Lecture 6 Model Selection II

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

January 26, 2021

Outline

Ridge Regression

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

$$\sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \beta_j^2,$$

penalty or regularizer

lambda>0

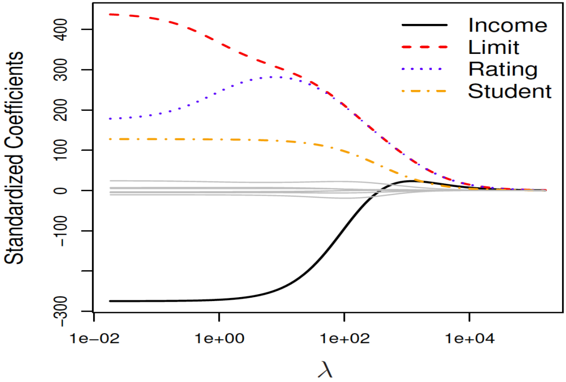
where λ is a tuning parameter, to be determined separately.

- The second term $\lambda \sum_j \beta_j^2$ called a shrinkage penalty, is small when $\beta_j,\ j \geq 1$ 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.
- Selecting a good value for λ is critical; cross-validation is used for this.

Different lambda will lead to different solutions to beta

Credit data example

lambda affects the model coefficients (and thus the complexity)



As λ increases, the coefficients are shrunken to zeros.

Scaling of predictors

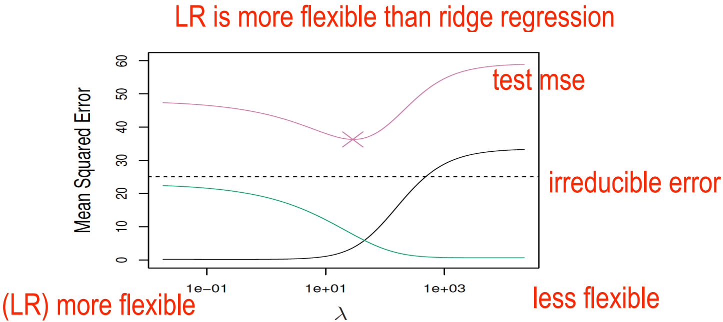
- The standard least squares coefficient estimates are scale equivariant: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the j-th predictor is scaled $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficient term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

predictor1: house sold price last time
predictor2: sqft
predictor3: monthly utility bill
response: value of the house

$$\tilde{x}_{ij} = x_{ij} \Big/ \sqrt{\sum_i (x_{ij} - \bar{x}_j)^2 / n}.$$

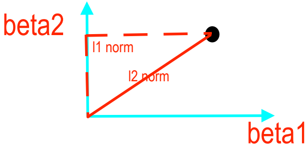
"std" of x_{ij} across all samples i

Credit data example



Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficient. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

The LASSO



- Ridge regression, unlike subset selection which will select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.
- The LASSO is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficient $\hat{\beta}^L$ minimize the quantity

street block distance

$$\text{minimize} \sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j|,$$

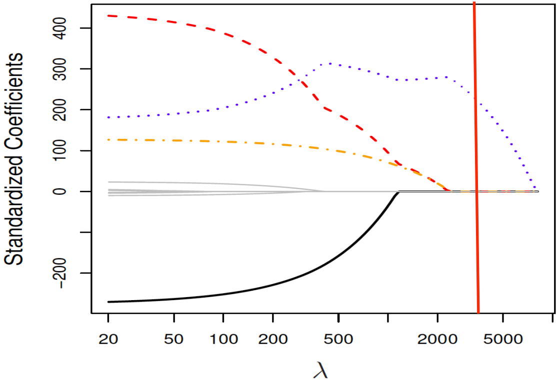
where λ is a tuning parameter.

- The LASSO uses l_1 penalty instead of l_2 (ridge regression).

The LASSO

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero as λ increases.
- However, in the case of the lasso, the l_1 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. Thus it performs variable selection.
- We say that the lasso yields sparse models — that is, models that involve only a subset of the variables.
- Selecting a good value for λ is critical; cross-validation is again used for this.

