Regularized Regression

Machine Learning FRS-2021

Yuri Balasanov

iLykei Teaching Tech Corp

Outline of the Session

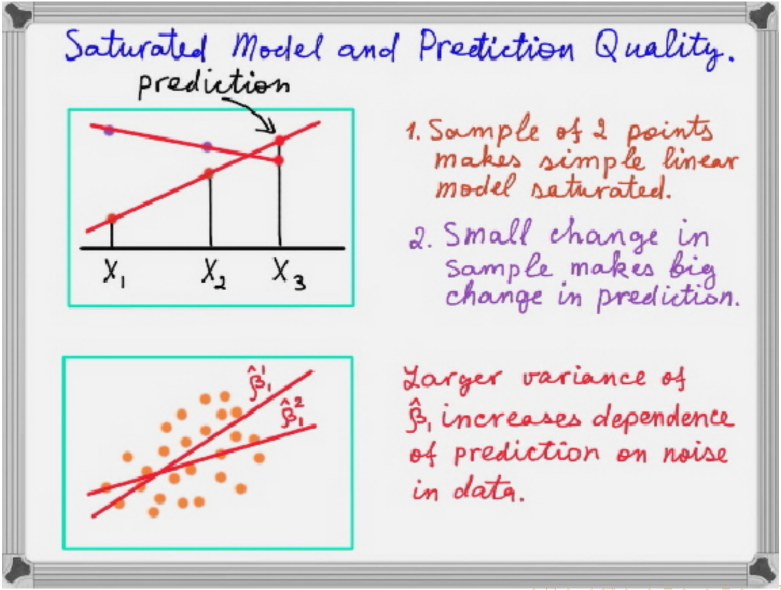
- Shrinkage and selection of predictors
  - Regularization
  - General problem
- Ridge regression
- Lasso regression

Text: ISLR, G.James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2013

Summary of Chapter 6

- Recall what is the main challenge of the age of Big Data and in what way it affects classical regression methods
- Connection between variance of estimators and accuracy of prediction: sensitivity of estimates to new data
- Methods recommended in Chapter 6 of ISLR:
  - Subset selection
  - Dimension reduction
  - Shrinkage
- MLS gives unbiased estimators of parameters, but large number of predictors make variance of the parameter estimators large
- Shrinkage method attempts improving predictive power (variance of estimators) by giving up some bias

How Increased Variance Affects Prediction

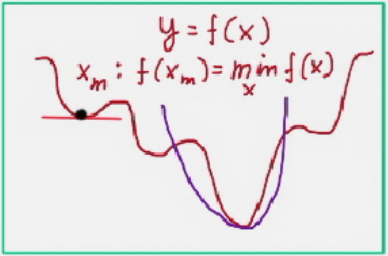


Shrinkage Methods

- When there are too many parameters and their number has not been reduced before fitting a model there is a need to remove some of them in the process of fitting
- Some methods called **shrinkage methods** are proposed in the book, section 6.2
- Shrinkage methods are based on technique constraining coefficient estimates, i.e. shrinks them to zero
- Shrinkage methods are part of a broader area of applied mathematics called **methods of regularization**
- The two methods proposed in the book for regularization of regression coefficients are called **ridge regression** and **lasso regression**

Regularization.

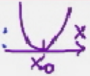
Problem of optimization.



Using a standard method, like Newton-Raphson may lead to a local minimum.

$\boxed{\text{Selecting } x_0 \text{ is the key}}$

Let $\alpha(x, x_0)$ be a function like this:



For example, $\alpha(x, x_0) = (x - x_0)^2$

Minimize $f(x) + \alpha(x, x_0)$ instead of $f(x)$

$\boxed{\alpha(x, x_0) \text{ is called regularizator}}$

Ridge Regression I

- In the process of fitting linear regression model we minimize the sum of squares

$$RSS\left(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}\right) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- In order to "encourage" the coefficients to be closer to zero ridge regression replaces minimization of RSS with minimization of

$$RSS\left(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}\right) + \alpha\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Function $\alpha\left(\boldsymbol{\beta}\right) = \lambda \sum_{j=1}^{p} \beta_j^2$ is called regularizator. Parameter $\lambda \geq 0$ is a tuning parameter.

Ridge Regression II

- As with least squares, ridge regression seeks coefficient estimates that fit the data, by making the RSS, the first term, small.
- The second term, called regularizator or shrinkage penalty, is small when $\beta_1, \ldots, \beta_p$ are close to zero. It forces the coefficients to be smaller in absolute value
- The objective is to eliminate small coefficients completely if they are not large enough
- The tuning parameter λ controls the relative impact of these two terms on the regression coefficient estimates
- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates
- However, as $\lambda \longrightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will all approach zero
- Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates for each value of λ. Selecting a good value for λ is critical

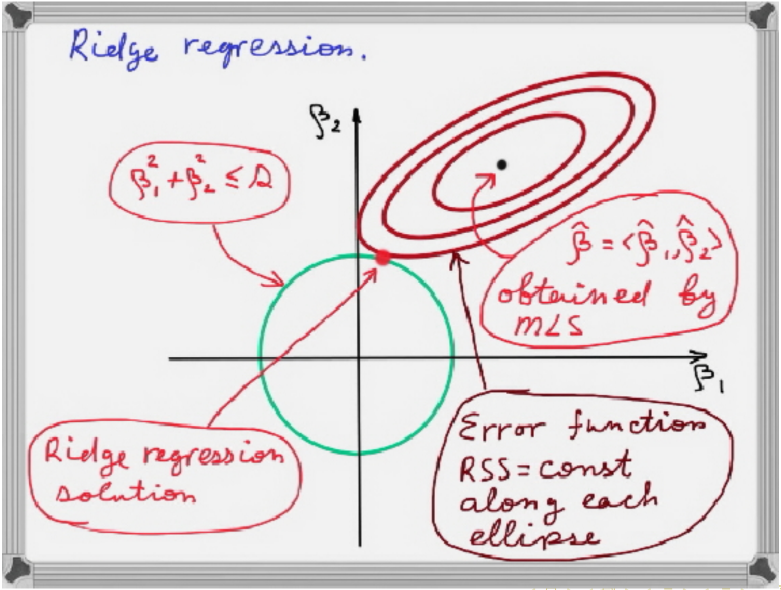
Ridge Regression III

- What ridge regression is trying to achieve is reduction in the variance of estimators
- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias for the coefficients
- This effect is called the bias-variance trade-off
- Regularization with L_2-norm regularizator which in fact is

$$\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

  also improves computational efficiency of least squares
- Effect of ridge regression can be illustrated on a simple graph

Ridge Regression IV



Lasso Regression I

- One disadvantage of ridge regression is that it shrinks all coefficients towards zero in a pretty uniform way and it may not make them exactly zero
- Lasso regression is designed to overcome that weakness of ridge regression. It replaces the L_2-norm regularizator of ridge regression with L_1-norm regularizator

$$RSS\left(\boldsymbol{\beta};\mathbf{y},\mathbf{x}\right)+\alpha\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

- As with ridge regression, lasso shrinks the coefficient estimates towards zero. However, in the case of lasso, the 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
- As in ridge regression, selecting a good value of λ for lasso is critical and done using cross-validation

Lasso Regression II