Ridge and Lasso

- Ridge and Lasso are efficient tools to do variable selection.
- Ridge regression:

$$\min f(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2 + \frac{\lambda}{2}\langle \beta, \beta \rangle$$

- Lasso:

$$\min \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1., \lambda \geq 0.$$

- They augment the least square loss in regression with an extra component: regularization, to give extra information of the coefficient estimates.

Ridge and Lasso

- Lasso solution is more sparse than Ridge solution, thus easier to interpret. Ridge solution tends to pick all p predictors.
- Ridge regression has a closed form solution, although in high dimensions, iterative methods are needed.
- Lasso seems to be a very attractive option with high dimensional data, since one expects only a small number of predictors to be important.

A simple special case for ridge and lasso

- Consider a special case: n=p, and X is the identity matrix. The regression problem is:

$$\frac{1}{2}\sum_{i=1}^{p}(y_i - \beta_i)^2 \text{ or } \frac{1}{2}\|y - \beta\|_2^2.$$

- The solution is simply $\hat{\beta}_i = y_i$.
- The ridge regression is :

$$\frac{1}{2}\sum_{i=1}^{p}(y_i - \beta_i)^2 + \frac{\lambda}{2}\sum_{i=1}^{p}\beta_i^2 \text{ or } \frac{1}{2}\|y - \beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2$$

A simple special case for ridge and lasso

- In this scenario the ridge solution is : $\hat{\beta}_i = \frac{y}{1+\lambda}$
- The lasso regression is:

$$\frac{1}{2}\sum_{i=1}^{p}(y_i - \beta_i)^2 + \lambda \sum_{i=1}^{p} |\beta_i| \text{ or } \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1.$$

- The lasso estimates take the form

$$\beta_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \\ 0 & \text{if } \mathit{otherwise} \end{cases}$$

Lasso solution

- Ridge regression is consider an easier problem. It has "closed form" solution. Moreover, the objective function is differentiable.
- Iterative methods such gradient method, Newton method can be easily implemented with differentiable function (evaluation of gradients and Hessian matrices).
- We have an optimality condition that be used to verify where we are in the search for the solution.
- Bad news: Lasso regression is not differentiable.

Lasso solution

- Absolute value function is not differentiable.

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

- We can verify that : $\lim_{x \to 0^+} |x| = \lim_{x \to 0^-} |x| = 0$. The function is continuous.
- However,

$$\lim_{h \to 0^+} \frac{|0 + h| - |0|}{h} = 1$$

and

$$\lim_{h \to 0^-} \frac{|0 + h| - |0|}{h} = -1$$

- Therefore the function $|\beta|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_p|$ is not differentiable at 0.

Gradient descent method revisited

- Gradient descent method with fixed steplength for a differentiable function f at an iteration k:

$$\begin{aligned} x^{k+1} &= x^k + \alpha d^k, \text{ with: } d^k = -\nabla f(x^k). \\ &= x^k - \alpha \nabla f(x^k) \end{aligned}$$

- The kth iteration is actually equivalent to solving a sub-problem: $x^{k+1} = \min_x f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\alpha} \|x - x^k\|_2^2$.
- Lets verify this: Consider the problem

$$\min_x g(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\alpha} \|x - x^k\|_2^2$$

Lasso solution

- Take the derivative in terms of x:

$$\begin{aligned}\nabla g(x) &= \nabla f(x^k) + 2\frac{1}{2\alpha}(x - x^k)\\ &= \nabla f(x^k) + \frac{1}{\alpha}(x - x^k).\end{aligned}$$

- Set this derivative to 0 we get

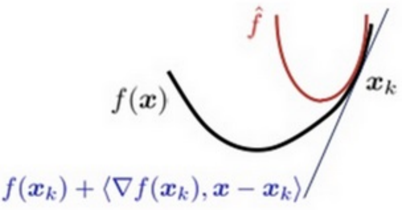
$$\begin{aligned}\nabla f(x^k) + \frac{1}{\alpha}(x - x^k) &= 0\\ x - x^k &= -\alpha\nabla f(x^k)\\ x &= x^k - \alpha\nabla f(x^k).\end{aligned}$$

- So taking a fixed steplength α is equivalent to solving the problem above.

