Assignment - 3

Name - Biswarup Karmakar
SR : - 21055

Question: 1

(I) The simple linear regression model is given by

$$\hat{y} = \hat{w}_0 + \sum_{j=1}^{M} x_j \hat{w}_j \quad \text{——— (1)}$$

For a given set of outputs y in the training set we minimize the cost function $L(\hat{w}) = \|y - X\hat{w}\|^2$ such that we get best w.

where X = feature matrix

$\hat{w}$ = coefficient vector $[\hat{w}_0 \ \hat{w}_1 \ \cdots \ \hat{w}_M]^T$

Now
$$\begin{aligned} L(\hat{w}) &= \|y - X\hat{w}\|^2 \\ &= (y - X\hat{w})^T (y - X\hat{w}) \\ &= (y^T - \hat{w}^T X^T)(y - X\hat{w}) \\ &= y^T y - y^T X\hat{w} - \hat{w}^T X^T y - \hat{w}^T X^T X \hat{w} \end{aligned}$$

we minimize $L(\hat{w})$ with respect to $\hat{w}$, i.e, we set the derivative w.r.t $\hat{w}$ is 0.

$$\nabla L(\hat{w}) = 0 \Rightarrow 0 - X^T y - X^T y - 2X^T X\hat{w} = 0$$

[$X^T X$ is positive definite, symmetric]
[Also $\hat{w}^T X^T y = y^T X \hat{w}$ is a scalar]

$$\Rightarrow 2X^T y = 2X^T X \hat{w}$$
$$\Rightarrow \hat{w} = (X^T X)^{-1} X^T y$$

[Here X is full column rank matrix, so $X^T X$ is invertible]

Hence proved.

(II) Now we consider the linear regression in one variable $\hat{y} = \hat{w}_0 + x\hat{w}_1$

Suppose the mean squared error be denoted by S_r i.e,

$$S_r = \frac{1}{N}\sum_{i=1}^{N}(y^i - \hat{y}^i)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y^i - \hat{w}_0 - x^i\hat{w}_1)^2$$

where (x^i, y^i) are i th data point for $i = 1, 2, \dots N$.

Now Differentiating w.r.t $\hat{w}_0$ we get

$$\frac{\partial S_r}{\partial \hat{w}_0} = \frac{1}{N}\, 2 \sum_{i=1}^{N}(y^i - \hat{w}_0 - x^i\hat{w}_1)(-1) = 0$$

(We minimizing S_r w.r.t $\hat{w}_0$)

$$\Rightarrow \sum_{i=1}^{N}(y^i - \hat{w}_0 - x^i\hat{w}_1) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} w_0 + \sum_{i=1}^{N} x^i\hat{w}_1} \quad \text{--- (1)}$$

Differentiating w.r.t $\hat{w}_1$, [minimizing w.r.t $\hat{w}_1$]

$$\frac{\partial S_r}{\partial \hat{w}_1} = \frac{2}{N}\sum_{i=1}^{N}(y^i - \hat{w}_0 - x^i\hat{w}_1)(-x^i) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{N} x^i y^i = \sum_{i=1}^{N} x^i\hat{w}_0 + \sum_{i=1}^{N}(x^i)^2\hat{w}_0} \quad \text{--- (2)}$$

These two equations are the normal equations with two variable.

From equation (1)

$$N\hat{w}_0 = \sum_{i=1}^{N} y^i - \sum_{i=1}^{N} x^i\hat{w}_1$$

$$\Rightarrow \hat{w}_0 = \frac{1}{N}\left(\sum_{i=1}^{N} y^i - \hat{w}_1\sum_{i=1}^{N} x^i\right) \quad \text{--- (3)}$$

Substituting ③ in ②, we get

$$\left(\frac{1}{N}\sum_{i=1}^{N} y^i - \hat{w}_1 \frac{1}{N}\sum_{i=1}^{N} x^i\right)\sum_{i=1}^{N} x^i + \hat{w}_1 \sum_{i=1}^{N} (x^i)^2 = \sum_{i=1}^{N} x^i y^i$$

$$\Rightarrow \left(\sum_{i=1}^{N} x^i\right)\left(\sum_{i=1}^{N} y^i\right) - \hat{w}_1 \left(\sum_{i=1}^{N} x^i\right)^2 + N\hat{w}_1 \sum_{i=1}^{N} (x^i)^2 = N\sum_{i=1}^{N} x^i y^i$$

