Lecture 10: LASSO

Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

University of Washington

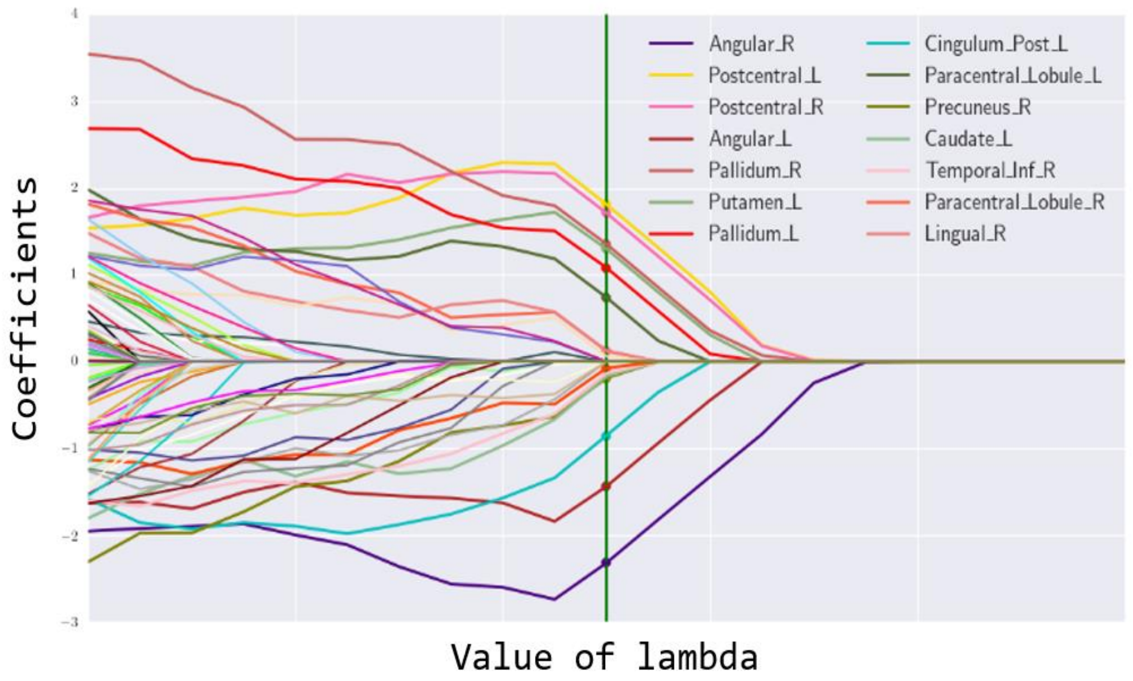
Feature selection in linear regression model

- LASSO was used to sparsify the linear regression model and allowed the regression model to select significant predictors automatically.
- The formulation of LASSO is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\},$$

- where $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ is the measurement vector of the response, $\mathbf{X} \in \mathbb{R}^{N\times p}$ is the data matrix of the N measurement vectors of the p predictors, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is the regression coefficient vector.
- Here, $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p}|\beta_i|$.

