STA 314: Statistical Methods for Machine Learning I

Lecture 5 - More on regularized linear regression

Xin Bing

Department of Statistical Sciences
University of Toronto

Review: why consider alternatives to the OLS estimator?

Recall the linear model is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon.$$

Alternative fitting procedures to OLS could yield **better prediction accuracy** and **model interpretability**.

- Prediction: OLS estimator has large variance when p is large. Especially, if $p > n$, then OLS estimator is not unique and its variance is very large.
- Interpretability: By removing irrelevant features – that is, by setting some coefficient estimates to zero – we can obtain a model that is more parsimonious hence more interpretable.

Review

- Best subset selection
  - Great! But computationally unaffordable (choose from 2^p models)!
- Stepwise subset selection
  - Forward stepwise selection
  - Backward stepwise selection
  - Computationally affordable, but greedy approaches
- Are there better alternatives?
  - Shrinkage Methods! In particular, the Lasso.

Magic of the Lasso

Why does the lasso, unlike ridge regression, yield coefficient estimates that have exact zero?

Another Formulation for Ridge Regression and Lasso

The lasso and ridge regression coefficient estimates solve the problems

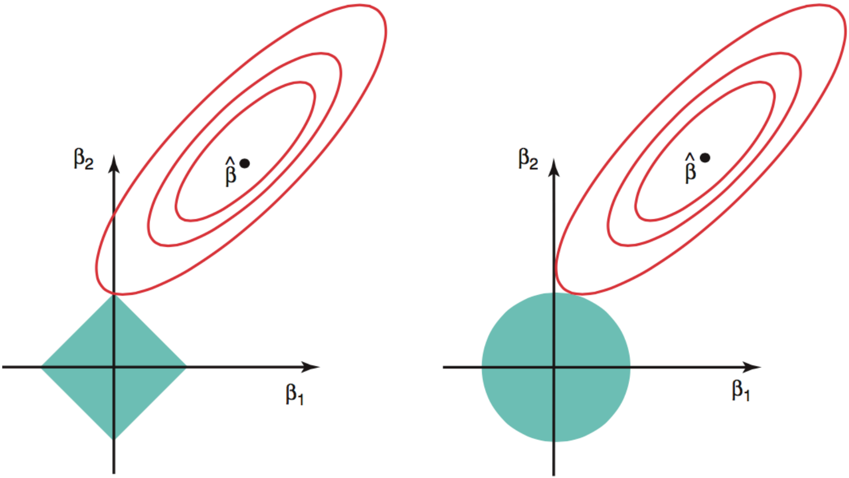
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Here $s \geq 0$ is some regularization parameter (connected with the original λ).

Understand why the Lasso yields zero estimates



The solid areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

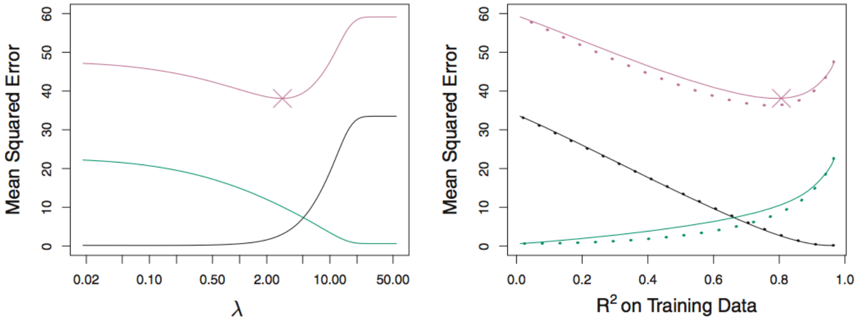
Lasso vs Ridge

- The ability of yielding a **sparse** model is a huge advantage of Lasso comparing to Ridge.

- A more sparse model means more interpretability!

- What about their prediction performance?

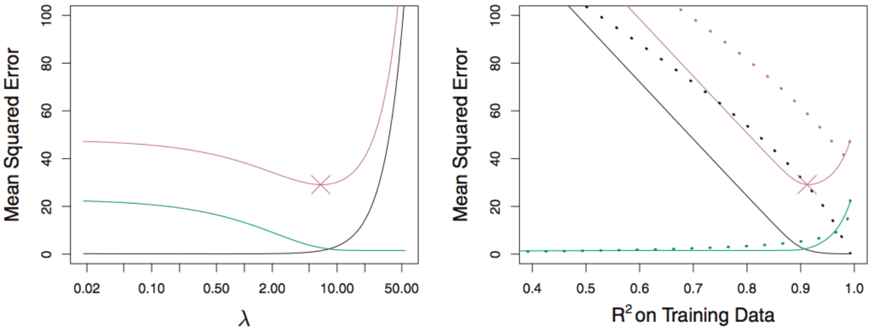
Comparing the MSE of Lasso and Ridge



Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.
Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.