$$\Rightarrow \hat{w}_1 \left[N\sum_{i=1}^{N} (x^i)^2 - \left(\sum_{i=1}^{N} x^i\right)^2\right] = N\sum_{i=1}^{N} x^i y^i - \left(\sum_{i=1}^{N} x^i\right)\left(\sum_{i=1}^{N} y^i\right)$$

$$\Rightarrow \boxed{\hat{w}_1 = \frac{N\sum_{i=1}^{N} x^i y^i - \left(\sum_{i=1}^{N} x^i\right)\left(\sum_{i=1}^{N} y^i\right)}{N\sum_{i=1}^{N} (x^i)^2 - \left(\sum_{i=1}^{N} x^i\right)^2}}$$

Hence derived.

Now we substitute $\hat{w}_1$ in ③

$$\hat{w}_0 = \frac{1}{N}\left[\sum_{i=1}^{N} y^i - \frac{N\sum_{i=1}^{N} x^i y^i - \left(\sum_{i=1}^{N} x^i\right)\left(\sum_{i=1}^{N} y^i\right)}{N\sum_{i=1}^{N} (x^i)^2 - \left(\sum_{i=1}^{N} x^i\right)^2}\sum_{i=1}^{N} x^i\right]$$

$$= \frac{1}{N}\,\frac{N\sum_{i=1}^{N} (x^i)^2 \sum_{i=1}^{N} y^i - \cancel{\left(\sum_{i=1}^{N} x^i\right)^2 \sum_{i=1}^{N} y^i} - N\sum_{i=1}^{N} x^i y^i \sum_{i=1}^{N} x^i + \cancel{\left(\sum_{i=1}^{N} x^i\right)^2 \sum_{i=1}^{N} y^i}}{N\sum_{i=1}^{N} (x^i)^2 - \left(\sum_{i=1}^{N} x^i\right)^2}$$

$$\hat{w}_0 = \frac{\sum_{i=1}^{N} (x^i)^2 \sum_{i=1}^{N} y^i - \sum_{i=1}^{N} x^i y^i \sum_{i=1}^{N} x^i}{N\sum_{i=1}^{N} (x^i)^2 - \left(\sum_{i=1}^{N} x^i\right)^2}$$

Question 2

(a) For ridge regression, the ridge co-efficient minimize a penalised residual sum of squares

$$Z(w_0, w_j) = \sum_{i=1}^{N} \left(y^i - w_0 - \sum_{j=1}^{M} x_j^i w_j\right)^2 + \lambda \sum_{j=1}^{M} w_j^2$$

Now with the vector Notation, this can be written as

$$Z(w_0, W) = (y - XW - w_0 1)^T (y - XW - w_0 1) + \lambda W^T W$$

where $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$, $W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix} \in \mathbb{R}^M$ ——— (1)

$1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^N$, $X = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_M^1 \\ \vdots & & & \\ x_1^N & x_2^N & \cdots & x_M^N \end{pmatrix} \in \mathbb{R}^{N \times M}$

We will minimize Z with respect to coefficient so, $\nabla_{w_0} Z(w_0, W) = 0$ and $\nabla_W Z(w_0, W) = 0$

From 1st ~~result~~, condition

$$\nabla_{w_0} Z = 0 \Rightarrow 2\,(y - XW - w_0 1)^T_{1\times N} \cdot 1_{N\times 1} = 0$$

$$\Rightarrow \sum_{i=1}^{N} y^i - \sum_{i=1}^{N} x_j^i w_j - w_0 N = 0$$

$$\Rightarrow w_0 = \frac{1}{N}\left(\sum_{i=1}^{N} y^i - w_j \sum_{i=1}^{N} x_j^i\right)$$

$$\Rightarrow w_0 = \bar{y} - w_j \bar{x}_j \quad \text{——— (2)}$$

∴ The intercept $\hat{w}_0 = \bar{y}$ (Given that $\bar{x} = 0$)

Now From the 2nd ~~result~~ condition

$$\nabla_W Z = 0 \Rightarrow 2(y - XW - w_0 1)^T \cdot (-X) + 2\lambda W^T = 0$$

$$\Rightarrow (y - XW - w_0 1)^T \cdot X = \lambda W^T$$

$$\Rightarrow y^T X - W^T X^T X - w_0 1^T \cdot X = \lambda W^T \quad \text{——— (3)}$$

we have

$w_0 1^T . X$

$= w_0 . (1, 1, \dots 1)_{1 \times N} \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_M^1 \\ \vdots & & & \\ x_1^N & x_2^N & \dots & x_M^N \end{pmatrix}$