Credit data example



As λ increases, the coefficients are shrunken to exact zeros.

Ridge regression and the LASSO

- Why is it that the lasso, unlike ridge regression, results in some coefficient estimates being exactly zero?
- One can show that the lasso and ridge regression coefficient estimates solve the problems

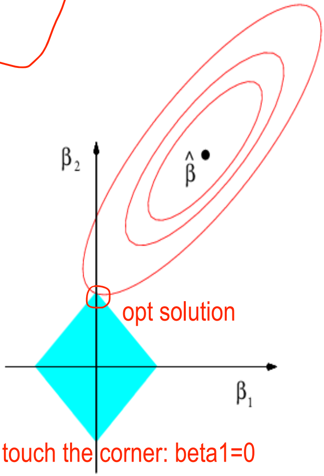
lagrangian theory

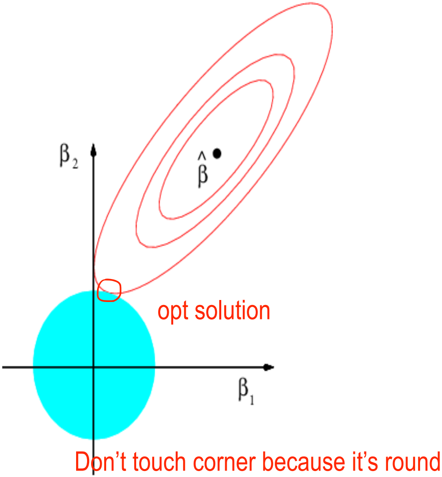
$$\min_{\beta} \sum_{i} \left(y_i - \beta_0 - \sum_{j} \beta_j x_{ij} \right)^2, \text{ subject to } \sum_{j} |\beta_j| \le c;$$

$$\min_{\beta} \sum_{i} \left(y_i - \beta_0 - \sum_{j} \beta_j x_{ij} \right)^2, \text{ subject to } \sum_{j} \beta_j^2 \le c;$$

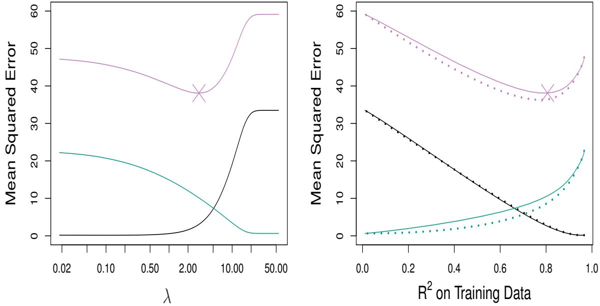
Ridge regression and the LASSO

Compressive Sensing (Candes/Tao)
Matrix Factorization/Sensing



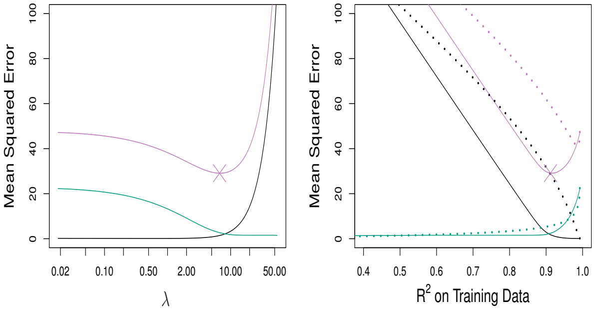


Credit data example



Left: Plots of squared bias (black), variance (green), and test mean squared error (purple) for the LASSO on a simulated data set. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). The purple crosses indicate the LASSO models for which the MSE is the smallest.

Credit data example



Left: Plots of squared bias (black), variance (green), and test mean squared error (purple) for the LASSO on another simulated data set. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). The purple crosses indicate the LASSO models for which the MSE is the smallest.

Conclusions

- 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 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.

and which hyperparameter lambda is better

Summary and Remark

- Ridge Regression
- The LASSO
- Ridge Regression and the LASSO
- Read textbook Chapter 3
- Do R lab