The path solution trajectory of LASSO

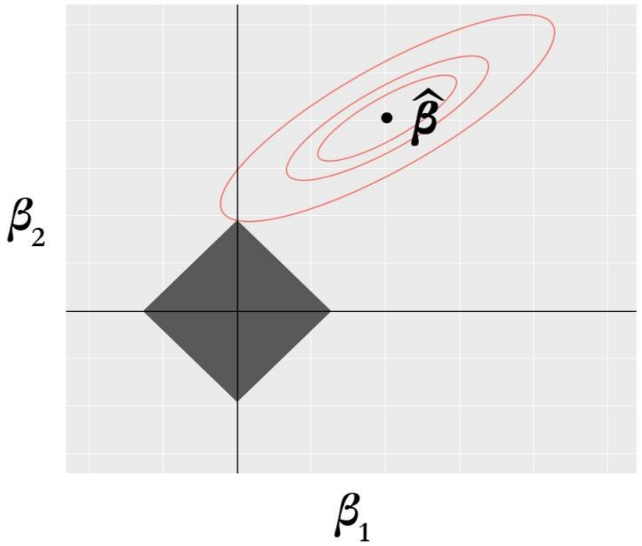


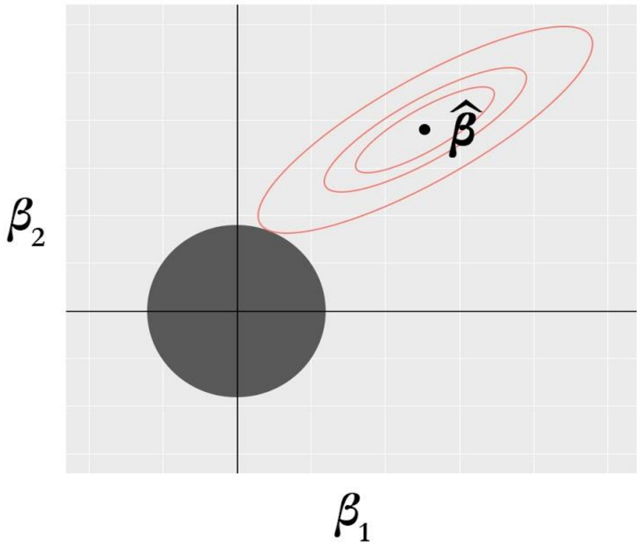
Why L1 norm?

- LASSO versus Ridge regression.
- The formulation of Ridge regression is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2\},$$

- where $\|\boldsymbol{\beta}\|_2 = \sum_{i=1}^{p}|\beta_i|^2$ is called the L_2 norm.





The shooting algorithm

- Let's first consider a simple case where there is only one predictor. Then, the objective function becomes

$$L(\beta) = \|\boldsymbol{y} - \mathbf{X}\beta\|_2^2 + \lambda|\beta|.$$

- To find the optimal solution, we can solve the equation as

$$\frac{\partial L(\beta)}{\partial \beta} = 0.$$

- The complication is the L1-norm term, $|\beta|$, which has no gradient when $\beta = 0$.

The shooting algorithm (cont'd)

We can discuss different scenarios and identify the solutions.

- If $\beta > 0$, then $\frac{\partial L(\beta)}{\partial \beta} = 2\beta - 2\mathbf{X}^T\boldsymbol{y} + \lambda$. Thus, $\frac{\partial L(\beta)}{\partial \beta} = 0$ will lead to the solution that $\beta = \frac{(2\mathbf{X}^T\boldsymbol{y} - \lambda)}{2}$. But if $2\mathbf{X}^T\boldsymbol{y} - \lambda < 0$, this will result in a contradiction, and thereby, $\beta = 0$.
- If $\beta < 0$, then $\frac{\partial L(\beta)}{\partial \beta} = 2\beta - 2\mathbf{X}^T\boldsymbol{y} - \lambda$. Similarly as above, we can conclude that $\beta = \frac{(2\mathbf{X}^T\boldsymbol{y} + \lambda)}{2}$. But if $2\mathbf{X}^T\boldsymbol{y} + \lambda > 0$, this will result in a contradiction, and thereby, $\beta = 0$.
- If $\beta = 0$, then we have had the solution and no longer need the calculate the gradient.

