HW-1

① Formulate and solve simplest case of linear regression

Consider N samples for X each of d dimensi

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

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 $E(W)$

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

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

for optimal weights we have

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

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

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

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

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

In matrix consolidated form, we have

$$2 X^T (Y - X W^*) = 0$$

W^* = 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

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

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

$N \times (m+1)$

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

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

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

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

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

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

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

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

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

$$\text{if} \quad \boxed{k_j = 2u_j = w_j}$$

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

(4) Consider the SSE cost function

$$CF = E(w) = \sum_{i=0}^{N} \left[y^{(i)} - \hat{y}^{(i)} \right]^2$$

for i^{th} row of y and k^{th} term (dimention)

$$CF_{ik} = \left(y_k^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_{jk} \right)^2$$

for optimal w^*

$$\frac{\partial CF_{ik}}{\partial w_{jk}} = 0 \Rightarrow -2 x_j^{(i)} \left(y_k^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_{jk} \right) = 0$$

$\Rightarrow$ In matrix format

$$X^T \left(Y - X w^* \right) = 0$$

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

In case of basis functions;

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

their orders being $d+1 \times K$ and

$(N+1) \times K$ respectively

(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 $X = Rx$

$Y = Ry$

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

$$R = \begin{bmatrix} \sqrt{\gamma_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\gamma_2} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \sqrt{\gamma_N} \end{bmatrix}.$$

$$Rx = \begin{bmatrix} \sqrt{\gamma_1} & \sqrt{\gamma_1}\, x_1' & \sqrt{\gamma_1}\, x_2' & \cdots \\ \sqrt{\gamma_2} & \sqrt{\gamma_2}\, x_1^2 & \sqrt{\gamma_2}\, x_2^2 & \cdots \\ \vdots & & & \end{bmatrix}$$

$$Ry = \begin{bmatrix} \sqrt{\gamma_1}\, y^1 \\ \sqrt{\gamma_2}\, y^2 \\ \vdots \\ \sqrt{\gamma_N}\, y^N \end{bmatrix}$$

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 l2 norm &

λ is the Lagrangian Multiplier

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

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

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

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

Regularization is applied when

→ To prevent overfitting 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 $\sum_{i=1}^{N} (R.V) = N\,\mathbb{E}(R.V)$

where R.V is a random variable

$\mathbb{E}_i \Rightarrow$ 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

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

$$+ N\mathbb{E}\left(\left(\sum_{j=0}^{d} N_j^i w_j\right)^2\right) =$$

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

α is basically a constant

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

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

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

and

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

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

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

(By expanding)

$$\Rightarrow C_F(X, w) = C_F(x, w) + \sigma^2 w^T w$$

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

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

⑧ Consider the maximum aposteriori 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 $\propto \frac{1}{(2\pi)^{p/2}\sigma^p} \exp\left(\frac{-(y-\hat{y})^T(y-\hat{y})}{2\sigma^2}\right) \frac{1}{(2\pi)^{p/2}\alpha^p}$

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

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

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

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

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

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

$$\Rightarrow \text{[illegible]}$$

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

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

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

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

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

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

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

this is similar to l2 normalized cost function for ridge regression

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