HW - 1

① Formulate and solve simplest case of linear regression

Consider N samples for X each of d dimension

$$X = \begin{bmatrix} 1 & x_1^1 & x_2^1 & x_3^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & & x_d^2 \\ \vdots & & & & & \\ 1 & x_1^N & x_2^N & \cdots & & x_d^N \end{bmatrix}_{N \times d+1}$$

and the labels Y be

$$Y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{bmatrix}_{N \times 1}$$

Consider the weights W

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}_{d+1 \times 1}$$

then we have the set of equations as

$$XW = \hat{Y}$$

where $\hat{Y}$ is the estimated labels for the given weights W.

Consider the cost function $E(W)$

$$E(W) = \sum_{i=1}^{N} \| y^{(i)} - \hat{y}^{(i)} \|^2 \quad = \text{sum of squared errors}$$

$$\Rightarrow E(W) = \| y - \hat{y} \|^2 \quad \| \ \| \text{ being the norm}$$

for optimal weights we have

$$\nabla E(W) = 0 \quad \left(\text{Gradient wrt } W \right)$$

$\Rightarrow \nabla \|y - \hat{y}\|^2 = 0$

$\Rightarrow \nabla \|Y - XW\|^2 = 0$

$$\Rightarrow \frac{\partial}{\partial w_j}\left(y^i - \sum_{j=0}^{d} x_j^i w_j\right)^2 = 0$$

$$\Rightarrow 2x_j\left(y^i - \sum_{j=0}^{d} x_j^i w_j\right) = 0$$

In matrix consolidated form, we have

$$2X^T(Y - XW^*) = 0 \qquad W^* = \text{optimal weights}$$

$$X^T Y = X^T X W^*$$

now, the first assumption is $X^T X$ is positive definite

$$\Rightarrow \boxed{W^* = (X^T X)^{-1} X^T Y.}$$

② ~~Bias~~ Basis function is a function ϕ which we apply on the matrix X to further continue and estimate weights.

So given data X, after applying bias function, we have

$$\cancel{\phi(X)} = \begin{bmatrix} \phi_0^1(x_0) & \phi_1^1(x_1^1) & \dots & \phi_M^1(x_d^1) \\ \vdots & \ddots & & \vdots \\ \phi_0^N & & & \phi_M(x_d^N) \end{bmatrix}$$

$$\phi(X) = \begin{bmatrix} \phi_0^1 & \phi_1^1 & \dots & \phi_M^1 \\ \phi_0^2 & \phi_1^2 & \dots & \phi_M^2 \\ \vdots & \ddots & & \vdots \\ \phi_0^N & \dots & \dots & \phi_M^N \end{bmatrix}$$

$N \times (M+1)$

Consider the weight matrix W

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}_{M+1 \times 1}$$

then we have

$$\hat{y} = \phi(x)\, W$$

for simplicity sake, drop x in $\phi(x)$

$$\hat{y} = \phi W$$

Considering cost function as SSE, we have

$$E(w) = \| y - \hat{y} \|^2$$

$$E(w) = \| y - \phi w \|^2$$

for optimal weights w^*

$$\nabla E(w) = 0$$

Similar to vanilla regression,

$\nabla E(w)$ doesn't depend on

X (or) in this case ϕ

hence

$\nabla E(w) = 0$

$\Rightarrow 2\phi^T(Y - \phi W^*) = 0$

$\Rightarrow \phi^T Y = \phi^T \phi w^*$

$$\boxed{W^* = (\phi^T \phi)^{-1} \phi^T Y}$$

(3) $\sigma(x) = \dfrac{1}{1+e^{-x}}$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}}$$

$$2\sigma(2x) = \frac{2}{1+e^{-2x}} \Rightarrow 2\sigma - 1 = \frac{2-1-e^{-2x}}{1+e^{-2x}}$$

