EE5600: Intro to AI and ML

1. Linear Regression:

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

$N \rightarrow$ number of training examples

$d \rightarrow$ length of input vector.

$$\bar{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}_{N\times 1} \rightarrow \text{outputs (labels)} \qquad \bar{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_d \end{bmatrix}_{(d+1)\times 1} \rightarrow \text{weights}$$

Cost function. (sum of squared error) $E(\bar{\omega}) = \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$

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

$$\hat{\bar{y}} = X.\bar{\omega} \rightarrow (N\times 1)$$

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

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

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

(weights that minimize the sum of squared error)

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

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

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

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

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

To show that this is minimum,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

take $2x = z$, $2\underline{u} = \underline{v}$ $\Rightarrow$ $\mu_j = \frac{k_j}{2}$

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

k_j = mean of v_j

[z replaced by x]

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

$$(w_1, w_2, \ldots w_M) = (2u_1, 2u_2, \ldots 2u_M)$$

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

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

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

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

$$\hat{Y} = X \cdot W \qquad \cancel{\hat{y}^{(i)} = x^{(i)} W = \sum_{p=0}^{d} x_p^{(i)} w}$$

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

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

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

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 $\sum_{j=1}^{k} E(\bar{\omega}^{(j)})$ [each of $E(\bar{\omega}^{(j)})$ is a ~~positive~~ non-negative quantity]

i.e. $\cancel{\bar{W}^{(j)*} = (X^T X)^{-1} X^T \bar{y}^{(j)}}$

optimizing,

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

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

So optimizing weights matrix:

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

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

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

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

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

$r_i > 0$

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

$$= (\bar{y} - \hat{\bar{y}})^T R (\bar{y} - \hat{\bar{y}})$$

$$R = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & & \\ \vdots & & r_3 & \\ 0 & & & \ddots \; r_N \end{bmatrix}$$

diagonal matrix

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

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

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

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

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

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

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

6. Regularization :-

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

* Bias is excluded.

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

$$\bar{w} = [w_1 \; w_2 \; \cdots \; w_d]^T$$

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

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

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

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

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

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

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

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

Regularization helps in avoiding overfitting.

8. When training data is noisy, the model won't be ~~ab~~ 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:

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

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

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

$$\cancel{= N E[\, y_i - w_1}$$

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

$$- 2 N E[(y_i - w_1 x_i - w_0)(w_1 n_i)]$$

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

$$+ N w_1^2 E[(\bar{n}_i)^2]$$

$$- 2 N E[(y_i - w_1 x_i - w_0)]\, E[\bar{n}] \cdot w_1$$

$E[\cdot] \rightarrow$ expectation

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

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

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

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

~~This is~~ cost function

$$\underbrace{\lambda w_1^2}_{l^2 \text{ norm regularization term.}}$$

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

$p(\bar{\omega}/\alpha) \sim N(0, \alpha^2 I)$

$\cancel{\Rightarrow p(\bar{\omega}/x,\bar{y},\alpha,\sigma) \sim N(0,\alpha^2 I)}$

~~By~~

~~R~~

We need to find $\bar{\omega}$ that maximizes

posterior distribution of $\bar{\omega}$ given $X, \bar{y}, \alpha, \sigma$

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

By Bayes' theorem,

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

$$\Rightarrow p(\bar{\omega}/x,\bar{y},\alpha,\sigma) \propto p(\bar{y}/x,\bar{\omega},\sigma)\, p(\bar{\omega}/\alpha)$$

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

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

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

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

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

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

~~as~~ with $\cancel{\lambda = \frac{\sigma^2}{\alpha^2}}$ $\boxed{\lambda = \frac{\sigma^2}{\alpha^2}}$