

## Ejercicio 1.1

Consideremos el perceptrón en 2 dimensiones con función de activación dada por  $h(x) = \text{sign}(w^T x)$ , donde  $x = [1, x_1, x_2]^T$  es el vector de entrada y  $w = [w_0, w_1, w_2]^T$  es el vector de pesos.

Supongamos que la regeón en el plano donde  $h(x) = 1$  está separada de la regeón donde  $h(x) = -1$  por una linea recta. Queremos demostrar que esta linea recta es perpendicular a  $w$  y pasa por el origen.

Sea la linea recta que separa las 2 regeones dadas por la ecuación  $x_2 = ax_1 + b$  donde  $a$  es la pendiente y  $b$  es la intersección en el eje y.

Consideremos 2 puntos  $x_1$  y  $x_2$  en esta linea recta

Entonces

$$x_2 = ax_1 + b$$

$$w^T x_1 = w_0 + w_1 x_1 + w_2 x_2 > 0 \text{ porque } x_1 \text{ está en la regeón donde } h(x) = +1$$

$$w^T x_2 = w_0 + w_1 x_1 + w_2 x_2 < 0 \text{ porque } x_2 \text{ está en la regeón donde } h(x) = -1$$

Restando estos desigualdades

$$w^T(x_1 - x_2) > 0$$

Como linea recta  $x_2 = ax_1 + b$  es perpendicular al vector  $[1, -a, 1]^T$  podemos escribir  $x_1 - x_2$  como un múltiplo escalar de este vector

$$x_1 - x_2 = k[1, -a, 1]^T$$

dónde  $k$  es una constante escalar. por lo tanto

$$w^T(x_1 - x_2) = \bar{w}^T k [1, -a, 1]^T = k(w_0 - w_1 a + w_2) > 0$$

ya que  $w^T x_0 = w_0 - w_2 \frac{b}{a} < 0$  (porque  $x_0$  está en la región donde  $h(x) = -1$ )

y se deduce que

$$w_0 < w_2 \frac{b}{a}$$

$$w_2 \frac{b}{a} - w_0 > 0$$

Como  $w_2$  es distinto de 0, podemos dividir ambos lados por  $w_2$ , obteniendo:

$$\frac{b}{a} - \frac{w_0}{w_2} > 0$$

$$\frac{b}{a} > \frac{w_0}{w_2}$$

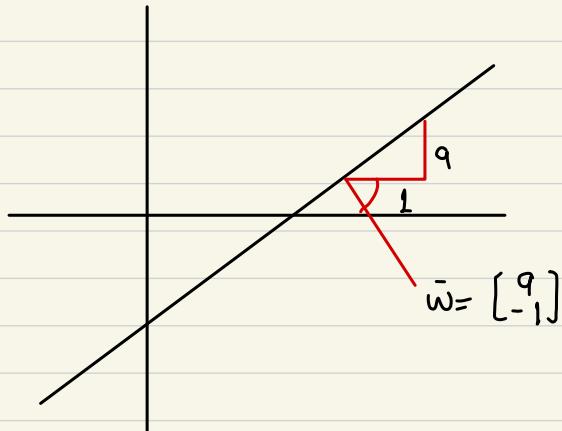
Por lo tanto la intersección en el eje  $y$   $b$  es mayor que  $\frac{w_0}{w_2}$  por esto es imposible si la línea recta que separa las regiones en el plano donde  $h(x) = 1$  y  $h(x) = -1$  pasa por el origen. Por lo tanto, concluimos que esta línea recta es perpendicular a  $w$  y pasa por el origen

entonces

$$x_2 = ax_1 + b \iff -ax_1 + x_2 + b = 0$$

$$x_2 \iff [a, -1] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b = 0$$

$$\iff [b - a] \begin{bmatrix} 1 \\ x_2 \end{bmatrix}$$



Si expresamos esta línea por la ecuación  $x_2 = ax_1 + b$ , ¿cuál es la pendiente a y la intercepción b en términos de  $w_0, w_1$  y  $w_2$ ?