$$\cancel{2\sigma(2x) + 1} = \cancel{\frac{2+1+e^{-2x}}{1+e^{-2x}}}$$

$$\boxed{2\sigma(2x) - 1 = \frac{1-e^{-2x}}{1+e^{-2x}}}$$

$$\Rightarrow \boxed{\tanh(x) = 2\sigma(2x) - 1.}$$

Consider

$$\hat{y}(x,w) = w_0 + \sum_{j=1}^{M} w_j \, \sigma\left(\frac{x-\mu_j}{s}\right)$$

$$\hat{y}(x,u) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x-\mu_j}{s}\right)$$

$$\tanh(x) = 2\sigma(2x) - 1.$$

$$\hat{y}(x,u) = u_0 + \sum_{j=1}^{M} u_j \left[2\sigma\left(2\left(\frac{x-\mu_j}{s}\right)\right) - 1\right]$$

$$= \underbrace{u_0 - \sum_{j=1}^{M} u_j}_{k_0} + \sum_{j=1}^{M} \underbrace{2u_j}_{k_1} \sigma\left(\frac{2x-2\mu_j}{s}\right)$$

$$\hat{y}(x,k) = k_0 + \sum_{j=1}^{M} k_j \, \sigma\left(\frac{2x-2\mu_j}{s}\right)$$

take $2x = z$ & $2\mu_j = \alpha_j$

$$\hat{y}(z, k) = k_0 + \sum_{j=1}^{M} k_j \, \sigma\left(\frac{z - \alpha_j}{s}\right)$$

which is equivalent to

$$\hat{y}(x, u)$$

& $\boxed{k_0 = w_0 = u_0 - \sum_{j=1}^{M} u_j}$

& $\boxed{k_j = 2u_j = w_j}$

$$\therefore \quad \boxed{w_i = \begin{cases} u_0 - \sum_{j=1}^{M} u_j & i = 0 \\ 2u_i & \text{else} \end{cases}}$$

(4) $E(W) = \sum_{i=1}^{N}\left(y^{(i)} - \sum_{k=1}^{K}\sum_{j=0}^{d} x_j^{(i)} w_{jk}\right)^2$

$\nabla E(w^*) = 0$

$\Rightarrow \dfrac{\partial}{\partial w_{jk}} E(w^*) = 0$

$\Rightarrow -\cancel{2} x_{(j)}^{(i)}\left(y^{(i)} - \sum_{k=1}^{K}\sum_{j=0}^{d} x_j^{(i)} w_{jk}^*\right) = 0$

in matrix form

$X^T(Y - X\vec{w}^*) = 0$

$\Rightarrow \boxed{W^* = (X^T X)^{-1} X^T Y}$

So for (1) the W^* has a dimension of $\boxed{d+1 \times K.}$

for (2) $W^* = (\phi^T \phi)^{-1} \phi^T Y$

& W^* has the order $\boxed{m+1 \times K}$

⑤ $E(w) = \sum_{i=1}^{N} r_i \left(y^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_j \right)^2.$

$$= \sum_{i=1}^{N} \left(Y^{(i)} - \sum_{j=0}^{d} X_j^{(i)} w_j \right)^2$$

where $Y^{(i)} = \sqrt{r_i}\, y^{(i)}$

$X_j^{(i)} = \sqrt{r_i}\, x_j^{(i)}$

for w^*, $\nabla E(w^*) = 0$

$$\Rightarrow 2 X^T (Y - X w^*) = 0$$

$$\Rightarrow \boxed{w^* = (X^T X)^{-1} X^T Y}$$

where $X = R x$

$Y = R y$

$$(x^T R^2 x)^{-1} x^T R^2 y$$

$$R = \begin{bmatrix} \sqrt{r_1^*} & 0 & \cdots & 0 \\ 0 & \sqrt{r_2^*} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \sqrt{r_N^*} \end{bmatrix}.$$

$$Rx = \begin{bmatrix} \sqrt{r_1^*} & \sqrt{r_1^*}\, x_1^1 & \sqrt{r_1^*}\, x_2^1 & \cdots \\ \sqrt{r_2^*} & \sqrt{r_2^*}\, x_1^2 & \sqrt{r_2^*}\, x_2^2 & \cdots \\ \vdots & & & \ddots \\ \vdots & & & \end{bmatrix}$$

$$Ry = \begin{bmatrix} \sqrt{r_1^*}\, y^1 \\ \sqrt{r_2^*}\, y^2 \\ \vdots \\ \sqrt{r_N^*}\, y^N \end{bmatrix}$$

