Problem 1:
Given the following training data X and Y, derive the vector ***w*** using linear regression (to fit the given data). The number of features in each data point is 4 (d=4). X_0 is not shown (always 1).

X_1=[3, 4.5, 6, 3.4] y_1= 60
X_2=[5, 7, 9, 5.2] y_2= 84
X_3=[8, 10, 12, 7] y_3=120
X_4=[1, 3, -4, 2] y_4=16
X_5=[0, 4, 10, 9] y_5=125
X_6=[2, -2, 3, 1] y_6=34

Show your ***w*** and also each step in the derivation.

$$X = \begin{bmatrix} 1 & 3 & 4.5 & 6 & 3.4 \\ 1 & 5 & 7 & 9 & 5.2 \\ 1 & 8 & 10 & 12 & 7 \\ 1 & 1 & 3 & -4 & 2 \\ 1 & 0 & 4 & 10 & 9 \\ 1 & 2 & -2 & 3 & 1 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 5 & 8 & 1 & 0 & 2 \\ 4.5 & 7 & 10 & 3 & 4 & -2 \\ 6 & 9 & 12 & -4 & 10 & 3 \\ 3.4 & 5.2 & 7 & 2 & 9 & 1 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 5 & 8 & 1 & 0 & 2 \\ 4.5 & 7 & 10 & 3 & 4 & -2 \\ 6 & 9 & 12 & -4 & 10 & 3 \\ 3.4 & 5.2 & 7 & 2 & 9 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 4.5 & 6 & 3.4 \\ 1 & 5 & 7 & 9 & 5.2 \\ 1 & 8 & 10 & 12 & 7 \\ 1 & 1 & 3 & -4 & 2 \\ 1 & 0 & 4 & 10 & 9 \\ 1 & 2 & -2 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 19 & 26.5 & 36 & 27.6 \\ 19 & 103 & 127.5 & 161 & 96.2 \\ 26.5 & 127.5 & 198.25 & 232 & 161.7 \\ 36 & 161 & 232 & 386 & 236.2 \\ 27.6 & 96.2 & 161.7 & 236 & 173.6 \end{bmatrix}$$

$$X^{\dagger} = (X^T X)^{-1} X^T = \begin{bmatrix} 1.1198 & 0.3936 & -0.7350 & 0.1640 & -0.1368 & 0.1944 \\ -0.3464 & -0.1183 & 0.2768 & 0.0573 & -0.0488 & 0.1793 \\ 0.2435 & 0.1144 & -0.1130 & 0.0078 & -0.0494 & -0.2034 \\ 0.1643 & 0.0731 & -0.0874 & -0.1282 & -0.0061 & -0.0157 \\ -0.4170 & -0.1731 & 0.2279 & 0.1208 & 0.1550 & 0.0864 \end{bmatrix}$$

$$W = X^{\dagger} \cdot Y = \begin{bmatrix} 1.1198 & 0.3936 & -0.7350 & 0.1640 & -0.1368 & 0.1944 \\ -0.3464 & -0.1183 & 0.2768 & 0.0573 & -0.0488 & 0.1793 \\ 0.2435 & 0.1144 & -0.1130 & 0.0078 & -0.0494 & -0.2034 \\ 0.1643 & 0.0731 & -0.0874 & -0.1282 & -0.0061 & -0.0157 \\ -0.4170 & -0.1731 & 0.2279 & 0.1208 & 0.1550 & 0.0864 \end{bmatrix} \begin{bmatrix} 60 \\ 84 \\ 120 \\ 16 \\ 125 \\ 34 \end{bmatrix}$$

$$= \begin{bmatrix} 4.1804 \\ 3.4170 \\ -2.2967 \\ 2.1647 \\ 12.0265 \end{bmatrix}$$

Problem 2:

$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2$. Show that $E_{in}(w) = \frac{1}{N}(w^T X^T X w - 2w^T X^T Y + Y^T Y)$ step by step. For each step, if you need to use any basic matrix operation, please specify it clearly.

$$\frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2 = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n x_n w + 2w^T x_n y_n + y_n y_n)$$
$$= \frac{1}{N}\left(\sum_{n=1}^{N}(w^T x_n x_n w) + \sum_{n=1}^{N}(2w^T x_n y_n) + \sum_{n=1}^{N}(y_n y_n)\right)$$

Since

$$\sum_{n=1}^{N}(w^T x_n x_n w) = w^T x_1 x_1 w + w^T x_2 x_2 w + \cdots + w^T x_n x_n w$$
$$= w^T [x_1 \quad \dots \quad x_n] \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} w$$
$$= w^T X^T X w$$

$$\sum_{n=1}^{N}(2w^T x_n y_n) = 2w^T x_1 y_1 + 2w^T x_2 y_2 + \cdots + 2w^T x_n y_n$$
$$= 2w^T [x_1 \quad \dots \quad x_n] \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$$
$$= 2w^T X^T Y$$

$$\sum_{n=1}^{N}(y_n y_n) = y_1 y_1 + y_2 y_2 + \cdots + y_n y_n$$
$$= [y_1 \quad \dots \quad y_n] \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$$
$$= Y^T Y$$

