STAT 154 Lab 7: Lasso

Yuansi Chen and Raaz Dwivedi

Apr 1, 2019

1 Derivations with one-dimensional Lasso (Slide 6 of Lec 17)

2 LASSO vs ridge picture

Recall the picture trying to explain why ℓ_1 regularization leads to sparsity, while ℓ_2 regularization does not. (Figure 3.11 in ESL book) In this problem, first we try to understand the details of this picture.

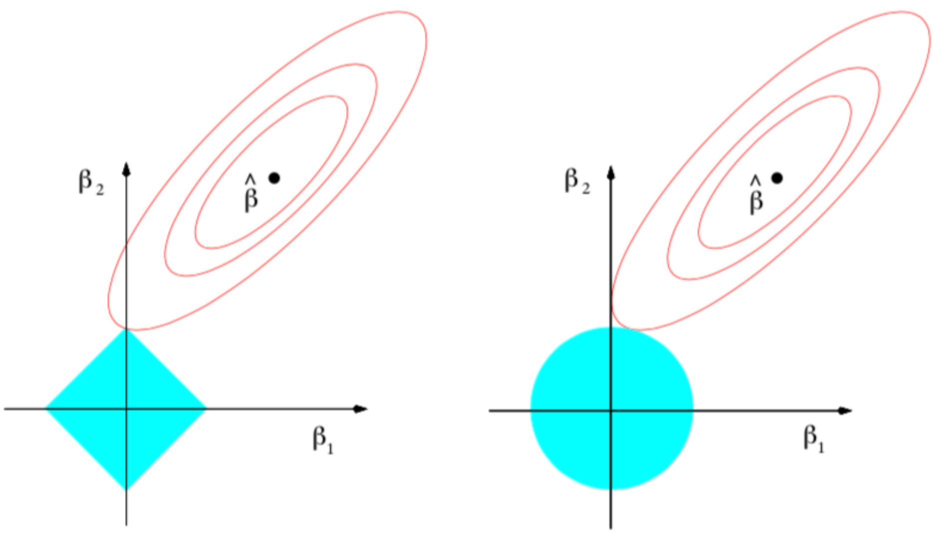

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

The mean squared error (MSE) on the training set in the linear regression problem with $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n$ $(n > d)$ is

$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^\top \beta - y_i\right)^2 = \frac{1}{n}\left\|\mathbf{X}\beta - \mathbf{y}\right\|_2^2.$$

Recall that $\hat{\beta}_{\mathrm{OLS}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$. The LASSO estimator $\hat{\beta}_{\mathrm{LASSO}}$ is the minimizer of the following minimization problem

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_{\text{LASSO}} \|\beta\|_1 , \tag{1}$$

and the ridge estimator $\hat{\beta}_{\text{ridge}}$ denotes the minimizer of the following minimization problem

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_{\text{ridge}} \|\beta\|_2^2 . \tag{2}$$

1. First we show that the level sets of the training loss are indeed **ellipsoids** centered at the training loss minimizer $\hat{\beta}$. Show that for any $\beta \in \mathbb{R}^d$, we have

$$R(\beta) = \frac{1}{n}\left(\beta - \hat{\beta}_{\text{OLS}}\right)^\top \mathbf{X}^\top \mathbf{X}\left(\beta - \hat{\beta}_{\text{OLS}}\right) + R(\hat{\beta}_{\text{OLS}}).$$

2. Give an expression for the set of β for which the empirical risk exceeds $R(\hat{\beta}_{\text{OLS}})$ by an amount $c > 0$. If $\mathbf{X}$ is full rank, then $\mathbf{X}^\top\mathbf{X}$ is positive definite, and this set is an ellipsoid. What is its center?

3. We now show that why it is conceptually okay to consider the constrained formulation of LASSO/Ridge. Show that for any $\lambda_{\text{LASSO}} > 0$, there exists μ_{LASSO} such that the LASSO problem (1) is equivalent to the following minimization problem,

$$\begin{aligned} &\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 \\ &\text{such that } \|\beta\|_1 \le \mu_{\text{LASSO}}. \end{aligned}$$

   Show that a similar result holds for ridge regression. We can solve this problem without introducing the Lagrange multipliers.

4. Show that $\beta = 0$ is the optimal Lasso estimator if $\lambda \geq ||X\theta - y||_\infty$.

5. When is $\beta = 0$ optimal for ridge regression?

We now see some comparisons between Lasso and Ridge.

4. What do the ridge regression coefficients look like when two feature columns $\mathbf{X}_{\cdot j}$ and $\mathbf{X}_{\cdot k}$ are identical?

5. What do the Lasso coefficients look like when two feature columns $\mathbf{X}_{\cdot j}$ and $\mathbf{X}_{\cdot k}$ are identical?

3 Discussion with Kernel Ridge Regression

4 Fun with OLS/Lasso

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda. The algorithm is fast, and can exploit sparsity in the input matrix $\mathbf{X}$.

glmnet solves the following minimization problem under option *family="gaussian"*.

$$\min_{\beta} \frac{1}{n} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \left[(1-\alpha)\|\beta\|_2^2/2 + \alpha \|\beta\|_1\right].$$

Try ridge and LASSO with **glmnet** in the following simulation settings.True model $\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\epsilon}$. Plot the training MSE v.s. CV MSE as a function of regularization parameter λ.

1. $\beta^* = (10, 10, 5, 5, \underbrace{1, \ldots, 1}_{10}, 0, \ldots, 0)^\top$, $d = 50 < n = 100$. $\mathbf{X}$ with entries i.i.d normal. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.

```r
library(glmnet)
library(MASS)
set.seed(123456)
d <- 50
n <- 100
ntest <- 200
X <- matrix(rnorm(d*n), ncol=d)
Xtest <- matrix(rnorm(d*ntest), ncol=d)
epsilon <- as.vector(rnorm(n))
epsilontest <- as.vector(rnorm(ntest))
betastar <- as.vector(rep(0, d))
```

2. $\beta^* = (10, 10, 5, 5, \underbrace{1, \ldots, 1}_{10}, 0, \ldots, 0)^\top$, $d = 50 < n = 100$. $Cov(\mathbf{X})_{ij} = (0.7)^{|i-j|}$. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.

```r
set.seed(123456)
d <- 50
n <- 100
ntest <- 20
CovMatrix <- outer(1:d, 1:d, function(x,y) {.7^abs(x-y)})
X <- as.matrix(mvrnorm(n, rep(0,d), CovMatrix))
```

3. $\beta^* = (\underbrace{1, \ldots, 1}_{15}, 0, \ldots, 0)^\top$, $d = 5000 > n = 1000$. $\mathbf{X}$ with entries i.i.d normal. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.

4. $\beta^* = (\underbrace{1, \ldots, 1}_{1500}, 0, \ldots, 0)^\top$, $d = 5000 > n = 1000$. $\mathbf{X}$ with entries i.i.d normal. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.