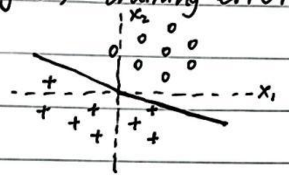
Homework 3

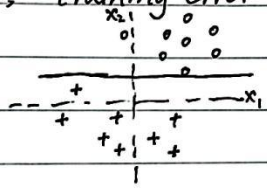
2. (1) $j=0$, training error $=0$

result:



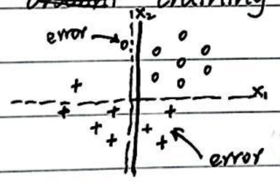
(2) $j=1$, training error $=0$

result:



(3) $j=2$, ~~traint~~ training error >0

result:



(2) a. A. Training error can be 0 when $w_1=0$.

b. $P(y=1|x,w) = P(y=0|x,w) = \frac{1}{2}$

$\Rightarrow -w_0 - w_1x_1 - w_2x_2 = 0$

So w_0 must be 0.

c. $P(y=1|x,w) > P(y=0|x,w)$

$\Rightarrow \exp(-w_0 - w_1x_1 - w_2x_2) < 1$

$\Rightarrow w_0 + w_1x_1 + w_2x_2 > 0$

So $w_0 > 0$.

3. (1) We want to minimize $\|y - Xw\|_2^2$.

So $\frac{\partial \|y-Xw\|_2^2}{\partial w} = 0$, that is, $-2X^T(y-Xw)=0$

So $\hat{w} = (X^TX)^{-1}X^Ty$

$\hat{y} = X\hat{w} = X(X^TX)^{-1}X^Ty$

No.
Date

(2) $P^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = P$

$\lambda x = Px = P^2x = \lambda^2 x$

$\Rightarrow \lambda = \lambda^2 \Rightarrow \lambda = 0 \text{ or } \lambda = 1$

(3) $E(\hat{w}) = E((X^TX)^{-1}X^Ty)$

$= E((X^TX)^{-1}X^T(Xw+\varepsilon))$

$= E((X^TX)^{-1}X^TXw)$

$= E(w)$

$= w$

$Var(\hat{w}) = E[(\hat{w}-E(\hat{w}))(\hat{w}-E(\hat{w}))^T]$

$= E[(\hat{w}-w)(\hat{w}-w)^T]$

$= E[((X^TX)^{-1}X^T\varepsilon)((X^TX)^{-1}X^T\varepsilon)^T]$

$= E[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}]$

$= E[(X^TX)^{-1}X^TX(X^TX)^{-1}]\cdot E(\varepsilon\varepsilon^T)$

$= E[(X^TX)^{-1}]\sigma^2$

$= (X^TX)^{-1}\sigma^2$

(4) $SS_{tot} = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$

$= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y}) + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$

$= SS_{res} + SS_{reg} + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$

$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = (y-\hat{y})^T(X\hat{w}-\bar{y}\mathbf{1})$

$= (y-\hat{y})^TX\hat{w} - \bar{y}(y-\hat{y})^T\mathbf{1}$

From (1), $\hat{y} = X(X^TX)^{-1}X^Ty$,

$X^T(y-\hat{y}) = X^Ty - X^TX(X^TX)^{-1}X^Ty = X^Ty - X^Ty = 0$

First column of X is $\mathbf{1}$, so $\mathbf{1}^T(y-\hat{y}) = 0$

So $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = 0$

$\Rightarrow SS_{tot} = SS_{res} + SS_{reg}$

No.

Date . .

4. (1) $\sum_{i=1, i\neq k}^{n} (y_i - x_i^T w)^2 + \lambda \|w\|_2^2 = \|y - Xw\|_2^2 - (y_k - x_k^T w)^2 + \lambda \|w\|_2^2$

$$\frac{\partial \left(\|y - Xw\|_2^2 - (y_k - x_k^T w)^2 + \lambda \|w\|_2^2 \right)}{\partial w} = -2X^T(y - Xw) + 2x_k(y_k - x_k^T w) + 2\lambda w = 0$$

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

$$\Rightarrow \hat{w}^{[k]} = (X^T X + \lambda I - x_k x_k^T)^T (X^T y - x_k y_k)$$

(2) Using Sherman-Morrison, we get

$$(X^T X + \lambda I - x_k x_k^T)^{-1} = (X^T X + \lambda I)^{-1} + \frac{(X^T X + \lambda I)^{-1} x_k x_k^T (X^T X + \lambda I)^{-1}}{1 - x_k^T (X^T X + \lambda I)^{-1} x_k}$$

And we have $x_k^T (X^T X + \lambda I)^{-1} x_k = p_{kk}$, $\hat{y}_k = x_k^T (X^T X + \lambda I)^{-1} X^T y$

