EE5600 Introduction to AI & ML

① Theory

(1) Simplest case of Linear regression

Linear regression is where the expression (function) for finding the labels from input is linear in parameter and labels are from a continuous set.

First,

In Linear regression

$$\hat{\vec{y}} = X\vec{w}$$

where

$\hat{\vec{y}}$ - estimated labels vector from input matrix and weights

X - Matrix containing N input $\vec{x}_i$ vectors
$\{ i \in \mathbb{N} [1, N]$
$\mathbb{N}$ - Natural numbers $\}$

$\vec{w}$ - weights vector.

$$\hat{\vec{y}} = [\hat{y}^{(1)} \; \hat{y}^{(2)} \cdots \hat{y}^{(N)}]^T.$$

$$\vec{w} = [w_0 \; w_1 \cdots \cdots w_d]^T.$$

$$X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(N)} \\ \vdots & \vdots & & \\ x_d^{(1)} & x_d^{(2)} & \cdots & x_d^{(N)} \end{bmatrix}$$

where $\vec{x}^{(i)} \in \mathbb{R}^d$ and $\hat{y}^{(i)} \in \mathbb{R} \; \forall i$

To find the optimal weights vector, we use sum of squared error as cost function,

$$E(\vec{w}) = \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)$$

Here $y^{(i)}$ is ground truth label associated with $\vec{x}^{(i)}$, $\boxed{y^{(i)} \in \mathbb{R} \; \forall i}$

$E(\vec{w})$ is thus the error fn (SSE).

$$E(\vec{w}) = \sum_{i=1}^{N} \left[y^{(i)} - \left(\sum_{j=1}^{d} x_j^{(i)} w_j + w_0\right)\right]^2$$

In Matrix notation,

$$E(\vec{w}) = (\vec{y} - X\vec{w})^T(\vec{y} - X\vec{w}) \quad \text{--- (1)}$$

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

where $\vec{w}^*$ is the optimal weight vector.

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

from (1)

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

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

$$\nabla E(\vec{w}) = -2X^T\vec{y} + 2(X^TX)\vec{w}^* = 0$$

$$X^T\vec{y} = X^TX\vec{w}^*$$

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

By taking 2nd derivative, we get $2X^TX$, which should be positive definite for the cost function to be minimum.

→ which is possible if X has full column rank.

A Matrix is said to be positive definite, if for any $\vec{a}$

$$\boxed{\vec{a}^T X \vec{a} > 0}$$

The best estimate for Y, would be

$$\boxed{\hat{Y}_{best} = X\vec{w}^{*}}$$

where $\vec{w}^{*} = (X^T X)^{-1} X^T Y$

(2) <u>Using basis function with (M+1) Parameters</u>

$$\hat{y}^{(i)} = \sum_{j=0}^{M} \phi_j(\vec{x}^{(i)}) w_j \quad -①$$

where

$\phi_0(\vec{x}^{(i)}) = 1$ for all i in $[1, N]$ $i \in \mathbb{N}$ ← Natural number

we can represent ① in Matrix form

$$\hat{\vec{Y}} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(N)} \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & \phi_1^{(1)} & \cdots & \phi_M^{(1)} \\ 1 & \phi_1^{(2)} & & \phi_M^{(2)} \\ \vdots & \vdots & & \vdots \\ 1 & \phi_1^{(N)} & \cdots & \phi_M^{(N)} \end{bmatrix}$$

$$\vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}$$

<u>Cost function (SSE):-</u>

$$E(\vec{w}) = \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=0}^{M} \phi_j(\vec{x}^{(i)}) w_j\right)^2$$

In Matrix form,

$$E(\vec{\omega}) = (\vec{y} - \phi\vec{\omega})^T(\vec{y} - \phi\vec{\omega})$$

Set,

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

to find $\vec{\omega}^*$,

$$\boxed{\nabla E(\vec{\omega}) = -2\phi^T\vec{y} + 2\phi^T\phi\vec{\omega}}$$

setting

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

$$-2\phi^T\vec{y} + 2\phi^T\phi\vec{\omega}^* = 0$$

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

the best estimate

$$\boxed{\hat{y}_{best} = \phi\vec{\omega}^*}$$

(6) Linear ridge regression with L_2 Regularization

→ Firstly ~~be~~ we do this,

we subtract the mean from the ~~label~~ label data thus we can ignore 'c' bias term.