(6) $E(w) = \|y - Xw\|^2 + \lambda w^T w$

where $\|y - Xw\|^2$ is the standard SSE

and $\lambda w^T w$ is the l_2 norm &

λ is the Lagrangian Multiplier

$$\nabla E(w^*) = 0 \Rightarrow -2X^T(Y - Xw) + 2\lambda I w = 0$$

$$\cancel{2}X^T Y = \cancel{2}X^T X w^* + \cancel{2}\lambda I w^*$$

$$\Rightarrow w^* = (X^T X + \lambda I)^{-1} X^T Y$$

$$\boxed{w^* = (X^T X + \lambda I)^{-1} X^T Y}$$

Regularization is applied when

→ To prevent over fitting of data

→ When data is noisy

(7) Let $\mathcal{N} \sim (0, \sigma^2)$ be added to x

$$X = x + \mathcal{N}$$

If x = matrix $N \times d+1$

$\mathcal{N}$ = matrix of $N \times d+1$

where $\mathcal{N}^{(i)} \sim (0, d)$

Now consider SSE cost function for this altered data

$C_F(X, W)$ = Cost function on X and W.

$$C_F(X,W) = \sum_{i=1}^{N} \left(\left(y^i - \sum_{j=0}^{d} X_j^i w_j \right)^2 \right)$$

$$= \sum_{j=\phi}^{N} \left(\left(y^i - \sum_{j=0}^{d} \left((x_j^i + N_j^i)(w_j) \right) \right)^2 \right)$$

$$= \sum_{i=1}^{N} \left(\left(\left(y^i - \sum_{j=0}^{d} x_j^i w_j \right) - \left\{ \sum_{j=0}^{d} N_j^i w_j \right) \right)^2 \right)$$

Let $y^i - \sum_{j=0}^{d} x_j^i w_j = \alpha$

$\sum_{j=0}^{d} N_j^i w_j = \beta$

$$C_F(X,W) = \sum_{i=1}^{N} \left((\alpha - \beta)^2 \right)$$

$$C_F(X,W) = \sum_{i=1}^{N} \left(\alpha^2 + \beta^2 - 2\alpha\beta \right)$$

Now $\sum_{i=1}^{N} (R \cdot V) = N\,\mathbb{E}(R \cdot V)$

where $R \cdot V$ is a random variable

$\boxed{\mathbb{E}_i \Rightarrow \text{expectancy over iterations of } i}$

$$\Rightarrow C_F(X,W) = N\,\mathbb{E}_i\left(\alpha^2 + \beta^2 - 2\alpha\beta \right)$$

$$= N\,\mathbb{E}(\alpha^2) + N\,\mathbb{E}(\beta^2) - 2N\,\mathbb{E}(\alpha\beta)$$

note that $N\,\mathbb{E}(\alpha^2) = C_F(x,W)$

i.e., $N\,\mathbb{E}(\alpha^2)$ = Cost function of data without gaussian noise

$$C_F(X,w) = N\,\mathbb{E}\Big(\big(y^i - \sum_{j=0}^{d} x_j^i w_j\big)^2\Big)$$

$$+ N\,\mathbb{E}\Big(\big(\sum_{j=0}^{d} N_j^i w_j\big)^2\Big) \cancel{- 2N}$$

$$-2N\,\mathbb{E}\Big(\big(y^i - \sum_{j=0}^{d} x_j^i w_j\big)\big(\sum_{j=0}^{d} N_j^i w_j\big)\Big)$$

