SUM OF SQUARED ERROR-BASED LINEAR CLASSIFIER

Problem: Determine the linear classifier that minimizes the sum of squared error for the binary classification problem between classes C_1 and C_2 where $\underline{x}_a \in C_1$ and $\underline{x}_B \in C_2$ such that $\underline{x}_a = [A, B]^T$ and $\underline{x}_B = [-B, -A]^T$ with $0 < A < B$.

Solution: Let $\mathcal{X} = \{\underline{x}_a, \underline{x}_B\}$ be the given dataset where the corresponding class labels are given within the set $Y = \{Y_a, Y_B\}$ where $Y_a = +1$ and $Y_B = -1$.

- Suppose that the exact functional form of the linear classifier is given by: $\boxed{f(\underline{x}) = \underline{w}^T \underline{x} + b \quad (1).}$

Function $f(\underline{x})$ provides the estimated class label $\hat{y}$ which in turn induces a per-pattern error of the following form:

$$\boxed{e(\underline{x}) = y - \hat{y} = y - f(\underline{x}) = y - \underline{w}^T \underline{x} - b \quad (2)}$$

- However, we need to define an overall cost functional taking into consideration the total misclassification cost such that:

$$\boxed{J(\underline{w}, b) = \sum_{\underline{x} \in \mathcal{X}} e^2(\underline{x}) = e^2(\underline{x}_a) + e^2(\underline{x}_B) \quad (3)}$$

- Imposing F.O.Cs in order to determine the optimal parameters of the decision hyperplane, yields:

$$\boxed{\begin{cases} \dfrac{\partial J}{\partial \underline{w}} = \underline{0} \\[2ex] \dfrac{\partial J}{\partial b} = 0 \end{cases} \quad (4)}$$

- Thus, we need to compute the following quantities:

$$\frac{\partial J}{\partial \underline{w}} = \sum_{\underline{x} \in \mathcal{X}} \frac{\partial}{\partial \underline{w}} \left\{ e^2(\underline{x}) \right\} \quad (5) \quad \text{and}$$

$$\frac{\partial J}{\partial b} = \sum_{\underline{x} \in \mathcal{X}} \frac{\partial}{\partial b} \left\{ e^2(\underline{x}) \right\} \quad (6)$$

- Eqs. (5) and (6), may be further expanded as:

$$\frac{\partial J}{\partial \underline{w}} = \sum_{\underline{x} \in \mathcal{X}} 2\, e(\underline{x}) \frac{\partial e(\underline{x})}{\partial \underline{w}} \quad (7) \quad \text{and}$$

$$\frac{\partial J}{\partial b} = \sum_{\underline{x} \in \mathcal{X}} 2\, e(\underline{x}) \frac{\partial e(\underline{x})}{\partial b} \quad (8)$$

- It also holds that:

$$\frac{\partial e(x)}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}} \left\{ y - \underline{w}^T \underline{x} - b \right\} = \frac{\partial}{\partial \underline{w}} \left\{ - \underline{w}^T \underline{x} \right\} = -\underline{x} \quad (9)$$

$$\frac{\partial e(x)}{\partial b} = \frac{\partial}{\partial b} \left\{ y - \underline{w}^T \underline{x} - b \right\} = \frac{\partial}{\partial b} \left\{ - b \right\} = -1 \quad (10)$$

#3

▷ Eqs. (7) and (8), may be written according to Eqs. (9) and (10) as:

$$\boxed{\begin{aligned} \frac{\partial J}{\partial \underline{w}} &= \sum_{\underline{x} \in \mathcal{X}} -2e(\underline{x})\,\underline{x} = \underline{0} \quad (11) \quad \text{and} \\ \frac{\partial J}{\partial b} &= \sum_{\underline{x} \in \mathcal{X}} -2e(\underline{x}) = 0 \quad (12) \end{aligned}}$$

▷ Eqs. (11) and (12) give:

$$\boxed{\left\{ \begin{aligned} &\sum_{\underline{x} \in \mathcal{X}} e(\underline{x})\,\underline{x} = \underline{0} \quad (13) \\ &\sum_{\underline{x} \in \mathcal{X}} e(\underline{x}) = 0 \quad (14) \end{aligned} \right.}$$

▷ Thus, we may write for Eq. (13) and (14):

$$\boxed{\begin{aligned} &e(\underline{x}_\alpha)\,\underline{x}_\alpha + e(\underline{x}_\beta)\,\underline{x}_\beta = \underline{0} \quad (15) \\ &e(\underline{x}_\alpha) + e(\underline{x}_\beta) = 0 \quad (16) \end{aligned}}$$