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_M^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_M^{(2)} \\ \vdots & \vdots & & \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_M^{(N)} \end{bmatrix}$$

$$\boxed{\vec{\hat{y}} = X\vec{\omega}}$$

$$\vec{\omega} = \begin{bmatrix} \omega_1 & \omega_2 & \cdots & \omega_M \end{bmatrix}^T$$

$$\vec{\hat{y}} = \begin{bmatrix} \hat{y}^{(1)} & \hat{y}^{(2)} & \cdots & \hat{y}^{(N)} \end{bmatrix}^T$$

where $\vec{\hat{y}}$ is estimated label vector from given inputs query

$$E(\vec{\omega}) = (\vec{y} - X\vec{\omega})^T(\vec{y} - X\vec{\omega}) + \lambda\vec{\omega}^T\omega$$

$$\nabla E(\vec{\omega}) = -2X^T(\vec{y} - X\vec{\omega}) + 2\lambda\vec{\omega}$$

Setting

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

$$2X^T(\vec{y} - X\vec{\omega}^*) = 2\lambda\vec{\omega}$$

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

The best estimate

$$\boxed{\vec{\hat{y}} = X\vec{\omega}^*}$$

Scanned by CamScanner

→ In most general cases No of observations(N) will be less than the total dimensions(d+1)

$N < (d+1)$

which leads to overfitting.

thus the model performs poorly on unseen data.

By using L2 regularization by trying to put a constraint on the squared norm of weights vector; along with SSE

we generate weights as close to origin as possible.

[Note :- We do this after subtracting the mean of labels from the each labels].

By varying λ, we can regulate the amount of shrinkage of weights.

→ A proper value of λ should chosen to ensure both less error and lesser squared norm.

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

and $r_i > 0$

Now,

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

In Matrix form

$$E(\vec{w}) = \left(\vec{y} - X\vec{w}\right)^T \left(R\vec{y} - RX\vec{w}\right)$$

where

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

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & & & & \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$$

$$\vec{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix}$$

$$E(\vec{w}) = (\vec{y}^T - \vec{w}^T X^T)(R\vec{y} - RX\vec{w})$$

$$= \vec{y}^T R \vec{y} - \vec{y}^T R X \vec{w} - \vec{w}^T X^T R \vec{y} + \vec{w}^T X^T R X \vec{w}$$

$$\nabla(E(\vec{w})) = -2X^T R \vec{y} + 2(X^T R X)\vec{w}^* = 0$$

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

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

(4) Linear regression

In multidimensional Labels

$$\boxed{\hat{y}_{ik} = \sum_{j=0}^{d} x_{ij} w_{jk}} \qquad \boxed{x_{i0} = 1 \ \forall i}$$

(or)

$$\boxed{\hat{y}^{(i)}_k = \sum_{j=0}^{d} x^{(i)}_j w_{jk}}$$

→ the weights matrix will be of dimension $(d+1) \times k$

→ $y^{(i)} \in R^k$

$$W = \begin{bmatrix} w_{01} & w_{02} & \cdots & w_{0k} \\ w_{11} & w_{12} & \cdots & w_{1k} \\ \vdots & \vdots & & \vdots \\ w_{d1} & \cdots & \cdots & w_{dk} \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_2^{(1)} & \cdots & \hat{y}_k^{(1)} \\ \hat{y}_1^{(2)} & \hat{y}_2^{(2)} & \cdots & \hat{y}_k^{(2)} \\ \vdots & & & \vdots \\ \hat{y}_1^{(N)} & \hat{y}_2^{(N)} & \cdots & \hat{y}_k^{(N)} \end{bmatrix}$$

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

$$\underset{N \times k}{\hat{Y}} = \underset{N \times d+1}{X} \cdot \underset{d+1 \times k}{W}$$

Applying the ~~R~~SSE cost function,

$$E(W) = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(y_{ik} - g_k(x_i)\right)^2$$

