NBA 4920/6921 Lecture 15

Shrinkage Methods: Elastic Net

Murat Unal

Johnson Graduate School of Management

10/21/2021

Agenda

Recap: Ridge & lasso regression

Elastic net

Application in R

Shrinkage methods

- Fit a model that contain all p predictors
- At the same time constrain or **regularize** the coefficient estimates
- 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

- 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

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

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

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 $\hat{\beta}_j = 0$

Can't we use both?

Elastic net combines Ridge and Lasso

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

Which increases the tuning parameters to two: α,λ

With $\alpha = 0$ we're back to Ridge

With $\alpha = 1$ we're back to Lasso

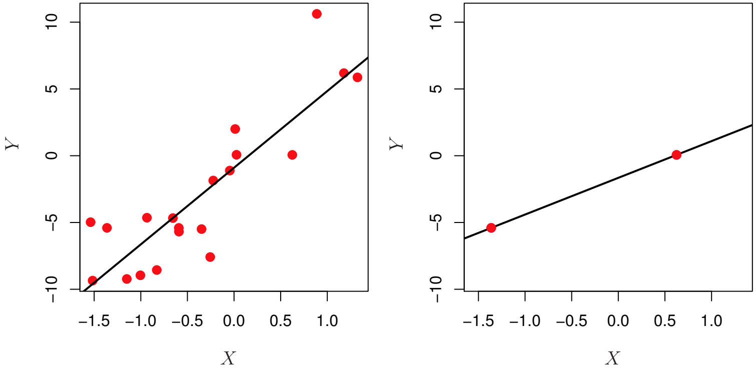
We need to tune both α and λ using cross-validation

Considerations in high dimensional settings

- When the number of features p is as large as, or larger than, the number of observations n, OLS, should not be used.
- Regardless of whether or not there truly is a relationship between the features and the response, OLS will yield a set of coefficient estimates that result in a perfect fit to the data, such that the residuals are zero.

Considerations in high dimensional settings

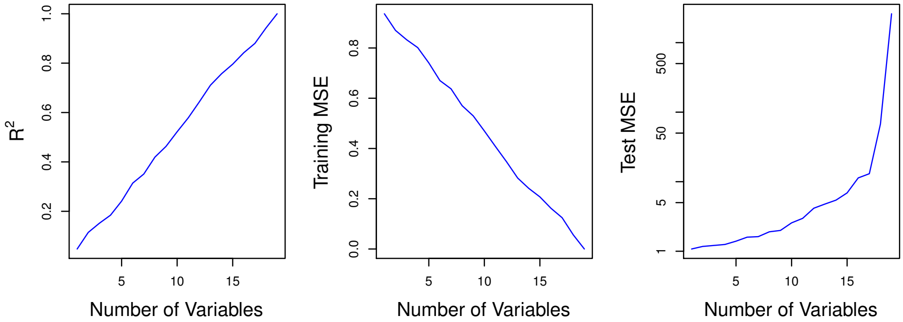
Left: Least squares regression in the low-dimensional setting.
Right: Least squares regression with n = 2 observations and two parameters to be estimated (an intercept and a coefficient).



Source: ISL

Considerations in high dimensional settings

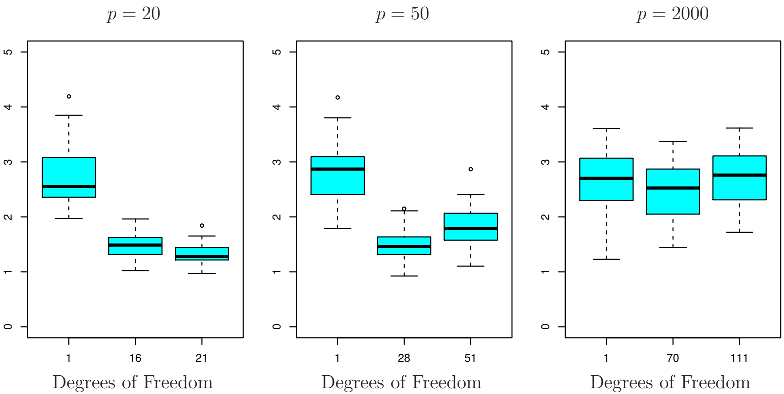
On a simulated example with n = 20 training observations, features that are completely unrelated to the outcome are added to the model. Left: The R2 increases to 1 as more features are included. Center: The training set MSE decreases to 0 as more features are included. Right: The test set MSE increases as more features are included



Source: ISL

Considerations in high dimensional settings

The lasso was performed with $n = 100$ and three values of p. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ.



Source: ISL

Considerations in high dimensional settings

1. Regularization or shrinkage plays a key role in high-dimensional problems
2. Appropriate tuning parameter selection is crucial for good predictive performance
3. The test error tends to increase as the dimensionality of the problem increases, unless the additional features are truly associated with the response.

Considerations in high dimensional settings

- In general, adding additional signal features that are truly associated with the response will improve the fitted model, in the sense of leading to a reduction in test set error.
- However, adding noise features that are not truly associated with the response will lead to a deterioration in the fitted model, and consequently an increased test set error.

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.