1) $SSE(\vec{W})$ = Cost function, the squared sum error

$$E(\vec{W}) = \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2.$$

where, $\hat{y} = f(\vec{x}^{(i)}, \vec{W})$

$$\Rightarrow E(\vec{W}) = (\vec{y} - X\vec{W})^T(\vec{y} - X\vec{W})$$

$$\vec{W}^* = \underset{\vec{W}}{\operatorname{argmin}}(E(\vec{W}))$$

$$\Rightarrow \nabla E(\vec{W}) = 0 \quad \text{—— ①}$$

$$E(\vec{W}) = \vec{Y}^T\vec{Y} - \vec{W}^T X^T \vec{Y} - \vec{Y}^T X \vec{W} + \vec{W}^T X^T X \vec{W}$$

$$\nabla E(\vec{W}) = \cancel{0 - 2X^T\vec{Y}} \; 0 - 2X^T\vec{Y} + 2X^T X \vec{W}$$

From ①,

$$-2X^T\vec{Y} + 2X^TX\vec{W}$$

$$\Rightarrow \boxed{\vec{W}^* = (X^TX)^{-1}X^TY}$$

$$\boxed{\hat{Y} = X\vec{W}^*}$$

Also, $\nabla^2 E(\vec{W})$ needs to be positive definite

$$\nabla^2 E(\vec{W}) = 2X^TX$$

2) Using Basis functions $\phi(x)$.

$$X = \begin{bmatrix} x_0^1 & x_1^1 & \cdots & x_d^1 \\ x_0^2 & & & \vdots \\ \vdots & & & \vdots \\ x_0^N & - & - & x_d^N \end{bmatrix}_{N \times (d+1)} \quad \vec{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}_{N \times 1}.$$

Let the basis function be

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

Here, ~~we are~~ $\phi_0^{(i)}(x) = 1$ for bias.

Squared Sum Error $\rightarrow SSE(\vec{w})$

$$SSE(\vec{w}) = |\vec{Y} - \hat{Y}|^2 = |\vec{Y} - \phi(X)\vec{w}|^2$$

Differentiating and equating to zero, we get

$$\nabla SSE(\vec{w}) = 0$$

$$\nabla |\vec{Y} - \phi(X)\vec{w}|^2 = 0$$

$$\Rightarrow \nabla |\vec{Y} - \phi(X)\vec{w}|^2$$

$$\Rightarrow \boxed{\vec{w} = (\phi^T(x)\phi(x))^{-1}\phi^T(x)\vec{Y}}$$

This can be solved similarly to the first problem.

3) Given,

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

We know that, ~~Tan~~

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

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

$$\boxed{\therefore\ \mathrm{Tanh}(x) = 2\sigma(2x) - 1}$$

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

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

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

Scanned by CamScanner

By comparision,

$$W_0 = U_0 - \sum_{j=1}^{M} U_j$$

$W_1 = U_1,\ W_2 = U_2 \ldots$

(or)

$$W_k = \begin{cases} U_0 - \sum_{j=1}^{M} U_j & \text{if } k=0 \\ U_k & \text{else} \end{cases}$$

This is true as the mean of the distribution is constant but only the varience is changing. Hence, W_j can be compared with U_j.

Problem number (7).

4) Since the Labels are multidimensional,

$$Y = \begin{bmatrix} y_1^1 & y_2^1 & \cdots & y_k^1 \\ y_1^2 & \cdots & & \\ \vdots & & & \\ y_1^N & \cdots & & y_k^N \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^N & \cdots & & x_d^N \end{bmatrix}$$

(a) $\cancel{E(w) = \|y - \hat{y}\|}$

$\cancel{E(w) = \sum_{i=0}^{N} \sum_{j=1}}$

$$E(w) = \sum_{i=0}^{N} \sum_{j=1}^{k} \left(y_j^i - \hat{y}_j^i\right)^2$$

$$E(w) = \sum_{i=1}^{N} \sum_{j=1}^{k} \left(y_j^i - \sum_{l=0}^{k} x_l^i w_i^l\right)^2 = |Y - XW|^2$$