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

Minimising this is similar to minimising prob ① for each column of y matrix

$$= \sum_{i=1}^{N}\left[\left(y_1^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_{j1}\right)^2 + \dots + \left(y_K^{(i)} - \sum_{j=0}^{d} x_j^{(i)} w_{jK}\right)^2\right]$$

$$E(W) = \left(y_1^{(1)} - \sum_{j=0}^{d} x_j^{(1)} w_{j1}\right)^2 + \dots + \left(y_1^{(N)} - \sum_{j=0}^{d} x_j^{(N)} w_{j1}\right)^2 +$$

$$\left(y_2^{(1)} - \sum_{j=0}^{d} x_j^{(1)} w_{j2}\right)^2 + \dots + \left(y_2^{(N)} - \sum_{j=0}^{d} x_j^{(N)} w_{j2}\right)^2$$

$$+ \vdots$$

$$\left(y_K^{(1)} - \sum_{j=0}^{d} x_j^{(1)} w_{jK}\right)^2 + \dots + \left(y_K^{N} - \sum_{j=0}^{d} x_j^{(N)} w_{jK}\right)^2$$

$$\nabla E(W) = \begin{bmatrix} \frac{\partial E}{\partial w_{01}} & \frac{\partial E}{\partial w_{02}} & \dots & \frac{\partial E}{\partial w_{0K}} \\ \vdots & & & \\ \frac{\partial E}{\partial w_{d1}} & \frac{\partial E}{\partial w_{d2}} & \dots & \frac{\partial E}{\partial w_{dK}} \end{bmatrix}$$

$$= \begin{bmatrix} -2\sum_{k=1}^{N}\left[\left(y_1^{(k)} - \sum_{j=0}^{d} x_j^{(k)} w_{j1}\right) x_0^{(k)}\right] & \dots \\ \vdots & \vdots \\ & -2\sum_{k=1}^{N}\left[\left(y_K^{k} - \sum_{j=0}^{d} x_j^{(k)} w_j\right) x_d^{k}\right] \end{bmatrix}$$

$$= -2X^T[Y - XW]$$

Scanned by CamScanner

On taking the $\nabla(E(w)) = 0$ similar to prob
① we get

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

For Basis function, M+1 parameters.

For multidimensional Labels,

$$\boxed{\hat{y}_{ik} = \sum_{j=0}^{d} \phi_j(\vec{x}^{(i)}) \, w_{jk}}$$

(or)

$$\boxed{\hat{y}^{(i)}_{k} = \sum_{j=0}^{d} \phi_j(\vec{x}^{(i)}) \, w_{jk}}$$

The weights matrix is of dimension $(d+1) \times (k)$

→ $y^{(i)} \in \mathbb{R}^k$

$w, \hat{y}$ are having similar matrix notation as above

$$\Phi = \begin{bmatrix} 1 & \phi_1^{(1)} & \cdots & \phi_M^{(1)} \\ 1 & \phi_1^{(2)} & \cdots & \phi_M^{(2)} \\ \vdots & \vdots & & \\ 1 & \phi_1^{(N)} & \cdots & \phi_M^{(N)} \end{bmatrix}$$

Now

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

$$E(w) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - \phi_k(x_i))^2$$

$$= tr\left[(y - \phi w)^T (y - x\phi) \right]$$

minimising this is similar to minimising prob ② for each column of y matrix.

on taking $\nabla(E(w)) = 0$ similar to prob ②

we get

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

⑦ Now

$$X_{ij} = x_j^{(i)} + n_{ij} \rightarrow ①$$

(or)

$$\boxed{X_i = x^{(i)} + n_i} \quad \forall\ i = 1 \text{ to } N.$$

$$L(\vec{\omega}) = \sum_{i=1}^{N} \left(y^{(i)} - \omega_0 - \sum_{j=1}^{d} \hat{x}_j^{(i)} \omega_j \right)^2$$

$L(\vec{\omega})$ is cost function.

$$\tilde{x}_j = x_j^{(i)} + n_{ij}$$

$$L(\vec{\omega}) = \sum_{i=1}^{N} \left(y^{(i)} - \omega_0 - \sum_{j=1}^{d} x_j^{(i)} \omega_j - \sum_{j=1}^{d} n_{ij} \omega_j \right)^2$$

$$= \sum_{i=1}^{N} \left[\left(y^{(i)} - \omega_0 - \sum_{j=1}^{d} x_j^{(i)} \omega_j \right) - \left(\sum_{j=1}^{d} n_{ij} \omega_j \right) \right]^2$$

$$= \sum_{i=1}^{N} \left[\left(y^{(i)} - \omega_0 - \sum_{j=1}^{d} x_j^{(i)} \omega_j \right) \right]^2 + \sum_{i=1}^{N} \left(\sum_{j=1}^{d} n_{ij} \omega_j \right)^2$$