▷ Eq. (16) yields: $e(\underline{x}_\beta) = -e(\underline{x}_\alpha)$, which is plugged into Eq. (15) to give:

$$e(\underline{x}_\alpha)\,\underline{x}_\alpha - e(\underline{x}_\alpha)\,\underline{x}_\beta = \underline{0} \Rightarrow e(\underline{x}_\alpha)\underbrace{(\underline{x}_\alpha - \underline{x}_\beta)}_{\neq \underline{0}} = \underline{0} \Rightarrow \boxed{e(\underline{x}_\alpha) = 0 \quad (17)}$$

▷ Thus, according to Eq. (16), we get that: $\boxed{e(\underline{x}_\beta) = 0 \quad (18)}$

- The system of linear equations (17) and (18) define an under-determined linear system of 2 Equations with 3 unknown variables:

$$\begin{cases} e(\underline{x_a}) = 0 & (19) \\ e(\underline{x_B}) = 0 & (20) \end{cases} \Rightarrow \begin{cases} \underline{w}^T \underline{x_a} + b = x_a & (21) \\ \underline{w}^T \underline{x_B} + b = Y_B & (22) \end{cases} \rightarrow \begin{cases} \underline{w}^T \underline{x_a} + b = +1 & (23) \\ \underline{w}^T \underline{x_B} + b = -1 & (24) \end{cases}$$

- Taking into consideration the fact that $\underline{w} = [w_1 \; w_2]^T$ and $\underline{x_a} = [A \;\; B]^T$ with $\underline{x_B} = [-B \; -A]^T$, Eqs (23) and (24) yield:

$$\begin{cases} [w_1 \; w_2]\begin{bmatrix} A \\ B \end{bmatrix} + b = +1 & (25) \\ [w_1 \; w_2]\begin{bmatrix} -B \\ -A \end{bmatrix} + b = -1 & (26) \end{cases} \rightarrow \begin{cases} Aw_1 + Bw_2 + b = +1 & (27) \\ -Bw_1 - Aw_2 + b = -1 & (28) \end{cases} \rightarrow$$

$$\begin{cases} (A+B)w_1 + (A+B)w_2 = 2 \quad (29) & \text{(Pairwise Subtraction)} \\ (A-B)w_1 + (B-A)w_2 + 2b = 0 \quad (30) & \text{(Pairwise Addition)} \end{cases}$$

- Eq.(29) yields that: $(A+B)w_2 = 2 - (A+B)w_1 \rightarrow \boxed{w_2 = \frac{2}{A+B} - w_1 \quad (31)}$

- Eq.(30) yields that: $2b = (B-A)w_1 + (A-B)w_2 \Rightarrow$

$$b = \frac{1}{2}(B-A)w_1 + \frac{1}{2}(A-B)w_2 \xrightarrow{\text{Eq.}(31)}$$

$$b = \frac{1}{2}(B-A)w_1 + \frac{1}{2}(A-B)\left[\frac{2}{A+B} - w_1\right] \Rightarrow$$

$$b = \frac{1}{2}(B-A)w_1 + \frac{A-B}{A+B} - \frac{1}{2}(A-B)w_1 \rightarrow$$

$$b = \frac{1}{2}(B-A)w_1 + \frac{1}{2}(B-A)w_1 + \frac{A-B}{A+B} \rightarrow$$

$$\boxed{b = (B-A)w_1 + \frac{A-B}{A+B} \quad (32)}$$

- Eqs. (31) and (32) provide the optimal parameters $\underline{w}^*$ and b^* for the decision hyperplane according to the Minimum Squared Error Criterion as:

$$\boxed{\begin{aligned} \underline{W}^*_{MSE} &= \left[w_1 \quad \frac{2}{A+B} - w_1 \right]^T \quad \text{where } w_1 \in \mathbb{R} \quad (33) \\ b^*_{MSE} &= (B-A)\,w_1 + \frac{A-B}{A+B} \quad \text{where } w_1 \in \mathbb{R} \quad (34) \end{aligned}}$$

- For this particular problem, the Minimum Squared Error Optimized Linear Classifier provides an infinite set of possible solutions since $w_1 \in \mathbb{R}$ is a free parameter.

- We have already established that the Hard-Margin Optimizing Linear SVM classifier will essentially acquire the following set of optimal parameters:

$$\begin{cases} \underline{W}^*_{SVM} = \left[\frac{1}{A+B} \quad \frac{1}{A+B} \right] & (35) \\ b^*_{SVM} = \emptyset & (36) \end{cases}$$

- So, the next question is whether the MSE-based Linear Classifier can actually acquire the SVM-based weights?

- The answer is positive, since by setting $w_1 = \frac{1}{A+B}$, we get:

$$w_2 = \frac{2}{A+B} - \frac{1}{A+B} = \frac{1}{A+B} \quad \text{and} \quad b = \frac{(B-A)}{A+B} + \frac{A-B}{A+B} = \emptyset.$$