NBA 4920/6921 Lecture 14

Shrinkage Methods: Lasso Regression

Murat Unal

Johnson Graduate School of Management

10/19/2021

Agenda

Recap: Ridge regression

Lasso

Application in R

Shrinkage methods

- Fit a model that contain all p predictors
- At the same time constrain or **regularize** the coefficient etimates
- Regularization **shrinks** the coefficients towards zero

1. Ridge regression
2. Lasso
3. Elastic net

Ridge regression

Recall that we estimate coefficients $\beta_0, \beta_1, \cdots, \beta_p$ by minimizing RSS

$$\min_{\hat{\beta}} \text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

Ridge regression makes a small change by adding a shrinkage penalty, the sum of squared coefficients ($\lambda \sum_j \beta_j^2$)

$$\min_{\hat{\beta}^R} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2 = \min_{\hat{\beta}} \text{RSS} + \lambda \sum_j \beta_j^2$$

Ridge regression

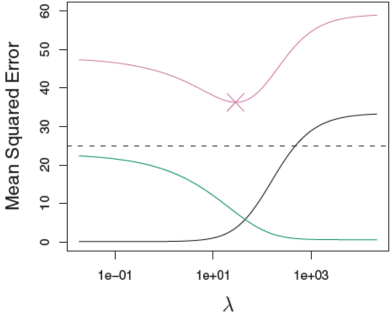
$$\min_{\hat{\beta^R}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2$$

$\lambda >= 0$ is a tuning parameter that determines the magnitude of the penalty

$\lambda = 0 \rightsquigarrow$ no penalty $\rightsquigarrow$ back to least squares

Ridge regression

- The optimal penalty balances reduced variance with increased bias. $p = 45, n = 50$
- OLS, $\lambda = 0$, will have low bias but high variance
- Ridge regression works best in situations where the least squares estimates have high variance



Source: ISL

Ridge regression

- While the shrinkage penalty has the effect of shrinking the estimates towards zero, it never forces them to be zero.
- As a result, we can end up with many tiny coefficients.
- This also means that Ridge regression can not be used for variable/feature/subset selection
- Enter the Lasso!

Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) replaces Ridge's squared coefficients with absolute values

$$\min_{\hat{\beta^L}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_j |\beta_j| = \min_{\hat{\beta^L}} \text{RSS} + \lambda \sum_j |\beta_j|$$

Lasso

$$\min_{\hat{\beta^L}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_j |\beta_j| = \min_{\hat{\beta^L}} \mathsf{RSS} + \lambda \sum_j |\beta_j|$$

$\lambda >= 0$ is a tuning parameter that determines the magnitude of the penalty

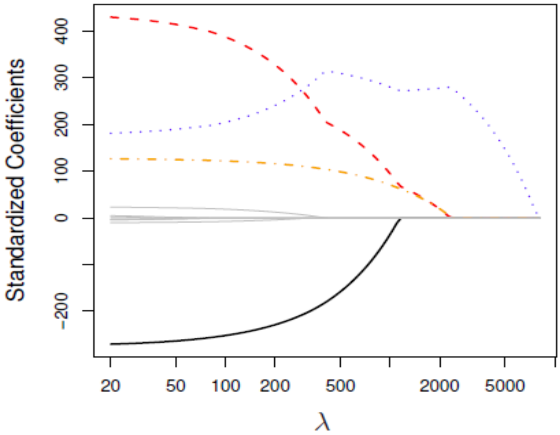
$\lambda = 0 \rightsquigarrow$ no penalty $\rightsquigarrow$ back to least squares

Lasso

- Similar to Ridge regression, the Lasso shrinks the coeffcient estimates towards zero
- However, the shrinkage penalty now has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- As such, the Lasso performs feature/subset selection
- Each value of λ results in different coefficient estimates, thus selecting a good value is critical

Lasso

The standardized lasso coefficients on the Credit data set are shown as a function of λ



Source: ISL

Selecting the tuning parameter λ

- We perform cross-validation to find the optimum λ
- Start by defining a grid of λ values, and compute the cross-validation error rate for each value of λ
- Select the tuning parameter value for which the cross-validation error is smallest
- Finally, the model is fit again using all of the available observations and the selected value of the tuning parameter.

Ridge or Lasso?

Ridge

- Shrinks $\hat{\beta}_j$ towards 0
- Many tiny $\hat{\beta}_j$
- Can not select features
- Harder to interpret
- Better to use when all $\hat{\beta}_j \neq 0$

Lasso

- Shrinks $\hat{\beta}_j$ towards 0
- Many $\hat{\beta}_j = 0$
- Can select features
- Easier to interpret
- Assumes some $\hat{\beta}_j = 0$

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017)
An Introduction to Statistical Learning
Springer.
https://www.statlearning.com/

Ed Rubin (2020)
Economics 524 (424): Prediction and Machine-Learning in Econometrics
Univ, of Oregon.