- *When the true coefficients are non-sparse, ridge and lasso have the same bias but ridge has a smaller variance hence a smaller MSE.*

Another Case



- *When the true coefficients are sparse, Lasso outperforms ridge regression of having both a smaller bias and a smaller variance.*

Conclusions on Lasso relative to Ridge

- These two examples illustrate that 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 only related with a relatively small number of predictors.
- As the ridge regression, when the OLS estimates have excessively high variance, the lasso solution can yield a reduction in variance at the expense of a small increase in bias, and consequently can lead to more accurate predictions.
- Unlike ridge regression, the lasso performs variable selection, and hence yields models that are easier to interpret.

A simple example of the shrinkage effects of ridge and lasso

- Assume that $n = p$ and $\mathbf{X} = \mathbf{I}_n$. We force the intercept term $\beta_0 = 0$.
- In this way,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}.$$

- We assume

$$\mathbb{E}[\epsilon_j] = 0, \qquad \mathbb{E}[\epsilon_j^2] = \sigma^2, \qquad \forall j \in [p].$$

The OLS estimator

- The OLS approach is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p} (y_j - \beta_j)^2.$$

This gives the OLS estimator

$$\hat{\beta}_j = y_j, \qquad \forall j \in \{1, \ldots, p\}.$$

The ridge estimator

- The ridge regression is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

This leads to the ridge estimator

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda}, \qquad \forall j \in \{1, \ldots, p\}.$$

Since $\lambda \geq 0$, the magnitude of each estimated coefficient is shrinked toward 0.

The lasso estimator

- The lasso is to find $\beta_1, \ldots, \beta_p$ that minimize

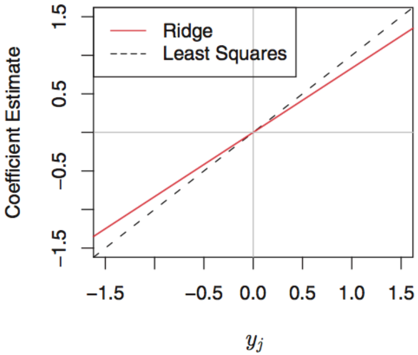
$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

This gives estimator

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases}$$

The estimated coefficients from Lasso are also shrinked. The above shrinkage is known as the **soft-thresholding**.

An illustrative figure



Bias and Variance of the OLS

Recall

$$y_j = \beta_j + \epsilon_j, \qquad \forall j \in [p].$$

For any $j \in [p]$, the OLS estimator $\hat{\beta}_j = y_j$ satisfies

- **Bias**:

$$\mathbb{E}[\hat{\beta}_j] = \mathbb{E}[y_j] = \mathbb{E}[\beta_j + \epsilon_j] = \beta_j$$

- **Variance**:

$$\mathrm{Var}(\hat{\beta}_j) = \mathrm{Var}(\epsilon_j) = \sigma^2$$

MSE of the OLS

- **Mean squared error** of the jth coefficient:

$$\mathbb{E}\left[\left(\hat{\beta}_j - \beta_j\right)^2\right] = \left(\mathbb{E}[\hat{\beta}_j] - \beta_j\right)^2 + \mathrm{Var}(\hat{\beta}_j) = \sigma^2$$

- **Mean squared error** of all p coefficients:

$$\mathbb{E}\left[\sum_{j=1}^{p}\left(\hat{\beta}_j - \beta_j\right)^2\right] = p\sigma^2.$$

Bias and Variance of the Ridge

Recall

$$y_j = \beta_j + \epsilon_j, \qquad \forall j \in [p].$$

For any $j \in [p]$, the ridge estimator with tuning parameter λ,

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda},$$

satisfies

- **Bias**:

$$\mathbb{E}[\hat{\beta}_j^R] = \mathbb{E}\left[\frac{y_j}{1+\lambda}\right] = \mathbb{E}\left[\frac{\beta_j + \epsilon_j}{1+\lambda}\right] = \frac{\beta_j}{1+\lambda}.$$

- **Variance**:

$$\mathrm{Var}(\hat{\beta}_j^R) = \mathrm{Var}\left(\frac{\epsilon_j}{1+\lambda}\right) = \frac{\sigma^2}{(1+\lambda)^2}$$

MSE of the ridge

- **Mean squared error** of the jth coefficient:

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\beta}_j^R - \beta_j\right)^2\right] &= \left(\mathbb{E}[\hat{\beta}_j^R] - \beta_j\right)^2 + \mathrm{Var}(\hat{\beta}_j^R) \\ &= \left(\frac{\beta_j}{1+\lambda} - \beta_j\right)^2 + \frac{\sigma^2}{(1+\lambda)^2} \\ &= \frac{\lambda^2\beta_j^2}{(1+\lambda)^2} + \frac{\sigma^2}{(1+\lambda)^2}.\end{aligned}$$

  Recall that $\mathbb{E}[(\hat{\beta}_j - \beta_j)^2] = \sigma^2$.