Lasso solution

- A different interpretation of gradient descent method.
- The sub-problem k is has two components. The first one: $f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$ is the first order Taylor approximation.
- It linearly approximate the original objective function. Solving a linear approximation of the original problem hopefully get you to the true optimal solution.
- The second component: $\frac{1}{2\alpha}\|x - x^k\|_2^2$ keeps the Taylor approximation in a neighborhood around the point x^k.
- Its logical to search for the next improved candidate solution in the neighborhood of the current point.

Gradient descent-ish method to find Lasso solution



Gradient descent-ish method to find Lasso solution

We will start using the language of linear approximation and subproblem to describe the solution method to Lasso regression.

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \lambda > 0$$

Also denote $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$, $g(\beta) = \lambda\|\beta\|_1$

- Initialization: A point β^0, iteration count k, α steplength:
- Repeat: Calculate $\nabla f(\beta^k) = X^T X\beta^k - X^T y$. Solve the sub-problem and set the solution to β^{k+1}

$$\beta^{k+1} = \min f(\beta^k) + \langle \nabla f(\beta^k), \beta - \beta^k \rangle + g(\beta) + \frac{1}{2\alpha}\|\beta - \beta^k\|_2^2.$$

Gradient descent-ish method to find Lasso solution

- This process is repeat until no improvement is made.
- In the k iteration the sub-problem is equivalent to:

$$\min_{\beta} \langle \nabla f(\beta^k), \beta - \beta^k \rangle + \lambda \|\beta\|_1 + \frac{1}{2\alpha} \|\beta - \beta^k\|_2^2$$

$$\min_{\beta} \langle \nabla f(\beta^k), \beta \rangle + \lambda \|\beta\|_1 + \frac{1}{2\alpha} (\langle \beta, \beta \rangle - 2 \langle \beta, \beta^k \rangle)$$

$$\min_{\beta} \lambda \|\beta\|_1 + \frac{1}{2\alpha} [\langle \beta, \beta \rangle - 2 \langle \beta, \beta^k - \alpha \nabla f(\beta^k) \rangle]$$

$$\min_{\beta} \lambda \|\beta\|_1 + \frac{1}{2\alpha} \|\beta - (\beta^k - \alpha \nabla f(\beta^k))\|_2^2.$$

Gradient descent-ish method to find Lasso solution

- Let $w = \beta^k - \alpha \nabla f(\beta^k)$, then the solution β^{k+1} is :

$$\beta_i^{k+1} = \begin{cases} w_i - \lambda\alpha & \text{if } w_i > \lambda\alpha \\ w_i + \lambda\alpha & \text{if } w_i < -\lambda\alpha \\ 0 & \text{if } otherwise \end{cases}$$

- This is called the soft-thresholding operator.
- This operator is used by almost every method for Lasso regression solution.

Gradient descent-ish method to find Lasso solution

- Initialization: A point β^0, iteration count k, α steplength:
- Repeat: Calculate $\nabla f(\beta^k) = X^T X \beta^k - X^T y$, $w = \beta^k - \alpha \nabla f(\beta^k)$ then the solution β^{k+1} is :

$$\beta_i^{k+1} = \begin{cases} w_i - \lambda\alpha & \text{if } w_i > \lambda\alpha \\ w_i + \lambda\alpha & \text{if } w_i < -\lambda\alpha \\ 0 & \text{if } \textit{otherwise} \end{cases}$$

- Until no improvement can be made.

Gradient descent-ish method to find Lasso solution

- The method in Matlab demonstration.

Selecting the tuning parameter

- Ridge and Lasso both have tuning parameters.
- Require a method for selecting a value for the tuning parameter λ.
- Split the whole data into 3 parts: training data, validating data, and testing data. Also choose a grid of values for λ ranging from small values to large.
- Example: $\textit{sequence} = -5 : 10/30 : 5$, $\lambda = e^{\textit{sequence}}$.
- Build models using training data and the values of λ above. Calculate MSE on the validating data. Choose the model with corresponding value of λ that has smallest validating MSE. Use the model on testing data.

Next class

- Resampling methods.
- Matlab built-in function for ridge and lasso.
- Examples with some real data sets using ridge and lasso.