ASST – I

Theory :-

1) Say

$$\hat{y} = f(x; \vec{w})$$

where $\vec{x} = [1 \;\; x_1 \;\; x_2 \;\; \cdots \;\; x_d]$

& $\vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$ be the ~~best~~

relation between y & x

We try to minimize the error

$$\sum_{i=1}^{N} (y^i - \hat{y}^i)^2$$ with respect to

$\vec{w}$ so that we can find the best fit for $\hat{y} = f(x; \vec{w})$

Let the optimal $\vec{w}$ be $\vec{w}^*$.

& the matrix of inputs be

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

where $N =$ no. of training samples
and d = dimension of input data (no. of parameters)

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

where $\boxed{\hat{y}^{(i)} = \sum_{j=0}^{d} x_j^{(i)} w_j}$

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

$$\Rightarrow \frac{\partial L(w)}{\partial w_k} = -2 \left\{ \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_j\right)\left(x_k^{(i)}\right) \right\} = 0$$

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

$$\Rightarrow \sum_{i=1}^{N} \cancel{x_k} (X^T)_{ki} \cdot (y - Xw)_i = 0$$

$$\Rightarrow X^T (y - Xw) = 0$$

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

$\therefore \vec{w}^*$ minimizes $L(w)$ if $X^T X$ is positive definite

---

2.) $\hat{y}^{(i)} = \sum_{j=0}^{M} \phi_j^{(i)} \cdot w_j$

where $\phi_j^{(i)}$ represents $\phi(x_j^{(i)})$, ϕ being basis function.

i.e., $\Phi = \begin{bmatrix} \phi_0(x^{(1)}) & \phi_1(x^{(1)}) & \phi_2(x^{(1)}) & \cdots & \phi_M(x^{(1)}) \\ \phi_0(x^{(2)}) & \phi_1(x^{(2)}) & \phi_2(x^{(2)}) & \cdots & \\ \vdots & & & & \\ \phi_0(x^{(N)}) & \cdots & & & \phi_M(x^{(N)}) \end{bmatrix}$

$N \times (M+1)$

$$L(w) = \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=0}^{M} \phi_j^{(i)} \cdot w_j \right)^2$$

$$\frac{\partial L(w)}{\partial w_k} = 0 \Rightarrow \sum_{i=1}^{N} \phi_k^{(i)} \left(y^i - \sum_{j=0}^{M} \phi_j^{(i)} \cdot w_j \right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} (\Phi^T)_{ki} (y - \Phi w)_i = 0$$

$$\Rightarrow \phi^T y = (\phi^T \phi)\, w$$

$$\Rightarrow \boxed{\vec{w}^{*} = (\phi^T \phi)^{-1} \cdot \phi^T y}$$

3.)

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

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

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

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

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

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

Scanned by CamScanner

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

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

$$- \sum_{j=1}^{M} u_j$$

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

Comparing this with

$\hat{y}(x, w)$., (also $2(x-\mu_i) = \cancel{y} \; x' - \mu_i'$)

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

$$\& \quad \boxed{w_i = 2u_i}$$

4.) This a similar argument as in ①, ② except

$\cancel{x_{ij}\, w_{jk}}$

$$\boxed{\hat{y}_{ik} = \sum_{j=0}^{d} x_{ij}\, w_{jk}} \Rightarrow \hat{Y}_{N\times k} = X_{N\times(d+1)} W_{(d+1)\times k}$$

The weight matrix will be a two dimensional matrix (dimensions :- $(d+1)\times k$)

$$L(w) = \sum_{i=1}^{K} \sum_{i=1}^{N} \left(y_{ik} - f_k(x_i)\right)^2$$

$$\cancel{= tr\left[(Y - XW)^T (Y - XW)\right]}$$

$$= tr\left[(Y - XW)^T (Y - XW)\right]$$

Minimizing this is similar to minimize individual terms as seen in prob ①, prob ② except Y is a $N\times k$ dimensional matrix

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

Scanned by CamScanner

W^* being optimal W

(for basis functions, $X = \phi$,

ϕ is the same matrix as in prob ②.

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

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

$r_i > 0$

$$\cancel{\frac{\partial L(w)}{\partial w_k} = 0}$$

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

$$\cancel{\Rightarrow \sum_{i=1}^{N} x_k^{(i)} \left(r^i y^i - r^i \sum_{j=0}^{d} x_j^{(i)} w_j \right) = 0}$$

$$\cancel{\Rightarrow \sum_{i=1}}$$

$$\cancel{\frac{\partial L(w)}{\partial}}$$

$$L(w) = \sum_{i=1}^{N} \left(\sqrt{r_i}\, y_i - \sqrt{r_i} \sum_{j=0}^{d} x_j^{(i)} w_j \right)^2$$

Let $y_i' = \sqrt{r_i}\, y_i$ & $(x_j^{(i)})' = \sqrt{r_i} \cdot x_j^{(i)}$

$$\Rightarrow L(w) = \sum_{i=1}^{N} \left(y_i' - \sum_{j=0}^{d} (x_j^{(i)})' w_j \right)^2$$

