EE5600: Intro to AI and ML

1. Linear Regression:

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

~~(d+1) x N~~

$N \rightarrow$ number of training examples

$d \rightarrow$ length of input vector.

$$\bar{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}_{N \times 1} \rightarrow \text{outputs (labels)} \qquad \bar{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}_{(d+1) \times 1} \rightarrow \text{weights}$$

Cost function. $E(\bar{w}) = \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$

(sum of squared error)

Here $\hat{y}^{(i)}$ is predicted output.

$$\bar{\hat{y}} = X \cdot \bar{w} \rightarrow (N \times 1)$$

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

$$= (\bar{y} - X\bar{w})^T (\bar{y} - X\bar{w})$$

optimal weight, $\bar{w}^* = \underset{\bar{w}}{\arg\min}\left(E(\bar{w})\right)$

(weights that minimize the sum of squared error)

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

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

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

$$\Rightarrow \quad -2X^T(\bar{y} - X\bar{w}) = 0$$

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

To show that this is minimum,

check second derivative ~~of~~. $\Rightarrow 2X^TX$

This is positive definite (if X has full column rank)

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

2. Basis functions: $\phi_j(\bar{x}^{(i)}) \rightarrow 0 \le j \le M$

$\phi_0(\bar{x}) = 1$

$$\bar{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}_{(M+1)\times 1}$$

assuming $\phi_j(\cdot)$ are $\mathbb{R}^{(d+1)} \rightarrow \mathbb{R}$

Prediction, $\hat{y}^{(i)} = \sum_{j=0}^{M} \phi_j(\bar{x}^{(i)})\, w_j$

Cost function, $E(\bar{w}) = \sum_{i=1}^{N} \left[y^{(i)} - \sum_{j=0}^{m} \left(\phi_j(\bar{x}^{(i)}) w_j \right) \right]^2$

Notation:- $\phi_j(X) = \begin{bmatrix} \phi_j(\bar{x}^{(1)}) \\ \vdots \\ \phi_j(\bar{x}^{(N)}) \end{bmatrix}$

Notation

$\hookrightarrow \Phi = \Phi(X) = \begin{bmatrix} \phi_0(\bar{x}^{(1)}) & \phi_1(\bar{x}^{(1)}) & \cdots & \phi_m(\bar{x}^{(1)}) \\ \phi_0(\bar{x}^{(2)}) & \phi_1(\bar{x}^{(2)}) & \cdots & \phi_m(\bar{x}^{(2)}) \\ \vdots & \vdots & & \vdots \\ \phi_0(\bar{x}^{(N)}) & \phi_1(\bar{x}^{(N)}) & \cdots & \phi_m(\bar{x}^{(N)}) \end{bmatrix}$

$E(\bar{w}) = (\bar{y} - \Phi \bar{w})^T (\bar{y} - \Phi \bar{w})$

minimizing the cost function,

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

$\Rightarrow \quad -2 \Phi^T \bar{y} + 2 \Phi^T \Phi \bar{w} = 0$

$\Rightarrow \quad \boxed{\bar{w}^* = (\Phi^T \Phi)^{-1} \Phi^T \bar{y}}$

3. $\sigma(x) = \frac{1}{1+e^{-x}}$ $\qquad$ $\text{Tanh}(x) = \frac{e^{+x} - e^{-x}}{e^x + e^{-x}}$

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

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

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

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

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

Take $2x = z$, $2\underline{u} = \underline{v}$ $\Rightarrow$ $v_j = \frac{k_j}{2}$

$\Rightarrow$ ~~$\hat{y}\left(\frac{z}{2}, \frac{1}{2}\underline{v}\right)$~~ $\hat{y}(x, \underline{v}) = \frac{v_0}{2} + \sum_{j=1}^{M} v_j\, \sigma\left(\frac{x - k_j}{s}\right) - \frac{v_j}{2}$

k_j = mean of v_j

[z replaced by x]

$$w_0 = \frac{1}{2}\left(v_0 - \sum_{j=1}^{M} v_j\right) = u_0 - \sum_{j=0}^{M} u_j$$

$$(w_1, w_2, \cdots w_M) = (2u_1, 2u_2, \cdots 2u_m)$$

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

output, $$Y = \begin{bmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_k^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_k^{(2)} \\ \vdots & & & \vdots \\ y_1^{(N)} & y_2^{(N)} & \cdots & y_k^{(N)} \end{bmatrix}_{N \times k}$$

$\bar{x}^{(i)}$ → input vector, $\bar{y}^{(i)}$ → output vector, $1 \le i \le N$

Weights $W = \begin{bmatrix} w_0^{(1)} & w_0^{(2)} & \cdots & w_0^{(k)} \\ w_1^{(1)} & w_1^{(2)} & \cdots & w_1^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ w_d^{(1)} & w_d^{(2)} & \cdots & w_d^{(k)} \end{bmatrix}_{(d+1) \times k}$

Cost function $E(W) = \sum_{i=1}^{N} \| \bar{y}^{(i)} - \hat{\bar{y}}^{(i)} \|^2$

$\hat{\bar{y}}^{(i)}$ → predicted output vector.

