Chapter 2 The Lasso for Linear Models

2.2 LS

LASSO coef bias to 0

LS subset coef debias away from 0 relaxed Lasso

2.3 CV

$\downarrow$

fit model for $\{t_i\}_{i=1}^{T}$

Test $= K$ $\quad ER_1^k \quad \cdots \quad ER_T^k$

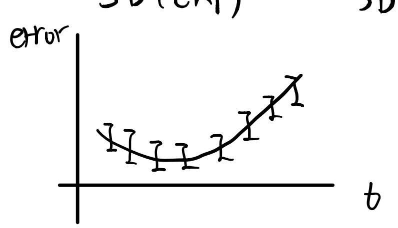
$\vdots$

Test $= 1$ $\quad ER_1^1 \quad \cdots \quad ER_T^1$

$\downarrow \qquad \downarrow$

$\overline{ER}_1 \quad \cdots \quad \overline{ER}_T$

$SD(ER_1) \qquad SD(ER_1)$



2.4 Computation of LASSO

QP problem $\min_{\beta} \left\{ \frac{1}{2N} \| y - X\beta \|_2^2 \right\}$

s.t. $\| \beta \|_1 \leq t$

Lagrangian $\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$

$$\min_{\beta} \left\{ \frac{1}{2N} \| y - X\beta \|_2^2 + \lambda \|\beta\|_1 \right\}$$

where $\frac{1}{N} \sum_i y_i = 0$ $\quad \frac{1}{N} \sum_i x_{ij} = 0$ $\quad \frac{1}{N} \sum_i x_{ij}^2 = 1$

EX 2.2 Derivation for LASSO by inspection

Since X has been standardized, the $\hat{\beta}_{LS} = (X^T X)^{-1} X^T y = X^T y$

Expanding the first term of the Lagrangian form,

$$\frac{1}{2N} \left[(y - X\beta)^T (y - X\beta) \right]$$

$$= \frac{1}{2N} \left[y^T y - (X\beta)^T y - y^T X\beta + (X\beta)^T (X\beta) \right]$$

$$= \frac{1}{2N} \left[y^T y - 2 (X\beta)^T y + \beta^T X^T X \beta \right]$$

$$= \frac{1}{N} \left[\underbrace{\frac{1}{2} y^T y}_{\text{no } \beta \text{ here}} - y^T X\beta + \frac{1}{2} \beta^T \beta \right]$$

The problem changes to

$$\min_{\beta} \quad \frac{1}{N}\left[-y^T X\beta + \frac{1}{2}\beta^T\beta\right] + \lambda \|\beta\|_1 \qquad \|\beta\|_1 = \sum_{i=1}^{N} |\beta_i|$$

$$\min_{\beta} \quad \frac{1}{N}\left[-\hat{\beta}_{LS}^T\beta + \frac{1}{2}\beta^T\beta\right] + \lambda \|\beta\|_1$$

$$\min_{\beta} \quad \frac{1}{N}\sum_{i=1}^{N}\left(-\hat{\beta}_{LS_i}\beta_i + \frac{1}{2}\beta_i^2 + N\lambda|\beta_i|\right)$$

So, the problem can be solved as individual problems indexed by i.

For a certain i, $\min L_i = -\hat{\beta}_{LS_i}\beta_i + \frac{1}{2}\beta_i^2 + N\lambda|\beta_i|$

If $\hat{\beta}_{LS_i} > 0$, we must have $\beta_i \geq 0$

< 0 $\qquad \leq 0$

Case 1, If $\hat{\beta}_{LS_i} > 0$, since $\beta_i \geq 0$,

$$L_i = -\hat{\beta}_{LS_i}\beta_i + \frac{1}{2}\beta_i^2 + N\lambda\beta_i$$

$$\frac{\partial L_i}{\partial \beta_i} = -\hat{\beta}_{LS_i} + \beta_i + \lambda N = 0$$

$$\beta_i = \hat{\beta}_{LS_i} - \lambda N$$

with the assumption,

$$\beta_i = (\hat{\beta}_{LS_i} - \lambda N)_+ = \operatorname{sgn}(\hat{\beta}_{LS_i})(|\hat{\beta}_{LS_i}| - \lambda N)$$