Clearly from prob ①, ②

the $\vec{w}^*$ which minimizes above function is

$$\vec{w}^* = \left((X')^T (X') \right)^{-1} (X')^T \cdot y'$$

where

$$X' = \cancel{R_{i \times i} \cdot X_{i \times j}} \; R_{N \times N} \cdot X_{N \times (d+1)}$$

& $$y' = \cancel{R_{i \times i} \; y_{i \times 1}} \; R_{N \times N} \cdot y_{N \times 1}$$

$$\cancel{R_{i \times i} = \begin{bmatrix} \sqrt{r_i} & & \\ & & \\ & & \end{bmatrix}}$$

$$R_{N \times N} = \begin{bmatrix} \sqrt{r_1} & 0 & 0 & 0 & \cdots \\ 0 & \sqrt{r_2} & 0 & 0 & \cdots \\ 0 & 0 & \sqrt{r_3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$\therefore \boxed{\vec{w}^* = (X^T \cdot R \cdot R \cdot X)^{-1} \cdot X^T \cdot R \cdot R \cdot y}$$

$$= (X^T (R^2) X)^{-1} \cdot X^T (R^2) y$$

$$R^2 = \begin{bmatrix} r_1 & 0 & 0 & 0 & \cdots \\ 0 & r_2 & 0 & 0 & \cdots \\ 0 & 0 & r_3 & 0 & \cdots \\ \vdots \end{bmatrix}_{N \times N}$$

6.)

$$E(\vec{w}) = (\vec{y} - X\vec{w})^T (\vec{y} - X\vec{w}) + \lambda \underbrace{\vec{w}^T \cdot \vec{w}}_{\text{regularizer}}$$

$$\nabla_w E(\vec{w}) = 2\lambda w - 2X^T y + 2X^T X w = 0$$

$$\Rightarrow (X^T X + \lambda I) w = X^T y$$

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

~~where X =~~

where $X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & \cdots & & \\ \vdots & \vdots & & \end{pmatrix}_{N \times d}$

& $y = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{pmatrix}_{N \times 1}$

Use of regularization :-

> Clearly $X^T X + \lambda I$ adds a non-zero term to ~~column~~ diagonals of $X^T X$

this makes the matrix $X^T X + \lambda I$ invertible for any matrix X.

> Also, this prevents weights from taking extreme values and reduces the high variance of model (better generalization) by constraining the norm of weight vector.

---

7. Here $x_i' = x_i + n_i$ is a random variable

$\Rightarrow \vec{x'}$ is a random vector

$\Rightarrow \vec{x'} \cdot \vec{w}$ is a random variable

and so ~~$y - \vec{x'} \cdot \vec{w}$~~ is

$y - w_0 - \vec{x'} \cdot \vec{w}$

$\boxed{\text{the dimension of } \vec{X}, \vec{N} \text{ is } 1 \times d}$ & ~~[illegible]~~

The loss function

$L(w)$ is given by

$$L(w) = E\left(\left(y - w_0 - \vec{x'}\cdot\vec{w}\right)^2\right)$$

$$= E\left(\left(y - w_0 - \vec{x}\cdot\vec{w} - \vec{N}\cdot\vec{w}\right)^2\right)$$

$$= E\left(\left(y - w_0 - \vec{x}\cdot\vec{w}\right)^2 - 2\left(y - w_0 - \vec{x}\cdot\vec{w}\right)\vec{N}\cdot\vec{w} + \left(\vec{N}\cdot\vec{w}\right)^2\right)$$

where $\vec{N}$ is a random vector where each $n_i \sim N(0, \sigma^2)$ & all are independent i.e., $E(n_i n_j) = 0$.

Same as normal regression

$$\Rightarrow L(w) = E\left(\left(y - w_0 - \vec{x}\cdot\vec{w}\right)^2\right)$$

~~constant~~ not random.

$$- 2\left(y - w_0 - \vec{x}\cdot\vec{w}\right)\cdot E\left(\vec{N}\cdot\vec{w}\right)$$

$$+ E\left(\left(\vec{N}\cdot\vec{w}\right)^2\right)$$

$$\boxed{E\left(\vec{N}\cdot\vec{w}\right) = E\left(\sum_{i=1}^{d} n_i w_i\right) = 0}$$

$$\boxed{\left(\because E(n_i) = 0\right)}$$

$$E\left((\vec{N}\cdot\vec{w})^2\right)$$

$$= E\left(\left(\sum_{i=1}^{d} n_i w_i\right)^2\right)$$

$$= E\left((n_1 w_1 + n_2 w_2 + \cdots + n_d w_d)^2\right)$$

$$= E\left(\sum_{i=1}^{d} n_i^2 w_i^2 + K'\right)$$

Here K' is $2\sum_{j=1}^{d}\sum_{i=1}^{j-1} n_i n_j w_i w_j$

$$\Rightarrow E(K') = 0$$

$$(\because E(n_i n_j) = 0)$$

$$\Rightarrow E\left((\vec{N}\cdot\vec{w})^2\right) = \sum_{i=1}^{d} E(n_i^2)\cdot w_i^2$$

$$= (\sigma^2)\,\|w_i\|^2$$

$$\therefore L(w) = E\left((y - w_0 - Xw)^2\right) + \sigma^2 \|w_i\|^2,$$

$$\|w\|^2 = \sum_{i=1}^{d} w_i^2$$

which is same as ridge regression loss function with $\boxed{\lambda = \sigma^2}$.

