Week 5: Shrinkage Methods and Model Selection

AI 539: Machine Learning for Non-Majors

Alireza Aghasi

Oregon State University

What are $\boxed{\text{Shrinkage}}$ Methods and Why Useful?

Shrinking Coefficient $\hat{\beta}$

You would probably hear **Ridge Regression** and **LASSO** quite often

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors
- As an alternative, we can fit a model containing all $\overset{\beta}{\text{p predictors}}$ using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly **reduce the model variance**

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \cdots, \beta_p$ using the values that minimize

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$RSS_{Ridge} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \boxed{+ \lambda \sum_{j=1}^{p} \beta_j^2} = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Here, λ is a tuning parameter

Ridge Regression

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small

less is better

- However, the second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, called a shrinkage penalty, encourages solutions that are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates (trade off between bias and variance)
- Selecting a good value for λ is critical; often cross-validation is used for this

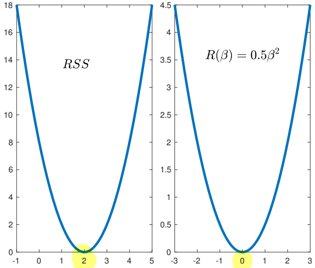
Using CV to select λ

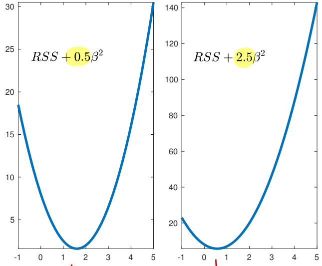
λ also controls the variance

Effect of Increasing λ on the β

- The figure below shows how increasing the Ridge penalty pushes the minimizers of the mixed RSS objective to zero

$y = \beta x + \varepsilon$ $\lambda\beta^2$





not 2 or 0

closer to 0 compared to $\lambda = 0.5$

Adding $\lambda\beta^2$ is pushing the stationary point to 0, the larger λ the closer to 0

Shrinkage Example

- Previously from the homework assignments you remember that the least squares solution to fit data point $(x_1, y_1), \cdots, (x_n, y_n)$ was obtained via the minimization:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta x_i)^2 \quad \therefore \quad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

- We can show that if we run the Ridge regression problem

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta x_i)^2 \boxed{+ \lambda \beta^2}$$

the new estimate becomes

$$\hat{\beta}^R = \frac{\sum_{i=1}^{n} x_i y_i}{\lambda + \sum_{i=1}^{n} x_i^2}$$

- Note how increasing λ pushes $\hat{\beta}^R$ towards zero

$$RR(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda \beta^2 , \quad \begin{cases} Var(\varepsilon) = \sigma^2 \\ Var(y_i) = \sigma^2 \end{cases}$$

to minimize $\beta \rightarrow \dfrac{dRR}{d\beta} = 2 \sum_{i=1}^{n} (\beta x_i - y_i) + 2\lambda\beta = 0$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\lambda + \sum_{i=1}^{n} x_i^2}$$

$$Var(\hat{\beta}) = Var\left(\frac{\sum_{i=1}^{n} x_i y_i}{\lambda + \sum_{i=1}^{n} x_i^2}\right) = \frac{1}{\left(\lambda + \sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} (x_i)^2 \cdot \underline{\underline{Var(y_i)}}$$

$$= \frac{\sum_{i=1}^{n} (x_i)^2}{\left(\lambda + \sum_{i=1}^{n} x_i^2\right)^2} \cdot \sigma^2$$

In Class Exercise

- For the simple regression problem of fitting $(x_1, y_1), \cdots, (x_n, y_n)$, to the model $y = \beta_0 + \beta_1 x$ show that the least-squares estimates for the Ridge regularized objective

$$\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 + \lambda(\beta_0^2 + \beta_1^2)$$

are

$$\hat{\beta}_1^R = \frac{\sum_{i=1}^n x_i y_i - \frac{n^2}{n+\lambda}\bar{x}\bar{y}}{\lambda + \sum_{i=1}^n x_i^2 + \frac{n^2}{n+\lambda}\bar{x}^2}, \quad \hat{\beta}_0^R = \frac{1}{n+\lambda}\left(\sum_{i=1}^n y_i - \hat{\beta}_1^R \sum_{i=1}^n x_i\right)$$

What Happens in Multiple Regression?