$$- 2 \sum_{i=1}^{N} \left[\left(y^{(i)} - \omega_0 - \sum_{j=1}^{d} x_j^{(i)} \omega_j \right) \left[\sum_{j=1}^{d} n_{ij} \omega_j \right] \right]$$

now

$$\sum_{i=1}^{N} f_i(x) = N E(f(x))$$

where E is expectation

$$L(\vec{\omega}) = N \left[E \left(y - \omega_0 - \sum_{j=1}^{d} x_j \omega_j \right)^2 \right] + N E \left[\sum_{j=1}^{d} n_j \omega_j \right]^2$$

$$- 2 N E \left(y - \omega_0 - \sum_{j=1}^{d} x_j \omega_j \right) E \left(\sum_{j=1}^{d} n_j \omega_j \right)$$

Taking termwise

$\rightarrow$
$$2NE\left(y-\omega_0-\sum_{j=1}^{d}\omega_j x_j\right)E\left[\sum_{j=1}^{d}n_j\omega_j\right]$$

$$=2NE\left(y-\omega_0-\sum_{j=1}^{d}\omega_j x_j\right)E\left(\vec{N}\cdot\vec{\omega}\right)$$

where $\vec{N}$ is a random vector where

each $n_j \sim N(0,\sigma^2)$ & all are

independent : $\boxed{E(n_i n_j)=E(n_i)\cdot E(n_j)=0}$ —(1)

$$=2N\left[E\left(y-\omega_0-\sum_{j=1}^{d}\omega_j x_j\right)\underline{E\left(\sum_{j=1}^{d}n_j\omega_j\right)}\right]$$

$$\boxed{E\left(n_1\omega_1+n_2\omega_2+\dots+n_d\omega_d\right)=E(n_1\omega_1)+\dots+E(n_d\omega_d) = 0+\dots+0=0}$$ (2)

$= 0$

$\rightarrow$
$$NE\left((\vec{N}\cdot\vec{\omega})^2\right)=NE\left(\left(\sum_{i=1}^{d}n_i\omega_i\right)^2\right)$$

$$=NE\left(\left(n_1\omega_1+n_2\omega_2+\dots+n_d\omega_d\right)^2\right)$$

$$=NE\left(\sum_{i=1}^{d}n_i^2\omega_i^2+2\sum_{j=1}^{d}\sum_{i=1}^{j-1}n_i n_j\omega_i\omega_j\right)$$

$$=NE\left(\sum_{i=1}^{d}n_i^2\omega_i^2\right)+\cancel{E\left(2\sum_{j=1}^{d}\sum_{i=1}^{j-1}n_i n_j\omega_i\omega_j\right)}\rightarrow 0$$

$$=N\sum_{i=1}^{d}E(n_i^2)\omega_i^2 \qquad \because E(n_i n_j)=E(n_i)E(n_j)=0$$

$$=N\sigma^2\sum_{j=1}^{d}\omega_j^2 = N\sigma^2\|\omega\|^2$$

Scanned by CamScanner

$\Rightarrow N\,E\left(\left(y - \omega_0 - \sum_{j=1}^{d} x_j \omega_j\right)^2\right)$ = cost fn for input without noise.

overall

$$L(\vec{\omega}) = \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=1}^{d} x_j^{(i)} \omega_j\right)^2 + N\sigma^2 \|\omega\|^2$$

$$= \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=1}^{d} x_j^{(i)} \omega_j\right)^2 + N\sigma^2 \omega^T \omega$$

→

(B) Given, $P(\vec{\omega} \mid x, y, \frac{1}{\sigma^2}, \frac{1}{\alpha^2})$ is Maximised if

$$\boxed{\vec{\omega}^* = \underset{\vec{\omega} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \left\{ P\left(\vec{\omega} \mid x, y, \frac{1}{\sigma^2}, \frac{1}{\alpha^2}\right) \right\}} \quad -(1)$$

Now Given information

$$P\left(\vec{\omega} \mid \frac{1}{\alpha^2}\right) \sim \mathcal{N}(0, \alpha^2 I)$$

$$P\left(\vec{y} \mid x, \vec{\omega}, \frac{1}{\sigma^2}\right) \sim \mathcal{N}(X\omega, \sigma^2 I)$$

$$P\left(\vec{\omega} \middle| x, \vec{\omega}, \frac{1}{\sigma^2}, \frac{1}{\alpha^2}\right) \propto P\left(\vec{y} \middle| x, \vec{\omega}, \frac{1}{\sigma^2}\right) \cdot P\left(\omega \middle| \frac{1}{\alpha^2}\right) \quad \text{--- (2)}$$

According to bayes theorem.

