4. Clasificación de toxicidad y pérdida máxima de grupo.

- Clasificador D: w = [−0.1, 1, 0]
- Clasificador T : w = [−0.1, 0, 1]

Comentario (x), Toxicidad (y), Presencia de menciones demográficas (d), Presencia de palabras tóxicas (t)

$$\phi(x) = [1, d, t]$$

Entonces tenemos cuatro grupos: (y = 1,d = 1), (y = 1,d = 0), (y = −1,d = 1), y (y = −1, d = 0)

a. En palabras, describe el comportamiento del Claisificador D y el clasificador T.

Clasificador D.

$$f_w = sign(w \cdot \phi(x)) = sign([-0.1{,}1{,}0] \cdot [1, d, t]) = sign([-0.1, d, 0]) \begin{cases} +1 \; si \; d = 1 \\ -1 \; si \; d = 0 \end{cases}$$

El clasificador D asignará una salida de $+1$ a un comentario con $y = 1$ si y solo si el comentario menciona identidades demográficas independientemente si hay palabras toxicas o no. Y si no hay menciones demográficas lo clasifica como tóxico.

$$Loss_{0-1}(x, y, w) = 1 \leftrightarrow bool[f_w(x) \neq 1]$$

Clasificador T.

$$f_w = sign(w \cdot \phi(x)) = sign([-0.1{,}0{,}1] \cdot [1, d, t]) = sign([-0.1, 0, t]) \begin{cases} +1 \; si \; t = 0 \\ -1 \; si \; t = 1 \end{cases}$$

$$Loss_{0-1}(x, y, w) = 1 \leftrightarrow bool[f_w(x) \neq 1]$$

El clasificador T produce una pérdida si no hay palabras tóxicas $(t = 0)$. El clasificador T asignará una salida de $+1$ a un comentario con $y = 1$ si y solo si el comentario contiene palabras tóxicas.

b. Calcula las siguientes tres cantidades sobre el Clasificador D usando el conjunto de datos de arriba:

$$f_w = sign([-0.1, 1, 0] \cdot [1, 0, 0]) = -1 \rightarrow 1[-1 \neq -1] = 0$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 0, 1]) = -1 \rightarrow 1[-1 \neq -1] = 0$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 1, 0]) = +1 \rightarrow 1[1 \neq -1] = 1$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 1, 1]) = +1 \rightarrow 1[1 \neq -1] = 1$$

$$TrainLoss_{-1}(w) = \frac{1}{4}(0 + 0 + 1 + 1) = 0.5$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 0, 0]) = -1 \rightarrow 1[-1 \neq 1] = 1$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 0, 1]) = -1 \rightarrow 1[-1 \neq 1] = 1$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 1, 0]) = +1 \rightarrow 1[1 \neq 1] = 0$$

$$f_w = sign([-0.1, 1, 0] \cdot [1, 1, 1]) = +1 \rightarrow 1[1 \neq 1] = 0$$

$$TrainLoss_{+1}(w) = \frac{1}{4}(1 + 1 + 0 + 0) = 0.5$$

c. Ahora calcula las siguientes cantidades sobre el Clasificador T usando el mismo con- junto de datos:

$$f_w = sign([-0.1, 0, 1] \cdot [1, 0, 0]) = -1 \rightarrow 1[-1 \neq -1] = 0$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 0, 1]) = +1 \rightarrow 1[1 \neq -1] = 1$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 1, 0]) = -1 \rightarrow 1[-1 \neq -1] = 0$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 1, 1]) = +1 \rightarrow 1[1 \neq -1] = 1$$

$$TrainLoss_{-1}(w) = \frac{1}{4}(0 + 1 + 0 + 1) = 0.5$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 0, 0]) = -1 \rightarrow 1[-1 \neq 1] = 1$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 0, 1]) = +1 \rightarrow 1[1 \neq 1] = 0$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 1, 0]) = -1 \rightarrow 1[-1 \neq 1] = 1$$

$$f_w = sign([-0.1, 0, 1] \cdot [1, 1, 1]) = +1 \rightarrow 1[1 \neq 1] = 0$$

$$TrainLoss_{+1}(w) = \frac{1}{4}(1 + 0 + 1 + 0) = 0.5$$