HomeWork # 3   ARINJAY JAIN
A20447307

Problem # 1 Ex. 3.12 Show that ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set.

Given

$$X_{new} = \begin{pmatrix} X_{n\times p} \\ \sqrt{\lambda} I_{p\times p} \end{pmatrix} \qquad Y_{new} = \begin{pmatrix} Y_{n\times 1} \\ 0_{p\times 1} \end{pmatrix}_{n+p}$$

$X_{new} \Rightarrow$ row $= n+p$
column $= p$

$Y_{new} \Rightarrow$ row $= n+p$
column $= 1$.

$$\hat{\beta}^{OLS} = (X^T X)^{-1} \cdot X^T \cdot Y$$

we have to show

$$\hat{\beta}^{OLS}_{(X_{new}, Y_{new})} = \hat{\beta}^{ridge}.$$

for example $\Rightarrow$ let take $n = 3$ and $p = 2$

$$X_{new} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ \sqrt{\lambda} I & \sqrt{\lambda} I \\ \sqrt{\lambda} I & \sqrt{\lambda} I \end{pmatrix} \qquad Y_{new} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ 0 \\ 0 \end{pmatrix}$$

$$(X_{new}^T \cdot X_{new}) =$$

$$X^T_{new} = \begin{pmatrix} \underbrace{\begin{matrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{matrix}}_{X^T} & \begin{matrix} \sqrt{\lambda} I & \sqrt{\lambda} I \\ \sqrt{\lambda} I & \sqrt{\lambda} I \end{matrix} \end{pmatrix}$$

$$X^T_{new} X_{new} = \begin{pmatrix} x_{11}^2 + x_{21}^2 + x_{31}^2 + 2\lambda I & x_{11}x_{12} + x_{21}x_{22} + x_{31}x_{32} + 2\lambda I \\ x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} + 2\lambda I & x_{11}^2 + x_{21}^2 + x_{31}^2 + 2\lambda I \end{pmatrix}$$

In general

$$\left\{ X^T_{new} X_{new} = X^T X + \cancel{2} p \lambda I \right\}$$

$p \cdot \lambda =$ constant. (λ)

$$\left[X^T_{new} X_{new} = X^T X + \lambda I \right]$$

$$X^T_{new} Y_{new} = \begin{pmatrix} \underbrace{\begin{matrix} x_{11} y_1 + x_{21} y_2 + x_{31} y_3 \\ x_{12} y_1 + x_{22} y_2 + x_{32} y_3 \end{matrix}}_{} & \underbrace{\begin{matrix} + \sqrt{\lambda} I \cdot 0 + \sqrt{\lambda} I \cdot 0 \\ + \sqrt{\lambda} I \cdot 0 + \sqrt{\lambda} I \cdot 0 \end{matrix}}_{0} \end{pmatrix}$$

$$X^T_{new} \cdot Y_{new} = X^T \cdot Y$$

$$\hat{\beta}^{OLS}_{new} = (X^T_{new} \cdot X_{new}) X^T_{new} \cdot Y_{new}$$

$$\left\{ \hat{\beta}^{OLS}_{new} = (X^T \cdot X + \lambda I) X^T \cdot Y = \hat{\beta}^{ridge} \right\}$$

#

Problem #2 Ex 3.30

Consider the elastic-net optimization problem:

$$\min_{\beta} \|y - X\beta\|^2 + \lambda[\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1]$$

Show how one can turn this into a lasso problem, using an augmented version of X and y.

Solution ⇒

Augmented version of X and y will be $\tilde{X}$ & $\tilde{y}$

$$\tilde{X} = \begin{bmatrix} X \\ \gamma I_p \end{bmatrix} \qquad \tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

$$\tilde{X}\beta = \begin{bmatrix} X\beta \\ \gamma\beta \end{bmatrix}$$

from given hint we know that

$$\left\{ \|\tilde{y} - \tilde{X}\beta\|_2^2 = \left\| \begin{bmatrix} y - X\beta \\ \gamma\beta \end{bmatrix} \right\|_2^2 = \|y - X\beta\|_2^2 + \gamma^2\|\beta\|_2^2 \right\}$$

elastic-net

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

$$= \|y - X\beta\|_2^2 + \underbrace{\lambda\alpha}_{\text{Constant} \to \gamma^2}\|\beta\|_2^2 + \underbrace{\lambda(1-\alpha)}_{\text{Constant} \to \tilde{\lambda}}\|\beta\|_1$$

$$= \underbrace{\|Y - X\beta\|_2^2 + \gamma^2 \|\beta\|_2^2} + \tilde{\lambda} \|\beta\|_1$$

$$= \|\tilde{y} - \tilde{X}\beta\|_2^2 + \tilde{\lambda} \|\beta\|_1 \quad \#$$

This is a lasso objective function in the form of augment $\tilde{X}$ and $\tilde{y}$

Q4 (3.16) Derive the entries in table 3.4 the explicit forms for estimators in the orthogonal case.