Differentiating and equating to zero,

$\cancel{\frac{\partial E(w)}{\partial w} =}$

$$\nabla E(w) = 0 \Rightarrow \nabla |Y - XW|^2 = 0$$

$$\Rightarrow \nabla\left((Y - XW)^T (Y - XW)\right) = 0$$

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

~~Where~~ $\boxed{\text{'W' has dimension '}(d+1) \times k\text{'}}$

(b) Similarly,

$$\phi(X) = \begin{bmatrix} \phi_0^1(x) & \phi_1^1(x) & \cdots & \phi_d^1(x) \\ \vdots & & & \vdots \\ \phi_0^N(x) & \cdots & \cdots & \phi_d^N(x) \end{bmatrix}_{N \times (d+1)}$$

$$\overline{W} = (\phi^T \phi)^{-1} \cdot \phi^T Y$$ as the parameters 'X' are replaced with their basis functions '$\phi(X)$'.

5) Given,

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

$$E(\vec{W}) = \sum_{i=0}^{N} \left((\sqrt{r_i}\, y^{(i)}) - \sum_{j=0}^{d} (x_j^{(i)} \sqrt{r_i}) w_j \right)^2.$$

This form is similar to the Sum of squared Error (SSE) form where the labels and inputs are

$$\vec{Y}' = \begin{bmatrix} y^1 \sqrt{r_1} \\ y^2 \sqrt{r_2} \\ \vdots \\ y^N \sqrt{r_N} \end{bmatrix} = \begin{bmatrix} \sqrt{r_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{r_2} & \cdots & & 0 \\ \vdots & & & & \\ 0 & \cdots & & & \sqrt{r_N} \end{bmatrix} \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{bmatrix}$$

$$X' = \begin{bmatrix} \sqrt{r_1} & 0 & \cdots & 0 \\ 0 & \sqrt{r_2} & \cdots & 0 \\ 0 & \cdots & & \sqrt{r_N} \end{bmatrix} \begin{bmatrix} x_0^1 & x_1^1 & \cdots & x_d^1 \\ \vdots & & & \vdots \\ x_0^N & \cdots & \cdots & x_d^N \end{bmatrix}$$

Scanned by CamScanner

Let $R = \begin{bmatrix} \sqrt{r_1} & 0 & \cdots & 0 \\ 0 & \sqrt{r_2} & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & \cdots & \sqrt{r_N} \end{bmatrix}$

$\therefore E(\vec{w}) = |\vec{Y}' - X'w|^2$

where, $X' = RX$, $\vec{Y}' = R\vec{Y}$

Minimising the cost function,

$\nabla E(\vec{w}) = 0$

$\Rightarrow \nabla |\vec{Y}' - X'\vec{w}|^2 = 0$

$\Rightarrow \vec{w} = ((X')^T X')^{-1} (X')^T \vec{Y}'$

$= ((RX)^T (RX))^{-1} (RX)^T (RY)$

$$\boxed{\vec{w} = (X^T R^T R X)^{-1} (X^T R^T R Y)}$$

This is the optimum set of weights for a weighted sum squared error case.

Scanned by CamScanner

b) SSE cost function with L2 regulisaisation is

$$E(\vec{W}) = |\vec{Y} - X\vec{W}|^2 + \lambda W^T W$$

$\lambda \rightarrow$ Lagrangian parameter

~~Sol~~ Minimising the cost, we get

$$\nabla E(\vec{W}) = \nabla(|\vec{Y} - X\vec{W}|^2 + \lambda W^T W) = 0$$

$$\Rightarrow -2X^T(\vec{Y} - X\vec{W}) + 2\lambda I \vec{W} = 0$$

$$\Rightarrow \vec{W} = (X^T X + \lambda I)^{-1} \cdot X^T \vec{Y}$$, where

X is the set of features not including

Uses of regularisation:

Overfitting - It means the model has more parameters than can be justified by the data. This results in the models capturing the irregularities and the noise in the data. Simply put, an overfitted model is too good to be true.

→ One way to overcome overfitting is regularisation. This significantly reduces the variance of the model without ~~affecting~~ ~~the higher bias~~ substantial increase in the bias.

→ By increasing the λ, it prevents overfitting but increasing it beyond a certain value, results in ~~improper~~ higher error rates.