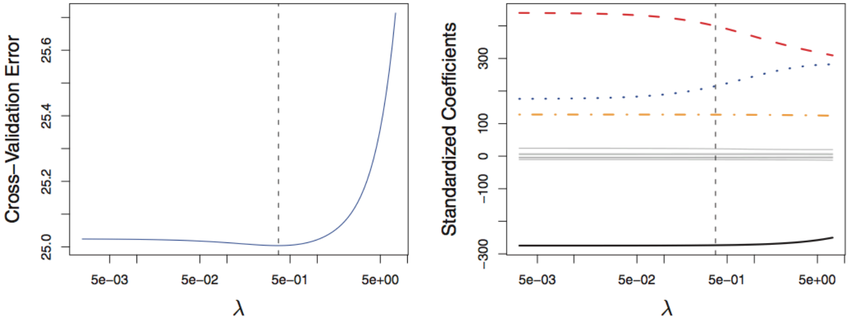
- **Mean squared error** of all p coefficients:

$$\mathbb{E}\left[\sum_{j=1}^{p}\left(\hat{\beta}_j^R - \beta_j\right)^2\right] = \frac{\lambda^2\sum_{j=1}^{p}\beta_j^2 + p\sigma^2}{(1+\lambda)^2}.$$

On selecting the tuning parameter

- Similar as the subset selection, for ridge and lasso, we require a systematic way of choosing the best model under a sequence of fitted models (from different choices of λ)
  - Equivalently, we require a method to select the optimal value of the tuning parameter λ.
- Cross-validation: we choose a grid of λ, and compute the cross-validation error rate for each value of λ.
- We then select the λ_* for which the cross-validation error is smallest.
- Finally, the model is re-fitted by using all of the available observations and the selected λ_*.

Credit Card Data Example



Cross-validation errors that result from applying ridge regression to the Credit data set for various choices of λ.

More choices of penalties

- There are many other penalties in addition to the ℓ_2 and ℓ_1 norms used by ridge and lasso.
  - the elastic net:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \left[(1-\alpha)||\boldsymbol{\beta}||_1 + \alpha||\boldsymbol{\beta}||_2 \right]$$

    for some tuning parameters $\lambda \geq 0$ and $\alpha \in [0, 1]$.
    - The ridge corresponds to $\alpha = 1$
    - The Lasso corresponds to $\alpha = 0$.

The group lasso

- If we suspect the model is nonlinear in X_1 or X_2, we can add quadratic terms, say

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \epsilon.$$

  The **group lasso** estimator minimizes

  $$RSS + \lambda\left(\sqrt{\beta_1^2 + \beta_2^2} + \sqrt{\beta_3^2 + \beta_4^2}\right).$$

  In this penalty, we view β_1 and β_2 (coefficient of X_1 and X_1^2) as if they belong to the same group. The group Lasso can shrink the parameters in the same group (both β_1 and β_2) exactly to 0 simultaneously.

- There are a lot more penalties out there ......

Regularization in more general settings

- The ridge and lasso regressions are not restricted to the linear models.

- The idea of penalization is generally applicable to almost all parametric models.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \underbrace{L(\boldsymbol{\beta}, \mathcal{D}^{train}) + Pen(\boldsymbol{\beta})}_{g(\boldsymbol{\beta};\mathcal{D}^{train})}.$$

  - OLS: $L(\boldsymbol{\beta}, \mathcal{D}^{train}) = ||\mathbf{y} - \mathbf{X}\beta||_2^2$, $Pen(\boldsymbol{\beta}) = 0$.
  - Ridge: $L(\boldsymbol{\beta}, \mathcal{D}^{train}) = ||\mathbf{y} - \mathbf{X}\beta||_2^2$, $Pen(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2^2$.
  - Lasso: $L(\boldsymbol{\beta}, \mathcal{D}^{train}) = ||\mathbf{y} - \mathbf{X}\beta||_2^2$, $Pen(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$.
  - In general,
    - L can be any loss function, i.e. negative likelihood, 0-1 loss.
    - Pen could be any penalty function.