Lecture 10: Penalized Regression

Naijia Liu

Feb. 22 2024

Recap

- Lasso can help us to select out variables.
- Lasso can shrink coefficients to optimize prediction performance.
- Lasso has a different loss function than OLS:

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i}^{N} \left(Y_i - \widetilde{\beta} X_i \right)^2 + \lambda |\widetilde{\beta}|$$

Why do we want to use Lasso?

- Multi-collinearity

  Education and Income, GDP and GDP per capita

Why do we want to use Lasso?

- Multi-collinearity

  Education and Income, GDP and GDP per capita

- Sparse data (text)

Document	meaning	of	life	$\cdots$
1	1	1	1	$\cdots$
2	0	0	0	$\cdots$
$\cdots$	0	0	0	$\cdots$

Why do we want to use Lasso?

- Multi-collinearity

  Education and Income, GDP and GDP per capita

- Sparse data (text)

Document	meaning	of	life	$\cdots$
1	1	1	1	$\cdots$
2	0	0	0	$\cdots$
$\cdots$	0	0	0	$\cdots$

- Optimize prediction power.

Lasso and its lambda

$$\hat{\beta} = \left(\beta_{\text{OLS}} - \text{sgn}_{\beta_{\text{OLS}}} \frac{\lambda}{B}\right) \mathbf{1}\left(|\beta_{\text{OLS}}| > \frac{\lambda}{\beta}\right)$$

- Shrink to zero if β_{OLS} is smaller than a threshold.
- Let's see how the magic λ works!

Toy Example

$$Y = 5X + 0$$

Toy Example

$$Y = 5X + 0$$

Unit	Y	X
1	5	1
2	10	2
3	15	3
4	20	4
5	25	5

Toy Example

$$Y = 5X + 0$$

Unit	Y	X
1	5	1
2	10	2
3	15	3
4	20	4
5	25	5

$$\begin{aligned}
\beta_{\text{OLS}} &= \frac{\sum_i X_i Y_i}{\sum_i X_i X_i} \\
&= \frac{5 \cdot 1 + 10 \cdot 2 + 15 \cdot 3 + 20 \cdot 4 + 25 \cdot 5}{1 + 4 + 9 + 16 + 25} \\
&= \frac{275}{55} \\
&= 5
\end{aligned}$$

OLS gets us unbiased estimate of β!

Toy Example

$$\lambda = 550$$

$$Y = 5X + 0$$

Unit	Y	X
1	5	1
2	10	2
3	15	3
4	20	4
5	25	5

Toy Example

$$\lambda = 550$$

$$Y = 5X + 0$$

This makes $\frac{\lambda}{\sum_i^N X_i X_i} = \frac{550}{55} = 10$

Unit	Y	X
1	5	1
2	10	2
3	15	3
4	20	4
5	25	5

Toy Example

$$\lambda = 550$$

$$Y = 5X + 0$$

Unit	Y	X
1	5	1
2	10	2
3	15	3
4	20	4
5	25	5

This makes $\frac{\lambda}{\sum_i^N X_i X_i} = \frac{550}{55} = 10$

$$\begin{aligned}\beta_{\text{Lasso}} &= \left(\frac{\sum_i X_i Y_i}{\sum_i X_i X_i} - \frac{\lambda}{\sum_i X_i X_i}\right) \cdot \mathbf{1}\left(|\beta_{\text{OLS}}| > 10\right) \\ &= (5 - 10) * 0 \\ &= 0\end{aligned}$$

ah oh, shrank.

Toy Example

Let's check if this really minimizes the loss function for LASSO.

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i}^{N} \Big(Y_i - \widetilde{\beta} X_i \Big)^2 + \lambda |\widetilde{\beta}|$$

- When $\hat{\beta} = 0$, the loss function is:

  $\frac{1}{2}\left((5-0)^2 + (10-0)^2 + (15-0)^2 + (20-0)^2 + (25-0)^2\right) + 0 = 687.5$

Toy Example

Let's check if this really minimizes the loss function for LASSO.

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i}^{N} \left(Y_i - \widetilde{\beta} X_i \right)^2 + \lambda |\widetilde{\beta}|$$

- When $\hat{\beta} = 0$, the loss function is:

  $\frac{1}{2}\left((5-0)^2 + (10-0)^2 + (15-0)^2 + (20-0)^2 + (25-0)^2\right) + 0 = 687.5$

- When $\hat{\beta} = 5$, the loss function is:

$$\frac{1}{2}(0) + 550 * 5 = 2750$$

Toy Example

- What if λ is 55 instead?
- Can you calculate $\hat{\beta}$ for both OLS and Lasso?

  $\hat{\beta}$ for OLS should not change!
- Can you check if $\hat{\beta}_{\text{lasso}}$ really minimizes the loss function in that case?