- In this case we previously had

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta)$$

which led to

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

- In the case of regularized problem ($\|.\|$ denotes the L-2 norm)

$$(\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda \|\beta\|^2$$

we will have

$$\hat{\beta}^R = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

where $\boldsymbol{I}$ is the identity matrix

$$\lambda = \infty \Rightarrow \beta = 0$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad , \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad , \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\Rightarrow \quad \vec{y} = \vec{\beta}^T \cdot \vec{X}$$

If multiple features, X is a matrix of data with n samples as the rows and p features as the columns.

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ \vdots & & & \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{bmatrix} \begin{array}{l} \rightarrow \text{Sample } 1 \\ \\ \rightarrow \text{Sample } n \end{array}$$

Norm Review

$$\|z\| = \sqrt{z_1^2 + z_2^2 + \dots z_p^2}$$

$$\|z\|^2 = z_1^2 + z_2^2 + \dots z_p^2$$

$$= z^T \cdot z$$

$$RR = (y - \beta X)^T (y - \beta X) + \lambda \|\beta\|^2$$

$$\frac{dRR}{d\beta} = \frac{d}{d\beta}\left[y^T y - 2 y^T X \beta + \beta^T X^T X \beta + \lambda \beta^T \beta \right]$$

$$= 0 - 2X^T y + 2X^T X \beta + 2\lambda\beta = 0 \Rightarrow (X^T X + \lambda I)\beta = X^T y$$

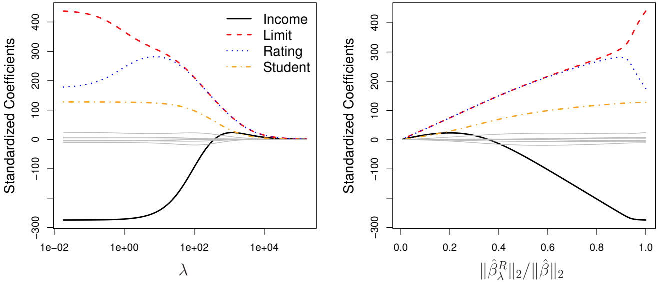
$$\Rightarrow \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$$

λ is very large $\rightarrow$ $(X^T X + \lambda I)^{-1} \simeq \frac{1}{\lambda} I$

$$\Rightarrow \hat{\beta} \simeq \frac{1}{\lambda} X^T y$$

Credit Data Example

- Left: each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we display $\|\hat{\beta}^R\|/\|\hat{\beta}\|$ (how much **shrinkage** happens by increasing λ)



β will never reach 0 since λ won't be ∞

Scaling of the Predictors

- In the standard least-squares if we scale a feature value by c, the corresponding coefficient scales by c^{-1}
- However when we have the Ridge regularized objective, this is no more the case
- To see a consistent behavior, for the Ridge regularized problem we often work with standardized features:

$$\tilde{x}_{i,j} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Basic RSS minimization

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \longrightarrow \beta_0, \beta_1, \cdots, \beta_p$$

$$y = \beta_0 + \beta_1 (2X_1) + \beta_2 X_2 + \cdots + \beta_p X_p \longrightarrow \beta_0, \frac{1}{2}\beta_1, \beta_2, \cdots, \beta_p$$

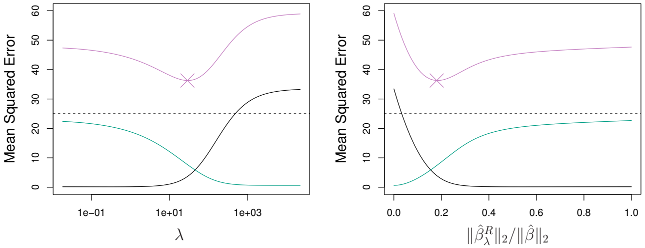
Ridge

$$y = \beta_0 + \beta_1 (2X_1) + \beta_2 X_2 + \cdots + \beta_p X_p \longrightarrow \text{Something different.}$$

Instead of dropping features, change the form of RSS to get better model.