| Hence proved |

* (Adding noise to input is same as regularisation)

---

8. $p(w) \sim \mathcal{N}(0, \alpha^2 I)$

& $p(\vec{y} \mid w, X) \sim \mathcal{N}(\vec{X}\vec{w}, \sigma^2 I)$

we find $\vec{w}$ maximize

$p(\vec{w} \mid X, y)$

8.) $p(w) \sim N(0, \alpha^2 I)$

we find w such that

~~$\Rightarrow p(\vec{w}) = \frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{\|w\|^2}{2\alpha^2}}$~~

$p(w \mid X, y)$ is maximized.

we know that

$$p(\vec{y} \mid X, w) \sim N(Xw, \sigma^2 I)$$

& also $y_i \in \vec{y}$ are independent.

$$\Rightarrow \boxed{p(\vec{y} \mid X, \vec{w}) = \prod_{i=1}^{N} \frac{1}{(\sqrt{2\pi})\sigma} e^{-\frac{\left(y^{(i)} - \sum_{j=0}^{d} x_j^i w_j\right)^2}{2\sigma^2}}}$$

where d - dimension of features

& N - No. of train samples.

$$p(\vec{w} \mid X, \vec{y}) \propto p(\vec{y} \mid X, \vec{w}) \cdot p(w)$$

(According to Baye's theorem)

~~$p(\vec{w} \mid X, \vec{y})$~~

$$p(y)\cdot p(\omega|y) = \cancel{p(\omega, y)} \quad p(y|\omega)\cdot p(\omega)$$

$$\Rightarrow p(\vec{\omega}|X,\vec{y}) \propto p(\vec{y}|X,\vec{\omega}) \cdot p(\vec{\omega})$$

$$\Rightarrow p(\vec{\omega}|X,\vec{y}) \propto \left(\prod_{i=1}^{N} \frac{1}{(\sqrt{2\pi})\,\sigma} e^{\frac{(y^i - (X\omega)^i)}{2\sigma^2}}\right) \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\,\alpha} e^{-\frac{\omega_j^2}{2\alpha^2}}$$

$$\Rightarrow p(\omega|X,y) \propto \frac{1}{2\pi\,\sigma^N \cdot \alpha^d} e^{-\left(\frac{\sum_{i=1}^{N}(y^i - (X\omega)^i)^2}{2\sigma^2} + \frac{\omega^2}{2\alpha^2}\right)}$$

$$\Rightarrow p(\vec{\omega}|X,\vec{y}) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{y^i - (X\omega)^i}{2\sigma^2}\right)}\right] \times \left[\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\,\alpha} e^{-\frac{\omega_j^2}{2\alpha^2}}\right]$$

$$\Rightarrow p(\vec{\omega}|X,\vec{y}) \propto \underbrace{\frac{1}{(2\pi)^{N/2}\cdot\sigma^N\cdot\alpha^d\cdot(2\pi)^{d/2}}}_{K} \cdot e^{-\left(\frac{\|\vec{y} - X\vec{\omega}\|^2}{2\sigma^2} + \frac{\|\vec{\omega}\|^2}{2\alpha^2}\right)}$$

$\Rightarrow$ maximizing $p(\vec{\omega}|X,\vec{y})$ is same as minimizing the mean which is

$$L(\vec{w}) = \frac{\|\vec{y} - X\vec{w}\|^2}{2\sigma^2} + \frac{1}{2\alpha^2}\left(\|\vec{w}\|^2\right)$$

~~Scalar~~

$$= \frac{1}{2}\left(\frac{\|\vec{y} - X\vec{w}\|^2}{\sigma^2} + \frac{\|\vec{w}\|^2}{\alpha^2}\right)$$

$$= \frac{1}{2\sigma^2}\left(\|\vec{y} - X\vec{w}\|^2 + \left(\frac{\sigma}{\alpha}\right)^2 \|\vec{w}\|^2\right)$$

Scalar

This is same as ridge regression loss function with

$$\boxed{\lambda = \left(\frac{\sigma}{\alpha}\right)^2}$$

It can also be observed that this loss function is actually the mean of the posterior distribution $P(\vec{w} \mid X, \vec{y})$ with $\boxed{\lambda = \left(\frac{\sigma}{\alpha}\right)^2}$

The same $L(w)$ is also the mode of the distribution (peak value in distribution)