$\hat{Y} = X \cdot W$

$\hat{\bar{y}}^{(i)} = \bar{x}^{(i)} W = \left[\sum_{p=0}^{d} x_p^{(i)} w_p^{(1)} \quad \cdots \quad \sum_{p=0}^{d} x_p^{(i)} w_p^{(k)} \right]$

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

$$= \sum_{i=1}^{N} \sum_{j=1}^{k} \left(y_j^{(i)} - \sum_{p=0}^{d} \left(x_p^{(i)} w_p^{(j)} \right) \right)^2$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{N} \left[y_j^{(i)} - \sum_{p=0}^{d} \left(x_p^{(i)} w_p^{(j)} \right) \right]^2$$

Let's write this as $E(W) = \sum_{j=1}^{k} E(\bar{w}^{(j)})$

$$E(W) = \sum_{j=1}^{k} E(\bar{w}^{(j)})$$

where $\bar{w}^{(j)} = \left(w_0^{(j)}, w_1^{(j)} \cdots w_d^{(j)} \right)^T$

i.e. jth ~~Row~~ Column in W.

Minimizing $E(W)$ is minimizing each term in $\sum_{j=1}^{k} E(\bar{w}^{(j)})$ [each of $E(\bar{w}^{(j)})$ is a ~~positive~~ non-negative quantity]

i.e. ~~$\bar{w}^{(j)*} = (X^T X)^T X^T \bar{y}^{(j)}$~~

optimizing,

$$\bar{w}^{(j)*} = (X^T X)^{-1} X^T \bar{y}_j$$

where $\bar{y}_j$ is jth column in Y.

So optimizing weights matrix:

$$W = \left[\left((X^T X)^{-1} X^T \bar{y}_1 \right)^* \; (X^T X)^{-1} X^T \bar{y}_2 \; \cdots \; (X^T X)^{-1} X^T \bar{y}_k \right]$$

$$= (X^T X)^{-1} X^T \left[\bar{y}_1 \; \bar{y}_2 \; \cdots \; \bar{y}_k \right]$$

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

5. Simple linear regression with weighted sum of squared error:

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

$r_i > 0$

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

$$= (\bar{y} - \bar{\hat{y}})^T R (\bar{y} - \bar{\hat{y}}) \qquad R = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & & \\ & & r_3 & \\ 0 & & & r_N \end{bmatrix}$$

$$= (\bar{y} - X\bar{w})^T R (\bar{y} - X\bar{w}) \qquad \text{diagonal matrix}$$

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

minimizing this $\Rightarrow \nabla E(\bar{w}) = 0$

~~$\nabla E(\bar{w}) = 2(RX)^T \bar{y} +$~~ .

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

$$\Rightarrow \bar{w}^* = (X^T R X)^{-1} (RX)^T \bar{y}$$

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

6. Regularization :-

put a constraint on the l2-norm of the weight vector $\bar{w}$.

* Bias is excluded.

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

$$\bar{w} = [w_1, w_2 \dots w_d]^T$$

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

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

$$\Rightarrow \nabla E(\bar{w}) = -2X^T(\bar{y} - X\bar{w}) + 2\lambda I \bar{w} = 0$$

$$\Rightarrow -2X^T\bar{y} + 2X^TX\bar{w} + 2\lambda I \bar{w} = 0$$

$$\Rightarrow \cancel{\bar{w}(X}$$

$$(XX^T + \lambda I)\bar{w} = X^T\bar{y}$$

$$\Rightarrow \cancel{\bar{w}^* = (XX^T + \lambda I)^{-1}\bar{y}}$$

$$\bar{w}^* = (XX^T + \lambda I)^{-1} \cdot X^T\bar{y}$$

Regularization helps in avoiding overfitting.

8. When training data is noisy, the model won't be ~~a~~ able to generalize well to future data. Regularization helps in shrinking these learned estimates to zero.