Thus

$$E_{in}(w) = \frac{1}{N}(w^T X^T X w - 2w^T X^T Y + Y^T Y)$$

Problem 3:

Given our objective function $\mathrm{E}_{in}(w) = \frac{1}{N}(w^T X^T X w - 2w^T X^T Y + Y^T Y)$, we obtained the gradient of E_{in} as $\nabla\mathrm{E}_{in}(\mathrm{w}) = \frac{2}{N}(X^T X w - X^T Y)$. To help you understand how we compute the gradient of E_{in} in linear regression, prove that:

$$\nabla_w(w^T A w) = (A + A^T)w$$

In this equation, A is a matrix of size (d+1) by (d+1). w is the weight vector of size (d+1) by 1. This will help you see how we obtain the first item in the gradient.

Solution:

$$w^T A w = w^T \begin{bmatrix} \sum_{j=0}^{d} a_{0j} w_j \\ \sum_{j=0}^{d} a_{1j} w_j \\ \cdots \\ \sum_{j=0}^{d} a_{dj} w_j \end{bmatrix} = \sum_{i=0}^{d} \sum_{j=0}^{d} w_i a_{ij} w_j$$

Since

$$\frac{\partial w^T A w}{w_k} = \sum_{j=0}^{d} a_{kj} w_k + \sum_{i=0}^{d} a_{ik} w_k$$
$$= A_k w + A_k^T w$$

Thus

$$\nabla_w(w^T A w) = \begin{bmatrix} \frac{\partial w^T A w}{w_0} \\ \frac{\partial w^T A w}{w_1} \\ \cdots \\ \frac{\partial w^T A w}{w_d} \end{bmatrix} = \begin{bmatrix} A_0 \\ A_1 \\ \cdots \\ A_d \end{bmatrix} w + \begin{bmatrix} A_0^T \\ A_1^T \\ \cdots \\ A_d^T \end{bmatrix} w = (A + A^T)w$$

Problem 4:

This is a problem about non-linear transformation function Φ (x)=(1,x_1^2, x_2^2). What kind of boundary in χ (original input space) does a hyperplane $\tilde{w}$ in Z correspond to in the following cases? Draw a picture that illustrates an example of each case.

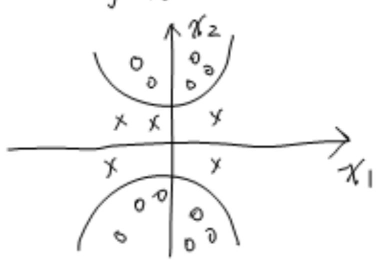
(a) $\tilde{w}_1 > 0$, $\tilde{w}_2 < 0$
(b) $\tilde{w}_1 > 0$, $\tilde{w}_2 = 0$
(c) $\tilde{w}_1 > 0$, $\tilde{w}_2 > 0$, $\tilde{w}_0 < 0$

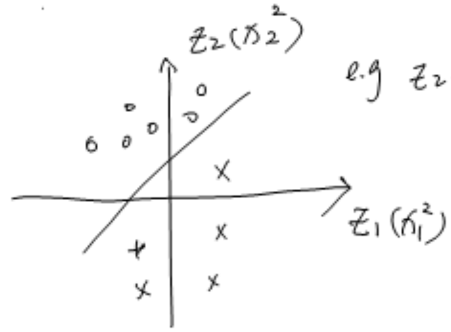
Suppose the hyperplane is $\tilde{w}_0 + \tilde{w}_1 \cdot z_1 + \tilde{w}_2 \cdot z_2 = 0$

$\Rightarrow \tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 \cdot x_2^2 = 0$ $\quad z_0 = 1$

(a). $\tilde{w}_1 > 0$, $\tilde{w}_2 < 0$. $\tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 \cdot x_2^2 = 0$ is a hyperbola in $\mathcal{X}$