Bias-Variance Trade-Off

- A toy example: squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R\|/\|\hat{\beta}\|$. The horizontal dashed lines indicate the minimum possible MSE (**the standard least squares,** $\lambda = 0$ **in nowhere close**). The purple crosses indicate smallest ridge regression model MSE values



- Remember (test error = bias + variance + noise variance)

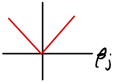
LASSO (Least Absolute Selection and Shrinkage Operator)

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model (none of the model coefficients become explicitly zero)
- The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients minimize the quantity

$$RSS_{LASSO} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \underline{RSS + \lambda \|\boldsymbol{\beta}\|_1}$$

Turns out to be a Convex Function

- We call $\|\boldsymbol{\beta}\|_1$ the L-1 norm of $\boldsymbol{\beta}$



LASSO

LASSO also shrinks β_i to 0

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero

$\|\beta_j\|$

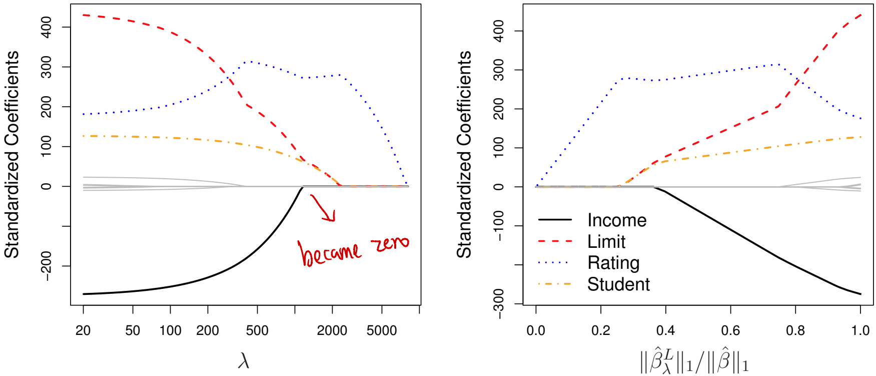
- However, in the case of the lasso, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large
- Technically the lasso performs the training and variable selection together
- We say that the lasso yields sparse models — that is, models that involve only a subset of the variables
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice

$$\begin{cases} y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \\ \text{Solve} \quad \min \quad RSS + \lambda \sum |\beta_j| \end{cases}$$

$\longrightarrow$ Some of $\hat{\beta}_j$ are zero, i.g. $\beta_3 = 0$, $\beta_{17} = 0$ ...

$\longrightarrow$ Which means drop some features !!

Example: Credit Data



Why LASSO promotes Sparsity?

- From a convex optimization perspective, one can show that the lasso and ridge regression coefficient estimates solve the problems

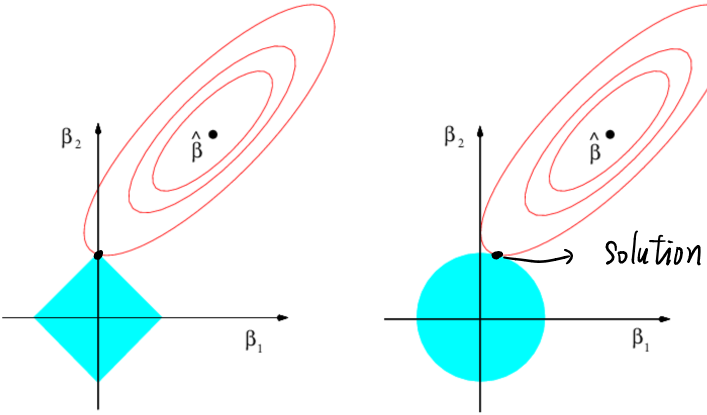
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \textit{subject to} \quad \underline{\|\boldsymbol{\beta}\|_1 \leq \tau}$$

and

Add constraints ?