$= w_0 . N (\bar{x}_1, \bar{x}_2, \dots \bar{x}_M)$

$= w_0 N \bar{x} = 0 \qquad (\because \bar{x} = 0)$

From (3) we get

$y^T X - W^T X^T X = \lambda W^T$ for the matrices A, B

$\Rightarrow W^T X^T X + \lambda W^T = y^T X$ $[\because (AB)^T = B^T A^T$

$\Rightarrow (XW)^T X + (\lambda W)^T = (X^T y)^T$

$\Rightarrow (X^T X W + \lambda W)^T = (X^T y)^T$ $[\because A^T + B^T = (A+B)^T]$

$\Rightarrow X^T X W + \lambda W = X^T y$

$\Rightarrow (X^T X + \lambda I) W = X^T y$ $(X \in \mathbb{R}^{N \times M}, y \in \mathbb{R}^N)$

Now $X^T X$ is $M \times M$ square matrix and invertible as X has full column rank.

So, $X^T X$ has non zero eigen value.

So, $X^T X + \lambda I$ also has non-zero eigen value.

Thus $(X^T X + \lambda I)^{-1}$ exists.

$\therefore \boxed{\hat{w}^c_{ridge} = (X^T X + \lambda I)^{-1} X^T y}$

Hence proved

(b) Now $\hat{w}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$

$\hat{y} = X \hat{w}_{ridge}$

SVD of X is USV^T

where $U \in \mathbb{R}^{N \times D}$ orthogonal matrix with column as u_j $(j = 1, \dots D)$ with $U^T U = I$

$S \in \mathbb{R}^{D \times D}$ Diagonal matrix with diagonal entries $s_1, s_2, \dots s_D$.

$V \in \mathbb{R}^{D\times D}$ orthogonal matrix ($\because V^TV = I$)

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

$= USV^T \left[(USV^T)^T USV^T + \lambda I \right]^{-1} (USV^T)^T y$

$= USV^T \left[VSU^TUSV^T + \lambda I \right]^{-1} VSU^T y$ [$S^T = S$ as S diagonal]

$= USV^T \left[VS^2V^T + \lambda I \right]^{-1} VSU^T y$ [$U^TU = I_{D\times D}$]

$= USV^T \left[(VS^2 + \lambda (V^T)^{-1}) V^T \right]^{-1} VSU^T y$

for two matrices $((AB)^{-1} = B^{-1}A^{-1}$

$= USV^T (V^T)^{-1} \left[VS^2 + \lambda (V^T)^{-1} \right]^{-1} VSU^T y$

$= US \left[V (S^2 + \lambda V^{-1} (V^T)^{-1}) \right]^{-1} VSU^T y$

$= US \left[S^2 + \lambda (V^TV)^{-1} \right]^{-1} V^{-1} VSU^T y$

$= US \left[S^2 + \lambda I \right]^{-1} SU^T y$

$= \sum_{j=1}^{D} US \left(\frac{1}{s_j^2 + \lambda} \right) SU^T y$

[Any diagonalizable matrix A has decomposition as $A = U \Lambda V^T$ $= \sum_{j=1}^{D} u_j \lambda_j v_j^T$ where $\lambda_j \in \mathbb{R}$, $u_j \in \mathbb{R}^D$, $v_j \in \mathbb{R}^D$

$= \left(\sum_{j=1}^{D} u_j s_j \frac{1}{s_j^2 + \lambda} s_j u_j^T \right) y$

$= \sum_{j=1}^{D} u_j \frac{s_j^2}{s_j^2 + \lambda} u_j^T \cdot y$

so, $\hat{y} = \sum_{j=1}^{D} u_j f(s_j, \lambda) u_j^T y$ where $f(s_j, \lambda) = \frac{s_j^2}{s_j^2 + \lambda}$

For the least sqaures we use

$$W_{ls} = (x^T x)^{-1} x^T y$$

Then $\hat{y} = x W_{ls}$

$$= x (x^T x)^{-1} x^T y$$

$$= USV^T [(USV^T)^T USV^T]^{-1} (USV^T)^T y$$

$$= USV^T [V S U^T U S V^T]^{-1} V S U^T y \quad [S^T = S \text{ as } S \text{ diagonal}$$