7. Cost function :

$$E(\bar{w}) = \sum_{i=1}^{N} (y_i - w_1 x_i - w_0 - w_1 \eta_i)^2$$

model:

$$\hat{y} = w_1 x' + w_0$$

$$= w_1(x + \eta) + w_0$$

$$\Rightarrow E(\bar{w}) = \sum_{i=1}^{N} \left\{ (y_i - w_1 x_i - w_0)^2 + (w_1 n_i)^2 - 2(y_i - w_1 x_i - w_0)(w_1 n_i) \right\}$$

~~$= N\mathbb{E}[y_i - w_1$~~

$$= \sum_{i=1}^{N} (y_i - w_1 x_i - w_0)^2 + N\mathbb{E}[(w_1 \bar{n}_i)^2] - 2N\mathbb{E}[(y_i - w_1 x_i - w_0)(w_1 n_i)]$$

$$\Rightarrow E(\bar{w}) = \sum_{i=1}^{N} (y_i - w_1 x_i - w_0)^2 + N w_1^2 \mathbb{E}[(\bar{n}_i)^2] - 2N\mathbb{E}[(y_i - w_1 x_i - w_0)]\,\mathbb{E}[\bar{n}] \cdot w_1$$

$\mathbb{E}[\cdot] \rightarrow$ expectation

$\mathbb{E}[x] = \frac{1}{N}\sum_{i=1}^{N} x_i$

$\bar{n} = \begin{bmatrix} n_1 \\ \vdots \\ n_N \end{bmatrix}$

$$= \underbrace{\sum_{i=1}^{N} (y_i - w_1 x_i - w_0)^2}_{\text{noise-free cost function}} + \underbrace{N^2 w_1^2 \sigma^2}_{\lambda w_1^2}$$

$\mathbb{E}[\bar{n}] = 0$

~~This is~~ cost function

λw_1^2 : l^2 norm regularization term.

8. We have $p(\bar{y} \mid x, \bar{w}, \sigma^2) \sim \mathcal{N}\left(\sum_{j=0}^{d} \bar{x}_j \cdot w_j, \sigma^2 I\right)$

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

~~$p(\bar{w} \mid x, \bar{y}, \alpha, \sigma^2) \sim \mathcal{N}(0, \alpha^2 I)$~~

~~By~~

We need to find $\bar{w}$ that maximizes posterior distribution of $\bar{w}$ given $X, \bar{y}, \alpha, \sigma$

i.e. $p(\bar{w} \mid X, \bar{y}, \alpha, \sigma)$

By Bayes' theorem,

$$p(\bar{w} \mid x, \bar{y}, \alpha, \sigma) = \frac{p(\bar{y} \mid x, \bar{w}, \sigma) \cdot p(\bar{w} \mid \alpha)}{p(x, \bar{w}, \sigma, \alpha)}$$

$$\Rightarrow p(\bar{w} \mid x, \bar{y}, \alpha, \sigma) \propto p(\bar{y} \mid x, \bar{w}, \sigma)\, p(\bar{w} \mid \alpha)$$

$$p(\bar{y} \mid x, \bar{w}, \sigma)\, p(\bar{w} \mid \alpha) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left(-\frac{(\bar{y} - \hat{\bar{y}})^T (\bar{y} - \hat{\bar{y}})}{2\sigma^2}\right).$$

$\hat{y} = \sum_{j=0}^{d} x_j \cdot w_j$

$$\frac{1}{(\sqrt{2\pi\alpha^2})^{d}} \exp\left(\frac{-\bar{w}^T \bar{w}}{2\sigma^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^N} \frac{1}{(\sqrt{2\pi\alpha^2})^{d}} \exp\left[-\left\{\frac{(\bar{y} - \hat{\bar{y}})^T (\bar{y} - \hat{\bar{y}})}{2\sigma^2} + \frac{\bar{w}^T \bar{w}}{2\alpha^2}\right\}\right]$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^N} \frac{1}{(\sqrt{2\pi\alpha^2})^d} \exp\left[-\frac{\left[(\bar{y}-\hat{\bar{y}})^T(\bar{y}-\hat{\bar{y}}) + \frac{\sigma^2}{\alpha^2}\bar{w}^T\bar{w}\right]}{2\sigma^2} \right]$$

Maximizing this probability is minimizing the expression in exponent (without -ve sign)

i.e. minimizing $(\bar{y}-\hat{\bar{y}})^T(\bar{y}-\hat{\bar{y}}) + \frac{\sigma^2}{\alpha^2}\bar{w}^T\bar{w}$

This expression is similar to l2 regularization with ~~$\lambda = \frac{\sigma^2}{\sigma^2}$~~ $\lambda = \frac{\sigma^2}{\alpha^2}$