in \mathbb{R}$,

$$S(a,b) = sign(a)(|a| - b)_+,$$

where

$$sign(a) = \begin{cases} -1 & \textit{if } a < 0 \\ 0 & \textit{if } a = 0 \\ 1 & \textit{if } a > 0, \end{cases}$$

and

$$(a)_+ = \begin{cases} a & \textit{if } a > 0, \\ 0 & \textit{if } a \leq 0, \end{cases}.$$

The function is to return β_j^*.[1]

(c) *Plot the functions $S(z,1)$ and $S(z,0)$ $z \in [-3,3]$. Interpret these functions.*[2]

3. *Implement a function for cyclic coordinate descent by normalizing the regressors, using this update for $j = 0$ (which λ is used here?) and then $j = 1, \ldots, p$, then again $j = 0$, $j = 1, \ldots, p$ etc until convergence. You can stop when the change in the objective function is below $\varepsilon = 0.1$.*

[1] One can avoid recomputing the sum in the first argument of S by updating the residual in place (e.g. by passing a reference). That is more memory-efficient, but since the notation on this problem set is already sufficiently involved, you are not asked to do this on this problem set. To do it, you would first have to add $x_j\tilde{\beta}_j$ to the residual, run the update for β_j and then subtract $x_j\beta_j^*$ from the residual passed in.

[2] Hints: One of these two functions relates to coordinate descent for OLS. Why? How "likely" is it that either function is zero (for example if z is uniformly distributed between -3 and 3)? How is this related to sparsity of the estimates?

4. *Add an option in this function for an active set strategy: After cycling through the βs, you cycle only through the non-zero parameters until you have convergence for these parameters. Only then you cycle through all βs again.*

5. *Compute the λ sequence of the glmnet package:*

   (a) *Choose a length l of the sequence, with default $l = 100$.*

   (b) *Define $\lambda_{\max}$ as (for standardized columns x_j of X)*

   $$\lambda_{\max} = \frac{1}{\sqrt{n}} \max_{j \in \{1,\ldots,p\}} \left| x_j^t y \right| .$$

   (c) *Choose a small $\delta > 0$, with default $\delta = 0.0001$ and set*

   $$\lambda_{\min} = \delta \lambda_1 .$$

   (d) *Take an equidistant sequence of length l from $\log(\lambda_{\min})$ to $\log(\lambda_{\max})$. The exponential of this sequence is the lambda sequence.*

6. *Add an option in your cross-validation sequence to use "warm-starts": As a starting point for the highest value of λ, use $\beta = 0$. As you move to smaller λ values, use the minimizer for the next higher λ as a starting value.*

7. *Run Lasso with this grid and 5-fold CV. Plot the parameter estimates as a function of λ (all in one plot). Report the running time with and without active set strategy and with and without warm starts for cross-validation.*

8. *Run Ridge with this grid and 5-fold CV. Plot the parameter estimates as a function of λ. Report the running time with and without active set strategy and with and without warm starts for cross-validation.*

9. *Now we consider the elastic net. As a grid for α, use $0, 0.1, 0.2, \ldots, 0.9, 1$. Run the elastic net where you cross-validate both α and λ with 5-fold CV. Report the running time with and without active set strategy and with and without warm starts for cross-validation.*

10. *Which estimator(s) leads to sparse estimates?*

11. *(**graduate students & groups of 3 only**) Now benchmark your implementation of coordinate descent for OLS (which λ do you have to choose?) with your implementation for gradient descent from problem set 4 using reg.csv from problem set 4. Redo the plot from problem set 4 with batch gradient descent and cyclic coordinate descent.* [3]

[3] Hint: It is important to think about what to use on the x-axis. What is the number of iterations for cyclic coordinate descent? A single coordinate update? 10 coordinate updates? A cycle of coordinate updates? One way to make these two comparable in the plot might be to use time on the x-axis.

Problem 2. *Kernel Ridge*

1. *Consider ps6.csv again. Consider the Ridge estimator for $\lambda = 1$. Show numerically that*

$$(X'X + \lambda I_p)^{-1}X'y = X'(XX' + \lambda I_n)^{-1}y.$$

   When do you prefer which formula?

2. *Now consider the DGP from problem 1 of problem set 1 again. Using an RBF kernel with $\sigma = 1$, use a kernel Ridge regression of y on x and plot the estimated function together with the true function and the data for $\lambda = 0.1, 0.5, 1, 5, 10$.*

3. *Use a kernel Ridge regression of y on x and plot the estimated function together with the true function and the data for $\sigma = 0.1$ and $\lambda = 0.005$. Briefly compare this plot to the plot of fitted functions in problem set 1.*