$U^T U = I_{D \times D}$

$$= USV^T [V S^2 V^T]^{-1} V S U^T y$$

for three matrices

$$= USV^T (V^T)^{-1} (S^2)^{-1} V^{-1} V S U^T y \quad [(ABC)^{-1} = C^{-1} B^{-1} A^{-1}$$

$$= US\, I_{D\times D} (S^2)^{-1} \; I_{D \times D} S U^T y$$

$$= US (S^2)^{-1} S U^T y$$

$$= US S^{-1} S^{-1} S U^T y = U I_{D\times D} I_{D \times D} U^T y$$

For least square prediction

$$= U U^T y$$

$$\hat{y} = \sum_{j=1}^{D} u_j u_j^T y$$

Here we that

$$\hat{y}_{ridge} = \sum_{j=1}^{D} u_j f(s_j, \lambda) u_j^T y$$

$$\hat{y}_{ls} = \sum_{j=1}^{D} u_j u_j^T y$$

The effective degrees of freedom defined as $df(\lambda) = \sum_{j=1}^{D} \frac{s_j^2}{s_j^2 + \lambda}$

where s_j is singular value of X,

(i) Now, when $\lambda = 0$ we have D parameters since there is no penalisation. This corresponds to no shrinkage, gives $df(\lambda) = D$.

This is the case for least square ridge regression.

(ii) When λ is large, the parameters heavily constrained and the degrees of freedom will effectively be lower, tending ~~as~~ to 0 as $\lambda \to \infty$. i.e all the weights w_j are shrink to zero.

Ridge regression shrinks the co-ordinates with respect to the orthonormal basis formed by the principal components, co-ordinates with respect to principal components u_j with smaller variance are shrunk more.

The larger λ is, the more the projection is shrunk in the direction of u_j.

Question 3

Given that $Z_1(W)$

$$= (y - XW)^T (y - XW) + \lambda_2 |W|^2 + \lambda_1 |W|_1$$

$$Z_2(W) = (\hat{y} - \tilde{X}W)^T (\hat{y} - \tilde{X}W) + c\lambda_1 |W|_1$$

where $c = \frac{1}{\sqrt{1+\lambda_2}}$, $\tilde{X} = c\begin{pmatrix} X \\ \sqrt{\lambda_2} I_d \end{pmatrix}$, $\hat{y} = \begin{pmatrix} y \\ 0_{d\times 1} \end{pmatrix}$

We want to show

$$Z_1(cW) = Z_2(W)$$

Now $X = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_d^1 \\ x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & \vdots \\ x_1^n & x_2^n & \cdots & x_d^n \end{bmatrix}_{n\times d}$

$y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{pmatrix}_{n\times 1} \in \mathbb{R}^n$ $\quad W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}_{d\times 1} \in \mathbb{R}^d$

where (x^i, y^i) $i = 1, 2, \ldots n$, $x^i \in \mathbb{R}^d$ are data points

Now the modified data

$$\tilde{X} = c\begin{pmatrix} X \\ \sqrt{\lambda_2} I_d \end{pmatrix} = c\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_d^1 \\ x_1^2 & x_2^2 & \cdots & x_d^2 \\ \cdots & \cdots & \cdots & \cdots \\ x_1^n & x_2^n & \cdots & x_d^n \\ \sqrt{\lambda_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_2} \end{bmatrix}_{(n+d)\times d}$$

$$\hat{y} = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{n+d}$$

Hence

$$Z_1(w) = (y^T - w^T x^T)(y - xw) + \lambda_2 w^T w + \lambda_1 |w|_1$$

$$= y^T y - y^T x w - w^T x^T y + w^T x^T x w + \lambda_2 w^T w + \lambda_1 |w|_1$$

$$Z_1(cw) = y^T y - y^T x (cw) - (cw)^T x^T y + (cw)^T x^T x (cw) + \lambda_2 (cw)^T cw + \lambda_1 |cw|_1$$

$$= y^T y - c\, y^T x w - c\, w^T x^T y + c^2 w^T x^T x w + c^2 \lambda_2 w^T w + c\lambda_1 |w|_1$$

$$Z_2(w) = (\tilde{y} - \tilde{x} w)^T (\tilde{y} - \tilde{x} w) + c\lambda_1 |w|_1$$

$$= \tilde{y}^T \tilde{y} - \tilde{y}^T \tilde{x} w - w^T \tilde{x}^T \tilde{y} + w^T \tilde{x}^T \tilde{x} w + c\lambda_1 |w|_1 \quad \text{—— (4)}$$