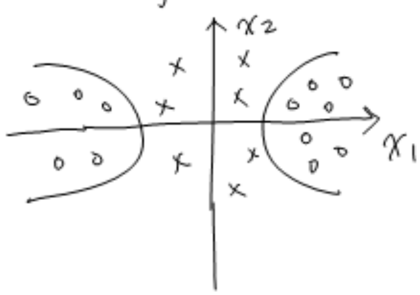
if $w_0 > 0$

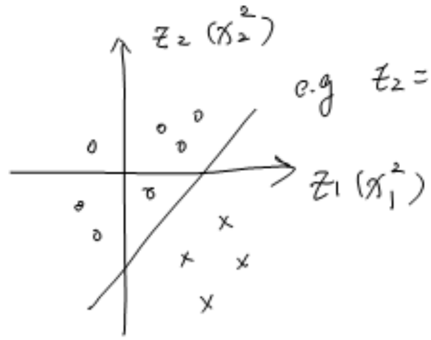




e.g. $z_2 = z_1 + 5$

if $w_0 < 0$

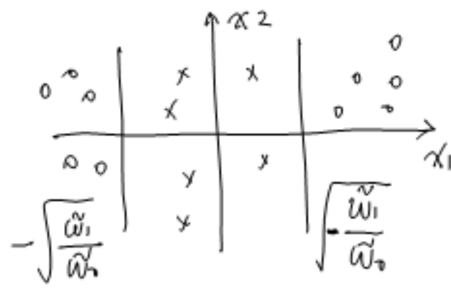


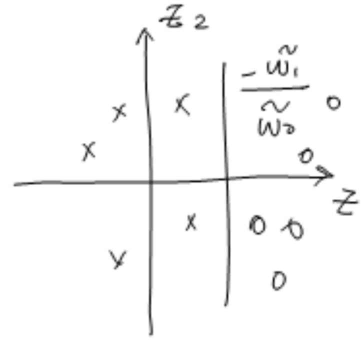


e.g. $z_2 = z_1 - 5$

(b) $\tilde{w}_1 > 0$. $\tilde{w}_2 = 0$ $\quad\quad$ $\tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 = 0 \Rightarrow x_1^2 = -\frac{\tilde{w}_1}{\tilde{w}_0}$

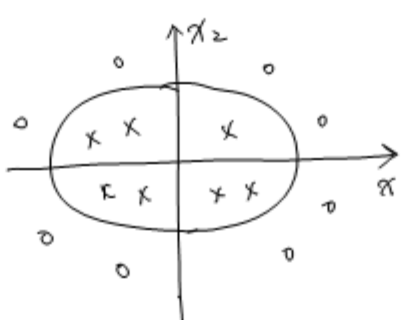
if $\tilde{w}_0 < 0$

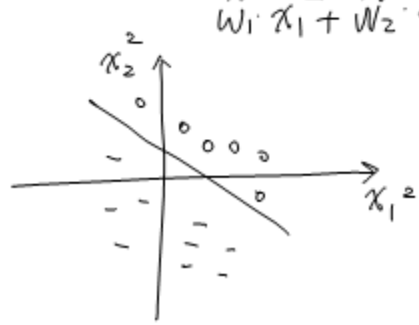




(c) $\tilde{w}_1 > 0$, $\tilde{w}_2 > 0$, $\tilde{w}_0 < 0$ $\quad\quad$ $\tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 \cdot x_2^2 = 0$

$\tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 \cdot x_2^2 = -\tilde{w}_0 > 0$





Problem 5:

Consider a new error measure: $e_n(\boldsymbol{w}) = \max(0, 1-y_n\boldsymbol{w}^T x_n)$. Show that : (1) $e_n(\boldsymbol{w})$ is an upper bound for $[\![\text{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!]$. (2) $\frac{1}{N}\sum_{n=1}^{N} e_n(\boldsymbol{w})$ is an upper bound for the in-sample classification error $E_{in}(\boldsymbol{w})$.

Solution (1):

$$[\![\text{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] = \begin{cases} 1, if\ y_n \neq sign(w^T x_n) \\ 0, if\ y_n = sign(w^T x_n) \end{cases}$$

$$\mathrm{e}_n(\boldsymbol{w}) = \max(0, 1 - y_n \boldsymbol{w}^T x_n) = \begin{cases} 1 + |y_n \boldsymbol{w}^T x_n|, & if\ y_n \boldsymbol{w}^T x_n < 0 \\ 1 - y_n \boldsymbol{w}^T x_n, & if\ 0 \le y_n \boldsymbol{w}^T x_n \le 1 \\ 0, & if\ 1 < y_n \boldsymbol{w}^T x_n \end{cases}$$

1. If $y_n \neq sign(w^T x_n)$, $y_n \boldsymbol{w}^T x_n < 0$. In this case, $[\![\text{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] = 1$ and $\mathrm{e}_n(\boldsymbol{w}) = 1 + |y_n \boldsymbol{w}^T x_n|$. Thus $[\![\text{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] < \mathrm{e}_n(\boldsymbol{w})$

2. If $y_n = sign(w^T x_n)$, $y_n \boldsymbol{w^T} x_n > 0$. In this case, $[\![\mathrm{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] = 0$ and $\mathrm{e}_n(\boldsymbol{w}) \geq 0$. Thus $[\![\mathrm{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] \leq \mathrm{e}_n(\boldsymbol{w})$

Thus, $\mathrm{e_n}(\boldsymbol{w})$ is an upper bound for $[\![\mathrm{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!]$.

Solution (2):

According to (1), we know that $[\![\mathrm{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] \leq \mathrm{e}_n(\boldsymbol{w})$ for each n. Thus, $\sum_{n=1}^{N} [\![\mathrm{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] \leq \sum_{n=1}^{N} \mathrm{e}_n(\boldsymbol{w})$. Since the in-sample classification error $\mathrm{E_{in}}(\boldsymbol{w})$ = $\frac{1}{N} \sum_{n=1}^{N} [\![\mathrm{sign}(\boldsymbol{w}^T x_n) \neq y_n]\!] \leq \frac{1}{N} \sum_{n=1}^{N} \mathrm{e}_n(\boldsymbol{w})$. Thus, $\frac{1}{N} \sum_{n=1}^{N} \mathrm{e}_n(\boldsymbol{w})$ is the upper bound for the in-sample classification error.