<u>Taking (1)</u> :-

(1) is maximised for if $P(\vec{y}/x,\omega,\frac{1}{\sigma^2}) \cdot P(\omega/\sigma^2)$ is maximised by (2).

Taking log on bothes of (2).

$$\log\left(P\left[\vec{y} \middle| x, \vec{\omega}, \frac{1}{\sigma^2}\right) \cdot P\left(\omega \middle| (1/\alpha^2)\right]\right)$$

$$= \log\left(P\left(\vec{y} \middle| x, \vec{\omega}, \frac{1}{\sigma^2}\right)\right) + \log P\left(\omega \middle| (1/\alpha^2)\right) \quad \text{--- (q.1)}$$

Now

Since $P\left(\vec{y} \middle| x, \omega, \frac{1}{\sigma^2}\right) \sim \mathcal{N}(X\vec{\omega}, \sigma^2 I)$

$\Rightarrow$ which says all elements of $\vec{y}$ given x, ω are <u>independent</u> since the covariance matrix is a <u>diagonal</u> matrix $(\sigma^2 I)$

Thus,

$$P\left(\vec{y}\,\middle|\,x,\omega,\frac{1}{\sigma^2}\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - (x\vec{\omega})_i)}{2\sigma^2}} \quad -③$$

Similarly,

$$P(\vec{\omega} \mid 1/\alpha^2) \sim N(0, \alpha^2 I)$$

Which says all elements of $\vec{\omega}$ are independ

Since the Covariance matrix is a diagonal

Matrix $(\alpha^2 I)$:

$$P\left(\vec{\omega}\,\middle|\,\left(\frac{1}{\alpha^2}\right)\right) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\,\alpha} e^{-(\omega_i)^2/2\alpha^2} \quad -④$$

Sub ③, ④ in Eq 1.

$$\vec{\omega}^* = \underset{\vec{\omega} \in \mathbb{R}^{d+1}}{\arg\max} \left\{ \log\left(P\left(\vec{y}\,\middle|\,x,\omega,1/\sigma^2\right)\right) + \log\left(P\left(\omega \mid 1/\alpha^2\right)\right)\right.$$

Thus we have to Maximise

$$f(\omega) = 2\times\left[\sum_{i=1}^{N}\left(\frac{-1}{2}\right)\log(2\pi)\right] - \sum_{i=1}^{N'}\left(\frac{1}{2}\right)\log(\sigma^2)$$

$$- \sum_{i=1}^{N} \frac{(y^{(i)} - (x\vec{\omega})_i)^2}{2\sigma^2} - \sum_{i=1}^{d}\frac{1}{2}\log(\alpha^2)$$

$$- \sum_{i=1}^{d} (\omega_i)^2 / 2\alpha^2$$

Maximising $f(\omega)$ means minimising

$$f(\omega) = \frac{\|\vec{y} - X\vec{\omega}\|}{2\sigma^2} + \frac{\|\omega\|^2}{\alpha^2}$$

$$\because \sum_{i=1}^{N} (y^{(i)} - (X\omega)_i)^2 = \|\vec{y} - X\vec{\omega}\|$$

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

$$\sum_{i=1}^{d} (\omega_i)^2 = \|\omega\|^2$$

$2\sigma^2 \downarrow$ scalar

This is same as ridge regression loss function with

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

It can also be seen that maximising the posterior distribution is minimising the L_2 regression of linear parameters or Ridge regression with $\boxed{\lambda = \left(\frac{\sigma}{\alpha}\right)^2}$

(3)

$$\tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{x}\left[1-e^{-2x}\right]}{e^{x}\left[1+e^{-2x}\right]} = \frac{1-e^{-2x}}{1+e^{-2x}}$$

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

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

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

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

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

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

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

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

comparing with

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

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

where $\boxed{t = S/2}$

thus it is equivalent to:

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

$$\boxed{w_i = 2u_j}$$