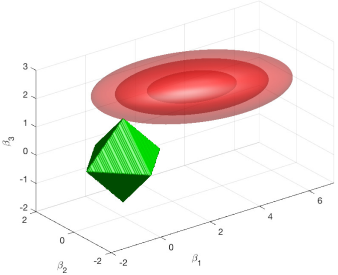
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \textit{subject to} \quad \underline{\|\boldsymbol{\beta}\|_2 \leq \tau'}$$

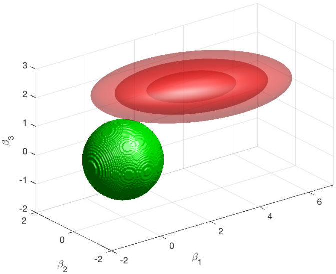
The Geometry of the Two Problems



The Geometry of the Two Problems In Higher Dimension

See the MATLAB code attached:



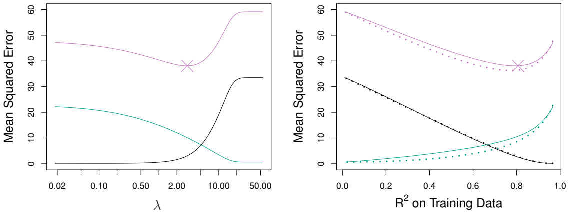


LASSO or Ridge?

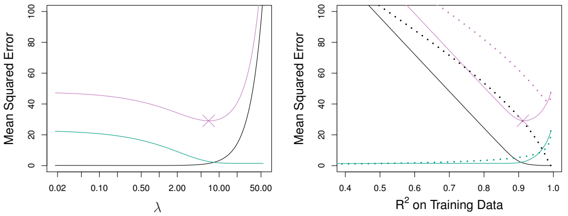
- Neither ridge regression nor the lasso will universally dominate the other
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors
- However, the number of predictors that is related to the response is never known a priori for real data sets
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set

Example

Previous example using all features (LASSO:Solid, Ridge: Dashed):



Using only two features:

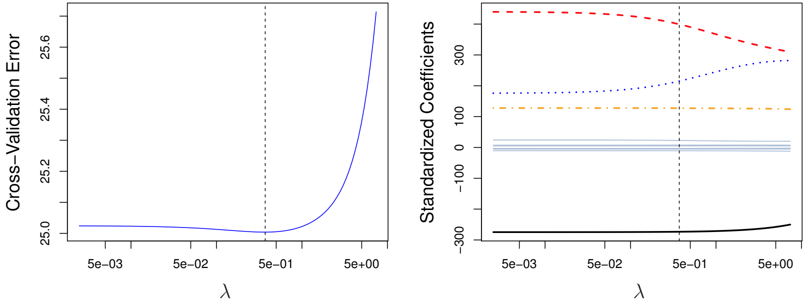


How to Determine λ?

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is the best
- That is, we require a method selecting a value for the tuning parameter λ
- Cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ
- We then select the tuning parameter value for which the cross-validation error is the smallest
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter

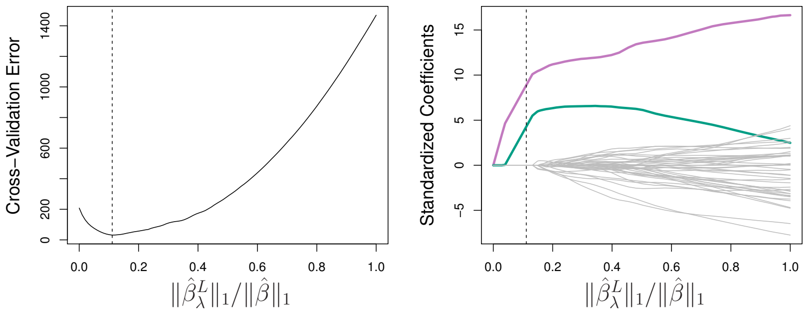
CV and Ridge Example

Determining λ via cross validation for the Ridge problem (credit data)



CV and LASSO Example

Determining λ via cross validation for the LASSO problem (simulated data)



Questions?

References

- https://www.alsharif.info/iom530, 2013.
- J. Friedman, T. Hastie, and R. Tibshirani.
  The elements of statistical learning.
  Springer series in statistics, 2nd edition, 2009.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/cv_boot.pdf, 2013.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/model_selection.pdf, 2013.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  An introduction to statistical learning: with applications in R, volume 112.
  Springer, 2013.