Calculating $\tilde{y}^T \tilde{y}$ in terms of y:

$$\tilde{y}^T \tilde{y} = \sum_{i=1}^{n+d} \tilde{y}^i \tilde{y}^i = \sum_{i=1}^{n} y^i y^i + \sum_{i=n+1}^{n+d} 0 \cdot 0 \quad \left[\tilde{y}^i = y^i \text{ for } i = 1, \dots n; \ = 0 \text{ for } i = n+1, \dots n+d \right]$$

$$\Rightarrow \boxed{\tilde{y}^T \tilde{y} = y^T y} \quad \text{—— (5)}$$

$$\tilde{y}^T \tilde{x} = (y^1, y^2, \dots y^n, 0 \dots 0)\, c \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_d^1 \\ x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & \\ x_1^n & x_2^n & \cdots & x_d^n \\ \sqrt{\lambda_2} & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_2} \end{pmatrix}$$

$$= c\left(\sum_{i=1}^{n} y^i x_1^i, \sum_{i=1}^{n} y^i x_2^i, \dots, \sum_{i=1}^{n} y^i x_d^i \right)_{1\times d}$$

$$\Rightarrow \boxed{\tilde{y}^T \tilde{x} = c\, y^T x} \quad \text{—— (6)}$$

Tranposing this $\boxed{\tilde{x}^T y = c\, x^T y}$ —— (7)

Now $\tilde{X}^T \tilde{X}$

$$= c\begin{bmatrix} x_1^1 & \cdots & x_1^{d+n} \\ x_2^1 & \cdots & x_2^{n+d} \\ \vdots & & \vdots \\ x_d^1 & \cdots & x_d^{n+d} \end{bmatrix}_{d\times(n+d)} c\begin{bmatrix} x_1^1 & \cdots & x_d^1 \\ x_1^2 & \cdots & x_d^2 \\ \vdots & & \vdots \\ x_1^{n+d} & \cdots & x_d^{n+d} \end{bmatrix}_{(n\times d)\times d}$$

$$= c^2\begin{bmatrix} \sum_{i=1}^{n+d} x_1^i x_1^i & \cdots & \sum_{i=1}^{n+d} x_1^i x_d^i \\ \cdots & & \\ \sum_{i=1}^{n+d} x_d^i x_1^i & \cdots & \sum_{i=1}^{n+d} x_d^i x_d^i \end{bmatrix}_{d\times d}$$

where $x_i^{n+1}, x_i^{n+2}, \ldots x_i^{n+d}$ are all zero for $i=1,\ldots d$ except $x_i^{n+i}=1$

For diagonal component

$$\sum_{i=1}^{n+d} x_j^i x_j^i = \sum_{i=1}^{n} x_j^i x_j^i + \sum_{i=n+1}^{n+d} x_j^i x_j^i$$

$$= \sum_{i=1}^{n} x_j^i x_j^i + \lambda_2$$

For Non-diagonal entries

$$\sum_{i=1}^{n+d} x_j^i x_k^i \quad k, j = 1, \ldots d \text{ and } k \neq j$$

$$= \sum_{i=1}^{n} x_j^i x_k^i + \sum_{i=n+1}^{n+d} 0 \qquad x_k^i = 0 \text{ for } k = 1, 2, \ldots d, \; i = n+1 \ldots n+d$$

$$= \sum_{i=1}^{n} x_j^i x_k^i$$

So, $$\tilde{X}^T\tilde{X} = c^2 \begin{bmatrix} \sum_{i=1}^{n} x_1^i x_1^i + \lambda_2 & \cdots & \sum_{i=1}^{n} x_1^i x_d^i \\ & & \\ \sum_{i=1}^{n} x_d^i x_1^i & \cdots & \sum_{i=1}^{n} x_d^i x_d^i + \lambda_2 \end{bmatrix}$$

$$= c^2 (X^T X + \lambda_2 I) \quad \text{—— (8)}$$

Now we substitute result (5), (6), (7), (8) in (4).

we get

$$Z_2(W) = y^T y - c y^T X W - c W^T X^T y + W^T c^2 (X^T X + \lambda_2 I) W + c\lambda_1 |W|_1$$

$$= y^T y - c y^T X W - c W^T X^T y + c^2 W^T X^T X W + c^2 \lambda_2 W^T W + c\lambda_1 |W|_1$$

$$= Z_1(cW)$$

Hence proved that the elastic problem can be reduced to a lasso problem on modified data.