Then $x_k^T \hat{w}^{[k]} - y_k = x_k^T \left[(X^T X + \lambda I)^{-1} + \frac{(X^T X + \lambda I)^{-1} x_k x_k^T (X^T X + \lambda I)^{-1}}{1 - x_k^T (X^T X + \lambda I)^{-1} x_k} \right] (X^T y - x_k y_k) - y_k$

$$= \hat{y}_k - p_{kk} y_k + \frac{p_{kk} \hat{y}_k}{1 - p_{kk}} - \frac{p_{kk}^2 y_k}{1 - p_{kk}} - y_k$$

$$= \frac{\hat{y}_k - y_k}{1 - p_{kk}}$$

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^{n} (x_k^T \hat{w}^{[k]} - y_k)^2 = \frac{1}{n} \sum_{n=1}^{k} \left(\frac{\hat{y}_k - y_k}{1 - p_{kk}} \right)^2$$

(3) $V(\lambda) = \frac{1}{n} \sum_{k=1}^{n} \left(\frac{1 - p_{kk}}{\frac{1}{n} tr(I - P)} \right)^2 \left(\frac{\hat{y}_k - y_k}{1 - p_{kk}} \right)^2$

$$= \frac{1}{n} \sum_{k=1}^{n} \left(\frac{\hat{y}_k - y_k}{\frac{1}{n} tr(I - P)} \right)^2$$

$$= \frac{1}{n} \left(\frac{1}{\frac{1}{n} tr(I - P)} \right)^2 \sum_{k=1}^{n} (\hat{y}_k - y_k)^2$$

$$= \frac{n}{(tr(I) - tr(P))^2} \|\hat{y}_0 - y\|^2$$

$$= \frac{n}{(n - tr(P))^2} \|Py - y\|^2$$

$$= \frac{\frac{1}{n} \|(I - P) y\|^2}{(1 - tr(P)/n)^2}$$

5. LASSO: $\min_{w,z} \left[\|y - Xw\|_2^2 + \lambda \|z\|_1 \right]$, subject to $w - z = 0$

$$L(w, z, u) = \|y - Xw\|_2^2 + \lambda \|z\|_1 + u^T (w - z) + \frac{1}{2} \|w - z\|_2^2$$

Step 1: $w^{(k+1)} = \arg\min_w L(w, z^{(k)}, u^{(k)})$

$$\frac{\partial L}{\partial w} = -2X^T (y - Xw) + u^{(k)} + (w - z^{(k)}) = 0$$

$$\Rightarrow w^{(k+1)} = (I + 2X^T X)^{-1} (2X^T y - u^{(k)} + z^{(k)})$$

↓ Next Page

No.

Date

Step 2: $z^{(k+1)} = \arg\min_{z} L(w^{(k+1)}, z, u^{(k)})$

$$\frac{\partial L}{\partial z} = \lambda\, \partial \|z\|_1 - u^{(k)} - (w^{(k+1)} - z)$$

$$\partial_{z_j} \|z\|_1 = \begin{cases} 1, & z_j > 0 \\ [-1, 1], & z_j = 0 \\ -1, & z_j < 0 \end{cases}$$

$$\frac{\partial L}{\partial z_j} = 0 \Rightarrow z_j^{(k+1)} = \begin{cases} u_j^{(k)} + w_j^{(k+1)} - \lambda, & u_j^{(k)} + w_j^{(k+1)} > \lambda, \\ 0, & u_j^{(k)} + w_j^{(k+1)} \in [-\lambda, \lambda], \\ u_j^{(k)} + w_j^{(k+1)} + \lambda, & u_j^{(k)} + w_j^{(k+1)} < \lambda \end{cases}$$

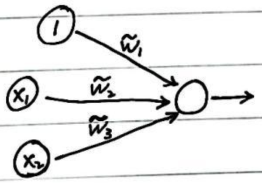
Step 3: Iterate $u^{(k+1)} = u^{(k)} + w^{(k+1)} - z^{(k+1)}$

6. (1) $z_1 = w_1 + w_3 x_1 + w_5 x_2$, $z_2 = w_2 + w_4 x_1 + w_6 x_2$

$$z' = w_7 + w_8 h_1 + w_9 h_2 = w_7 + w_8 c z_1 + w_9 c z_2$$

$$P(y=1|x,w) = g(z') = \frac{1}{1 + e^{-(w_7 + w_8 c(w_1 + w_3 x_1 + w_5 x_2) + w_9 c(w_2 + w_4 x_1 + w_6 x_2))}}$$

(2)



$$w_7 + w_8 c(w_1 + w_3 x_1 + w_5 x_2) + w_9 c(w_2 + w_4 x_1 + w_6 x_2)$$

$$= (w_7 + w_8 c w_1 + w_9 c w_2) + (w_8 c w_3 + w_9 c w_4) x_1 + (w_8 c w_5 + w_9 c w_6) x_2$$

$$\Rightarrow \begin{cases} \tilde{w}_1 = w_7 + w_1 w_8 c + w_2 w_9 c \\ \tilde{w}_2 = w_3 w_8 c + w_4 w_9 c \\ \tilde{w}_3 = w_5 w_8 c + w_6 w_9 c \end{cases}$$

(3) No. Not all activation function is linear. But if all activation functions are linear, it can be without hidden layer.