The shooting algorithm (cont'd)

- In summary, we can derive the solution of β as

$$\hat{\beta} = \begin{cases} \frac{\left(2\mathbf{X}^T\boldsymbol{y}-\lambda\right)}{2}, & if 2\mathbf{X}^T\boldsymbol{y} - \lambda > 0 \\ \frac{\left(2\mathbf{X}^T\boldsymbol{y}+\lambda\right)}{2}, & if 2\mathbf{X}^T\boldsymbol{y} + \lambda < 0 \\ 0, & if\ \lambda \geq |2\mathbf{X}^T\boldsymbol{y}| \end{cases}.$$

Generalize it to more general settings

- Let's contemplate an iterative structure that updates each β_j at a time when fixing all the other parameters as their latest values
- Suppose that we are now at the tth iteration and we are trying to optimize for β_j, we can rewrite the general optimization problem's objective function as a function of β_j

$$L(\beta_j) = \left\| \boldsymbol{y} - \sum_{k \neq j} \mathbf{X}_{(:,k)} \beta_k^{(t-1)} - \mathbf{X}_{(:,j)} \beta_j \right\|_2^2 + \lambda \sum_{k \neq j} \left| \beta_k^{(t-1)} \right| + \lambda |\beta_j|.$$

- Here, $\beta_k^{(t)}$ is the value of β_k in the tth iteration. The objective function above can be simplified as

$$L(\beta_j) = \left\| \boldsymbol{y} - \mathbf{X}_{(:,j)} \beta_j \right\|_2^2 + \lambda |\beta_j|,$$

- which just resembles the structure as the one-predictor special case we discussed. Thus, we can readily derive that

$$\hat{\beta}_j^{(t)} = \begin{cases} q_j - \lambda/2\,, & if\, q_j - \lambda/2 > 0 \\ q_j + \lambda/2\,, & if\, q_j + \lambda/2 < 0 \\ 0, & if\ \lambda \geq |2q_j| \end{cases},$$

- where $q_j = \mathbf{X}_{(:,j)}^T \left(\boldsymbol{y} - \sum_{k \neq j} \mathbf{X}_{(:,k)} \beta_k^{(t-1)} \right)$.

A simple example

- The dataset of Y is actually randomly sampled from the true model,

$$Y = 0.8X_1 + \varepsilon,\ \text{where } \varepsilon \sim N(0,0.5).$$

- The objective function of LASSO on this case is

$$\sum_{n=1}^{N}\left[y_n - \left(\beta_1 x_{n,1} + \beta_2 x_{n,2}\right)\right]^2 + \lambda(|\beta_1| + |\beta_2|).$$

X_1	X_2	Y
-0.707	0	-0.77
0	0.707	-0.33
0.707	-0.707	0.62

Note that, here, for simplicity, we don't need to include the offset parameter β_0 in the model as the predictors are standardized with mean as zero.

A simple example (cont'd)

- Suppose that we choose $\lambda = 0.88$. First, we initiate the parameters as $\hat{\beta}_1^{(0)} = 0$ and $\hat{\beta}_1^{(0)} = 0$.
- In the first iteration, we aim to update $\hat{\beta}_1$. We can obtain that

$$\boldsymbol{y} - \mathbf{X}_{(:,2)}\hat{\beta}_2^{(0)} = \begin{bmatrix} -0.71 \\ -1.037 \\ 1.327 \end{bmatrix}.$$

- Thus, $q_1 = \mathbf{X}_{(:,1)}^T \left(\boldsymbol{y} - \mathbf{X}_{(:,2)}\hat{\beta}_2^{(0)} \right) = 1.44.$
- As $q_1 - \lambda/2 = 1 > 0$, we know that $\hat{\beta}_1^{(1)} = q_1 - \lambda/2 = 1.$
- Similarly, we can update $\hat{\beta}_2$. We can obtain that

$$\boldsymbol{y} - \mathbf{X}_{(:,1)}\hat{\beta}_1^{(0)} = \begin{bmatrix} -1.477 \\ -0.33 \\ -0.087 \end{bmatrix}.$$

- Thus, $q_2 = \mathbf{X}_{(:,1)}^T \left(\boldsymbol{y} - \mathbf{X}_{(:,1)}\hat{\beta}_1^{(0)} \right) = -0.178.$
- As $\lambda \geq |2q_2|$, we know that $\hat{\beta}_2^{(1)} = 0.$

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets