Lecture 3: Ridge Regression

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Read: ISR Ch. 6.2 and ESLII Ch. 3.4

Outline

1 Ridge Regression

The next section would be . . . . . .

1 Ridge Regression

- When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance.
- A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin.

Ridge regression

- $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$; $X_i^T = (1, X_{i1}, \ldots, X_{ip})$.
- Linear regression OLS estimator minimizes the RSS:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- The ridge regression estimator chooses the $\boldsymbol{\beta}$ that minimizes

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

  for some tuning (regularization) parameter $t > 0$, to be determined separately.
- Term $\lambda\sum_{j=1}^{p}\beta_j^2$ is called a *shrinkage penalty*.
- λ controls the relative impact of both terms.
- Note the penalty is applied to $\beta_1, \ldots, \beta_p$, but not β_0 (why?)

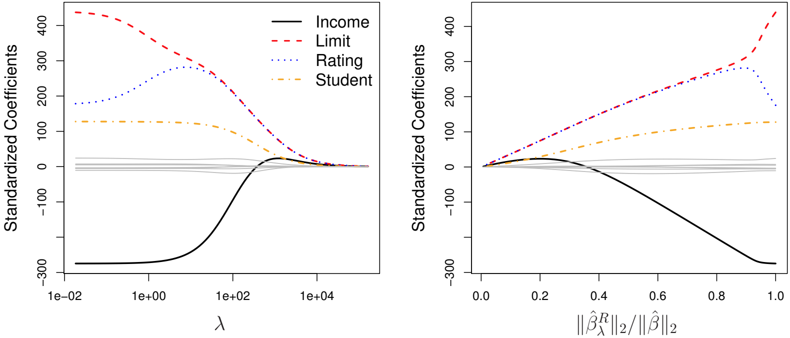
Constrained Least Square

- Alternatively, the ridge regression can be viewed as a constrained least square problem, that is to minimize

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

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t.$$

- These two formulas are equivalent because the former is the Lagrangian of the constrained form.
- There is a one-to-one correspondence between the parameters λ and t.


Standardized Coefficients
400
300
200
100
0
−100
−300
1e−02
1e+00
1e+02
1e+04
λ
Income
Limit
Rating
Student
Standardized Coefficients
400
300
200
100
0
−100
−300
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$

Practical Consideration

- The intercept term is never penalized. Why?
- The ridge regression solution is not invariant to scaling of the input variables.
- Hence one typically has to standardize the inputs so that they have same/similar scales.

Centering the data

- Objective function

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

- Easy to show (ESLII Ex. 3.5) that the ridge regression solution to the coefficients (except β_0) is the same as

$$\underset{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \bar{y} - \sum_{j=1}^{p} \beta_j(x_{ij} - \bar{x}_{\cdot j}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

  whereas the solution to β_0 is $(\bar{y} - \bar{x}^T \hat{\boldsymbol{\beta}}_{Ridge})$.

- There is no more intercept in the optimization.
- From now on, we assume that the data are always centered.

Closed form solution

- Assume that the data are centered. For centered data, the ridge regression line always goes through the origin.

$$\underset{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i^c - \sum_{j=1}^{p} \beta_j x_{ij}^c)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Matrix notation form: $\mathbf{X}$ has p columns (not $(p+1)$).

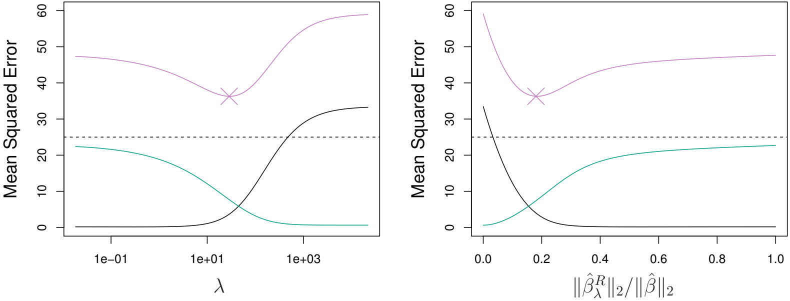
$$(\boldsymbol{y} - \mathbf{X}\beta)^T(\boldsymbol{y} - \mathbf{X}\beta) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

- Gradient equation: $-2\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\beta) + 2\lambda\boldsymbol{\beta} = 0$
- Closed form solution:

$$\begin{aligned}\mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= \lambda\boldsymbol{\beta} \\ \mathbf{X}^T\boldsymbol{y} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})\boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}}_{Ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\boldsymbol{y} \in \mathbb{R}^p\end{aligned}$$

Why Ridge Regression Improves Over OLS?

bias-variance trade-off!



Squared bias (black), variance (green), and test mean squared error (purple). Recall bias-variance decomposition.

Calculate the variance

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$
$$\text{Cov}(\hat{\boldsymbol{\beta}}_{Ridge}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\sigma^2$$

Compared to variance of the OLS:

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

It can be shown that

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) - \text{Cov}(\hat{\boldsymbol{\beta}}_{Ridge}) \text{ is positive definite}$$

so that ridge regression always reduces the uncertainty of the solution (i.e. variance of $\hat{\boldsymbol{\beta}}$).

Calculate the Bias

- As the ridge regression estimator is a linear estimator and outperforms OLS in terms of variance, it must be biased (otherwise it would refute the Gauss-Markov theorem).

$$
\begin{aligned}
\mathsf{E}(\hat{\boldsymbol{\beta}}_{Ridge}) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
\mathrm{Bias} &= \mathsf{E}(\hat{\boldsymbol{\beta}}_{Ridge}) - \boldsymbol{\beta} \\
&= [(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{X} - \mathbb{I}]\boldsymbol{\beta}
\end{aligned}
$$

The Existence Theorem

Theorem

There exists $\lambda > 0$ *such that* $MSE(\hat{\beta}_{Ridge}(\lambda)) < MSE(\hat{\beta}_{OLS})$

In other words, there must exist a ridge regression solution whose increases in squared bias is small than the reduction in variance.

Note that the RSS of ridge regression is bound to be greater than that of the OLS.

Effective degrees of freedom

- $\hat{\boldsymbol{y}}_{Ridge} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\boldsymbol{y}$
- $S_\lambda := \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T$ is similar to the hat matrix in OLS.
- The trace of S_λ is called the effective degrees of freedom.
- Recall that the degrees of freedom in OLS is $\text{Trace}(S_0) = p$ (assuming no intercept).
- Motivated by the effort we made to fit the model (or how well we have effectively done so.) See universal definition in (3.60) in ESLII.

A singular value decomposition analysis

- SVD: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{N\times p}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $\mathbf{D}$ diagonal (with diagonal elements $d_1 \geq d_2 \geq \ldots d_p \geq 0$).
- Hat matrix $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{U}\mathbf{U}^T = \sum_{j=1}^p \boldsymbol{u}_j\boldsymbol{u}_j^T$
- Hence $\hat{\boldsymbol{y}}_{OLS} = \mathbf{U}\mathbf{U}^T\boldsymbol{y} = \sum_{j=1}^p \boldsymbol{u}_j(\boldsymbol{u}_j^T\boldsymbol{y})$
- Note that $\boldsymbol{u}_j^T\boldsymbol{y}$ is the coordinate of $\boldsymbol{y}$ for axis $\boldsymbol{u}_j$ with respect to the orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$.
- In comparison, for the ridge regression,

$$\begin{aligned}\hat{\boldsymbol{y}}_{Ridge} = \mathbf{X}\hat{\beta}_{Ridge} &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\boldsymbol{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbb{I})^{-1}\mathbf{D}\mathbf{U}^T\boldsymbol{y} \\ &= \sum_{j=1}^p \boldsymbol{u}_j\left\{\frac{d_j^2}{d_j^2+\lambda}\boldsymbol{u}_j^T\boldsymbol{y}\right\}\end{aligned}$$

- Like linear regression, ridge regression computes the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\mathbf{U}$.
- It then shrinks these coordinates by the factors $\frac{d_j^2}{d_j^2+\lambda}$.
- Coordinates of basis vectors with smaller d_j^2 receives more shrinkage.
- Recall: $d_j^2 \sim$ variance of the jth principal component.
- The most shrinkage along PC direction with the smallest var.
- Using SVD, the effective degrees of freedom is

$$\text{Trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Tikhonov regularization

- Ridge regression uses the ℓ_2 norm of $\boldsymbol{\beta}$ as the regularizer (penalty function).
- A special case of the so-called Tikhonov regularization.
- Other form of regularizer will be explored later in the course.

Data Example

Lecture 3 R code

- Verify ridge regression closed-form solution.
- Draw the solution path.
- Why some coefficient becomes zero in the middle?

Summary

- Ridge regression has both ℓ_2 regularization and constrained OLS formulations.
- Closed-form solution is available.
- Lower variance, increase bias. May have better MSE than OLS.
- Shrink more along directions with smaller variance (or input variables that contribute to such directions).
- Some of the figures in this presentation are taken from ISLR with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.