Case 2. If $\hat{\beta}_{LS_i} < 0$, since $\beta_i \leq 0$

$$L_i = -\hat{\beta}_{LS_i}\beta_i + \frac{1}{2}\beta_i^2 - N\lambda\beta_i$$

$$\beta_i = (\hat{\beta}_{LS_i} + \lambda N)_- = \mathrm{sgn}(\hat{\beta}_{LS_i})(|\hat{\beta}_{LS_i}| - \lambda N)_+$$

To combine them together,

$$\hat{\beta} = \begin{cases} \frac{1}{N}\hat{\beta}_{LS} - \lambda & \text{if } \frac{1}{N}\hat{\beta}_{LS} > \lambda \\ 0 & \text{if } \frac{1}{N}|\hat{\beta}_{LS}| \leq \lambda \\ \frac{1}{N}\hat{\beta}_{LS} + \lambda & \text{if } \frac{1}{N}\hat{\beta}_{LS} < -\lambda \end{cases}$$

$$\hat{\beta} = S_\lambda\left(\frac{1}{N}\hat{\beta}_{LS}\right)$$

$$S_\lambda(x) = \mathrm{sign}(x)\,(|x| - \lambda)_+$$

By KKT conditions (subgradient)

$$\min\left\{\frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\}$$

by KKT cond.

$$-\frac{1}{N}X^T(y - X\beta) + \lambda s = 0 \qquad ①$$

where s is the subgradient of $\|\cdot\|_1$ ℓ_1 norm,

$$s_j = \begin{cases} \mathrm{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \in [-1, 1] & \text{if } \beta_j = 0 \end{cases}$$

when $X^TX = I$, ① becomes

$$-\frac{1}{N}(\hat{\beta}^{LS} - \beta) + \lambda s = 0$$

Consider the case where the solution would be $\beta_j = 0$. For this to be true we must have $\frac{1}{N}\hat{\beta}_j^{LS} = \lambda s \in [-\lambda, \lambda]$

$$\left|\frac{1}{N}\hat{\beta}_j^{LS}\right| \leq \lambda \iff \beta_j = 0$$

(KKT is sufficient)

$\beta_j \neq 0$, if $\beta_j > 0$,

$$\frac{1}{N}(\hat{\beta}^{LS} - \beta) = \lambda$$

$$\hat{\beta}^{LS} - \beta = N\lambda$$

$$\beta = \hat{\beta}^{LS} - N\lambda$$

Multiple Parameters: Cyclic Coordinate Descent

Repeatedly cycle through the predictors in some fixed order (say $j=1,\dots,p$), where at the jth step, we update the coefficient β_j by minimizing the objective function in this coordinate while holding fixed all other coefficients $\{\beta_k, k\neq j\}$ at their current values.

Writing the objective function as

$$\min_{\beta_j} \frac{1}{2N}\sum_{i=1}^{N}\Big(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\Big)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|$$

partial Residual $\quad r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat\beta_k$

$$r_i = y_i - \sum_{k\neq j} x_{ik}\hat\beta_k - x_{ij}\hat\beta_j$$

$$\frac{1}{N}x_j^T r^{(j)} = \frac{1}{N}\sum_{i=1}^{N} x_{ij} r_i^{(j)}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_{ij}y_i - x_{ij}\sum_{k\neq j} x_{ik}\hat\beta_k$$

$$\frac{1}{N}x_j^T r = \frac{1}{N}\sum x_{ij}y_j - x_{ij}\sum_{k\neq j} x_{ik}\hat\beta_k - x_{ij}^2\hat\beta_j$$

$$= \_\ \ \ldots\ \ \_\ \ \ \_\ \hat\beta_j$$

2.5 Degree of Freedom

Adaptive Model: Use degree of freedom more ~~than~~ the number of its parameters.

LASSO's DoF is unbiased

2.6 Uniqueness of the LASSO Solutions

X is full rank: the solution of LASSO is Unique
$\lambda > 0$

when $p \geq N$: the solution of LASSO is Unique when # nonzero coefficient is not larger than N.

X is not full rank: LS fitted values are unique, but coefs are not unique
(Caused by ① $p \leq N$ collinearity, ② $p > N$)
In ②, there're infinite number of solutions yield 0 training error.

LASSO, the fitted value $X\hat{\beta}$ are unique, but $\hat{\beta}$ may not be unique