α is basically a constant

$$\Rightarrow \quad \mathbb{E}(\alpha\beta) = \alpha\,\mathbb{E}(\beta)$$

$$\& \ \mathbb{E}(\beta) = \mathbb{E}\Big(\sum_{j=0}^{d} N_j^i w_j\Big)$$

$$= \sum_{j=0}^{d} w_j\,\mathbb{E}(N_j^i) = 0$$

~~$C_F(X,w)$~~ and

$$\mathbb{E}\Big(\big(\sum_{j=0}^{d} N_j^i w_j\big)^2\Big) = \|w\|^2\,\mathbb{E}\big((N_j^i)^2\big)$$

$$= \sigma^2 \|w\|^2$$

$$= \sigma^2 w^T w.$$

(By expanding)

$$\Rightarrow C_F(x, w) = C_F(x, w) + \sigma^{-2} w^T w$$

$$= \sum_{i=1}^{N} \left(\left(y^i - \sum_{j=0}^{d} x_j^i w_j \right)^2 \right) + \lambda w^T w$$

$\therefore$ $C_F(x, w)$ is equivalent to l2 regularized cost function of $C_F(x, w)$

(8) Consider the maximum aposterior probability expression

$$p(W / X, Y, \alpha, \beta)$$

We need to maximize this to get optimal weights

but,

$$p(W / X, Y, \alpha, \beta) \propto p(Y / X, W, \alpha, \beta) \cdot p(W / \alpha)$$

from Bayes theorem

and

$$p(Y / X, W, \alpha, \beta) \sim \mathcal{N}(\hat{Y}, \sigma^2 I)$$

$$p(W / \alpha) \sim \mathcal{N}(0, \alpha^2 I)$$

Max aposterior prob $\propto \frac{1}{(2\pi)^{p/2}\sigma^{p}} \exp\left(\frac{-(y-\hat{y})^T(y-\hat{y})}{2\sigma^2}\right)\frac{1}{(2\pi)^{p/2}\alpha^{p}}$

$$\exp\left(\frac{-W^T W}{2\alpha^2}\right)$$

these expressions follow from PDF's of a Gaussian Random Vector PDF with independent entries.

Max Apost Prob $\propto \frac{1}{(2\pi)^{p}\sigma^{p}\alpha^{p}} \exp\left(-\left[\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2}+\frac{W^T W}{2\alpha^2}\right]\right)$

Max aposterior prob $= K \exp\left(-\left[\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2}+\frac{W^T W}{2\alpha^2}\right]\right)$

for maximum aposterior probability gradient w.r.t W must be zero.

$$\Rightarrow \nabla\left(K \exp\left(-\left[\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2} + \frac{w^T w}{2\alpha^2}\right]\right)\right) = 0$$

$$\Rightarrow \cancel{K \exp\left(-\left[\right.\right.}$$

Let $\dfrac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2} + \dfrac{w^T w}{2\alpha^2} = CF$

$$\nabla K \exp(-CF) = 0$$

$$\Rightarrow K \exp(-CF) \cdot (-\nabla(CF)) = 0$$

$$\Rightarrow \nabla(CF) = 0$$

$$\Rightarrow \nabla\left(\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2} + \frac{w^T w}{2\alpha^2}\right) = 0$$

$$\Rightarrow \nabla\left((y-\hat{y})^T(y-\hat{y}) + \left(\frac{\sigma^2}{\alpha^2}\right) w^T w\right) = 0$$

$$\Rightarrow \nabla\left((y-\hat{y})^T(y-\hat{y}) + \lambda w^T w\right) = 0$$

this is similar to l2 normalized cost function for ridge regression

$$\Rightarrow \boxed{\lambda = \frac{\sigma^2}{\alpha^2}}$$