1) Given,

$x' = x + \text{Noise}$

$E(N) = 0,\ E(N^2) = \sigma^2$

$$SSE(\vec{w}) = \sum_{i=1}^{N}\left[y^i - \sum_{j=1}^{d} x_j^i w_j\right]^2,\ \text{here,}$$

Adding the noise, we get

$$SSE'(\vec{w}) = \sum_{i=1}^{N}\left[y^i - \sum_{j=1}^{d}(x_j^i + N_j^i)w_j\right]^2$$

$$= \sum_{i=1}^{N}\Big[\underbrace{y^i - \sum_{j=1}^{d} x_j^i w_j}_{a} - \underbrace{\sum_{j=1}^{d} N_j^i w_j}_{b}\Big]^2$$

$$E[SSE'(\vec{w})] = \sum_{i=1}^{N} E[a^2] + E[b^2] - E[2ab]$$

$$\boxed{NE[a^2] = E[SSE(\vec{w})]}$$

$$E[b^2] = E\left[\left(\sum_{j=1}^{d} N_j^i w_j\right)^2\right]$$

$$= .E\Big[(N_1^1 w_1)^2 + (N_2^1 w_2)^2 + \dots (N_d^1 w_d)^2 + (N_1^1 w_1 N_2^1 w_2) + \dots\Big]$$

$\longrightarrow$ 0

$$\boxed{E[N_i N_j] = \delta_{ij}\sigma^2}$$

$$= .E\left[(N_1^1 w_1)^2 + \dots [N_d^1 w_d]^2\right]$$

$$= \left[E[(N_1^1 w_1)^2] + E[(N_2^1 w_2)^2] \dots + E[(N_d^1 w_d)^2]\right]$$

$$\boxed{\sigma^2\left[\sum_{j=1}^{d} W_j^2\right] = E\left[b^2\right]}$$

$$\cancel{E\left[2ab\right] = E\left[2 \ldots\right]}$$

$$E\left(2ab\right) = E\left[2Err(\vec{W})\left(\sum_{j=1}^{d} N_j^i W_j\right)\right]$$

Here, $Err(\vec{W})$ is constant, W_j are constant

$$E\left[2ab\right] = 2Err(\vec{W})\, E\left[N_1' W_1 + N_2' W_2 \ldots + N_d' W_d\right]$$

$$= 2Err(\vec{W})\left[E(N_1')W_1 + E\left[N_2'\right]W_2 \ldots\right]$$

$$= 0 \qquad \boxed{\because E\left[Noise\right] = 0}$$

$$\therefore E\left[SSE'(\vec{W})\right] = E\left[SSE(\vec{W})\right] + \sigma^2\left[\sum_{j=0}^{d} W_j^2\right]$$

$$= |Y - X\vec{W}|^2 + \sigma^2 W^T W$$

This is simalato L2 norm where

$$\boxed{\lambda = \sigma^2}$$

Scanned by CamScanner

8) Maximum aposterior expression

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

↳ from Bayes theorem

~~We know~~ Given,

$$\vec{Y} \sim \mathcal{N}(X\vec{W}, \sigma^2 I)$$

$$\Rightarrow P(Y/X, W, \alpha, \beta) \sim \prod_{i=1}^{N} \mathcal{N}(\hat{Y}, \sigma^2 I)$$

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

$$P(W/X,Y,\alpha,\beta) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(\frac{-(y-\hat{y})^T(y-\hat{y})}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^d$$

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

$$\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N+d} \exp\left(-\left(\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2} + \frac{W^TW}{2\alpha^2}\right)\right)$$

$$\Rightarrow P(W/X,Y,\alpha,\beta) = C\exp\left(-\left[\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2} + \frac{W^TW}{2\alpha^2}\right]\right)$$

for Maximising $P(W/X,Y,\alpha,\beta)$, since the exponent is negative because maximising the power increases the exponent.

$$\underset{W}{\text{ArgMin}}\left(\frac{(y-\hat{y})^T(y-\hat{y})}{2\sigma^2} + \frac{W^TW}{2\alpha^2}\right) = \underset{W}{\text{Argmax}}\, P(W/X,Y,\alpha,\beta)$$

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

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

This is similar to the ridge regression

where $\lambda = \frac{\sigma^2}{\alpha^2}$

Scanned by CamScanner