Solution :→

Given table :→

Estimator	Formula.
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_M\rvert)$
Ridge	$\hat{\beta}_j / (1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

by the defination of orthonormal

$$X^T \cdot X = I$$

OLS

$$\hat{\beta} = (X^T X)^{-1} \cdot X^T \cdot y \Rightarrow (I)^{-1} X^T \cdot y$$

$$\hat{\beta} = X^T \cdot y$$

① Best subset ⇒ will take the m predictor with smallest residual sum of square. (RSS).

we know that columns of X are orthronormal we can Construct a basis of euclidean Space R^n equiped with the standard inner product. This will be happen by using the first 'p' columns of X and the extending there to 'N-p' linearly indepen-dent additional orthonormal vectors.

$$y = \sum_{j=1}^{p} \hat{\beta}_j x_j + \sum_{j=p+1}^{N} \gamma_j \tilde{x}_j$$

where $\hat{\beta}_j$ = Component of $\hat{\beta}$ in eq

γ_j = cofficients of 'y' w.r.t extended basis vector.

Best Subset Selection method estimatge of y can be written as

$$\hat{y} = \sum_{j=1}^{p} I_j \hat{\beta}_j x_j$$ { where $I_j = 1$ if the predictor X_i are kept or zero other wise.

~~As~~ X_i & $\tilde{X}_j$ are orthonormal

So,

$$\|y - \hat{y}\|_2^2 = \|y - X\hat{\beta}\|_2^2$$

$$\|y - \hat{y}\|_2^2 = \| \underbrace{\sum_{j=1}^{p} \hat{\beta}_j x_j + \sum_{j=p+1}^{N} \gamma_j \tilde{x}_j}_{\Downarrow \; y} - \underbrace{\sum_{j=1}^{p} I_j \hat{\beta}_j x_j}_{\Downarrow \; \hat{y}} \|$$

$$\|y - \hat{y}\|_2^2 = \| \sum \hat{\beta}_j \cancel{x_j} (1 - I_j) x_j + \sum \gamma_j \tilde{x}_j \|_2^2$$

$$= \sum_{j=1}^{p} \hat{\beta}_j^2 (1 - I_j)^2 \|x_j\|_2^2 + \sum_{j=p+1}^{N} \gamma_j^2 \|\tilde{x}_j\|_2^2$$

$$\|y - \hat{y}\|_2^2 = \sum_{j=1}^{p} \hat{\beta}_j^2 (1 - I_j)^2 + \sum_{j=1}^{N} \gamma_j^2$$

We can minimize $\|y - \hat{y}\|_2^2$ we will those m values of I_j that are equal to one which have the largest values of $\hat{\beta}_j^2$

Indicator function build a reation between A and x.

1 = all the elements of x in A

0 = all the elements of x not in A.

By the definition of Indicator function ~~our~~ we can sort the values of $|\beta_j|$ ~~easily~~ and get Only those values with the indices of largest M measing where $I_j = 1$ and remaining indices with $I_j = 0$ are taken out.

by using Indicator function

$$\hat{\beta}_j^{\text{best subset (size } M)} = \hat{\beta}_j \times I(\text{rank}(|\hat{\beta}_j| \leq M))$$

$$\left\{ \hat{\beta}_j^{bs} = \hat{\beta}_j^{ls} = X^T \cdot y \right\}.$$

for Ridge Regression :-

we know that.

$$\hat{\beta}^{ridge} = (\overbrace{X^T \cdot X}^{\text{orthonormal,}} + \lambda I)^{-1} X^T \cdot y.$$

$$= (I + \lambda I)^{-1} X^T \cdot y$$

$$= \frac{X^T \cdot y}{\cancel{I}(1+\lambda)} = \frac{\overbrace{X^T \cdot y}^{\hat{\beta}^{ls}}}{1+\lambda}$$

$$\boxed{\hat{\beta}^{ridge} = \frac{\hat{\beta}^{ls}}{1+\lambda}}$$

for Lasso :

we know that

$$L(\beta) = (y - x\beta)^T (y - x\beta) + \lambda |\beta|$$

first order derivative. w.r.t β.

$$\frac{\partial f(\beta)}{\partial \beta} = -x^T y + x^T x \beta + \lambda . \operatorname{Sign}(\beta)$$

for max (β) will $\frac{\partial L(\beta)}{\partial \beta} = 0$

$$-x^T y + x^T x \hat{\beta} + \lambda \operatorname{Sign}(\beta) = 0$$

$$\underbrace{x^T . x}_{\text{orthonormal}} \hat{\beta} = x^T y - \lambda \operatorname{Sign}(\beta)$$

$$I \hat{\beta} = x^T y - \lambda \operatorname{Sign}(\beta)$$

$$\hat{\beta} = I^{-1} (x^T y - \lambda \operatorname{Sign}(\beta)) \quad \therefore I^{-1} = I$$

$$= I (x^T y - \lambda \operatorname{Sign}(\beta))$$

$$\hat{\beta}^{lasso} = \operatorname{Sign}(\beta) (|x^T y| - \lambda)$$

$$\left\{ \hat{\beta}^{lasso} = \operatorname{Sign}(\beta) (|x^T y| - \lambda)_+ \right\}$$