La ecuación de la línea es  $x_2 = ax_1 + b$ . Podemos despejar a y b en términos de  $w_0, w_1$  y  $w_2$  de la siguiente manera:

$$x_2 = ax_1 + b$$

$$w_0 + w_1x_1 + w_2x_2 = 0 \leftarrow \text{cuando } x_2 \text{ está en la línea}$$

$$w_1x_1 + w_2x_2 = -w_0 - w_2x_2$$

$$w_2x_2 = w_1x_1 - w_0$$

$$x_2 = (-w_1/w_2)x_1 - (w_0/w_2)$$

Por lo tanto, la pendiente a es  $-w_1/w_2$  y la intercepción b es  $-w_0/w_2$  en términos de  $w_0, w_1, w_2$ .

$$y(x^T \bar{x})$$

Definimos que:

$$y(x) = 0$$

$$\Leftrightarrow (w_0, w_1, w_2) \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = 0$$

$$w_0 + w_1x_1 + w_2x_2 = 0$$

Este hiperplano separa el espacio  $\mathbb{R}_2$ , en 2 semiespacios definidos por

$$h(x) = 1 \quad y \quad h(x) = -1$$

Esto es

$$H_{+1} = \{x | y(x) > 0\}$$

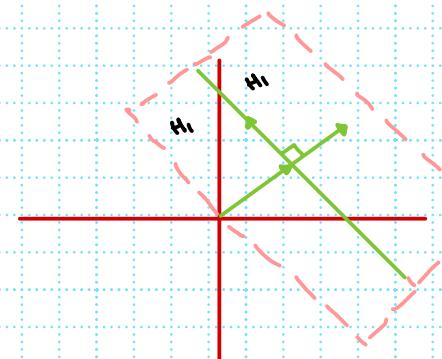
$$H_{-1} = \{x | y(x) < 0\}$$

Por lo tanto  $h(x) = \text{sign}(y(x)) = 1$  para

$$x \in H_{+1} \quad y \quad h(x) = \text{sign}(y(x)) = -1 \quad \text{para}$$

$$x \in H_{-1}$$

El vector  $\vec{w}$ , comienza desde cualquier lugar en el hipérplano, apunta a  $H+1$



(2)

Cuando estamos obteniendo el valor esperado del error para un nuevo dataset, y se tiene un entrenamiento previo, puede ser escrito como:

$$\mathbb{E}_x [\mathbb{E}_D [(g^{(0)}(x) - \bar{g}(x))^2] - (\bar{g}(x) - f(x))^2]$$

donde:

$\bar{g}(x)$  = Función promedio de hipótesis para  $n$  datasets entrenados

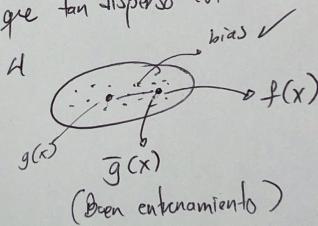
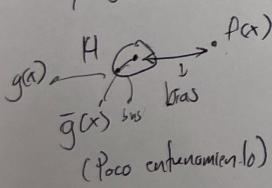
$f(x)$  = Función objetivo

$g^{(0)}(x)$  = Hipótesis para nuevo dataset

Si tenemos a bias como:

$$\text{bias} = \mathbb{E}_x [\mathbb{E}_D [(g^{(0)}(x) - \bar{g}(x))^2]]$$

que indica la separación en nuestro  $H$  de un nuevo dataset con respecto a  $\bar{g}(x)$ , por lo que buscamos que  $\bar{g}(x) \approx f(x)$  para poder utilizar con precisión el modelo. Si tenemos poco entrenamiento el bias  $(g^{(0)}(x) - \bar{g}(x))^2$  será grande, mientras que si existe buen entrenamiento,  $\bar{g}(x) \approx f(x)$  y podemos aplicar este modelo para medir que tan disperso está nuestro nuevo dataset y su hipótesis  $g^{(0)}(x)$ .



La varianza está dada por  $\mathbb{E}_x [(g(x) - f(x))^2]$  lo cual nos indica el sesgo de información para nuestro entrenamiento, si se tiene muy poca o nula varianza, sabremos que no hay mucho o bien entrenamiento, aunque tampoco puede ser muy grande porque entonces nuestra  $\bar{g}(x)$  está alejada de  $f(x)$ .

En general, cuando hemos entrenado nuestro modelo con muchos datasets ( $g(x)$ ) podemos utilizar ese modelo de descomposición de bias-varianza para ver lo lejano de nuestra nueva  $g(x)$  con respecto a nuestro  $\bar{g}(x) \approx f(x)$  para un análisis rápido.

### EJERCICIO 3

1) Muestre que si  $H$  es cerrado bajo la combinación lineal ( $g \in H$ ) entonces  $\bar{g} \in H$  donde

$$\bar{g} = \frac{1}{k} \sum_{k=1}^K g_k(x)$$

→ Para cualquier conjunto finito de hipótesis  $\{g_1, g_2, \dots, g_K\} \in H$  y cualquier constante  $\{a_1, a_2, \dots, a_K\}$ , la combinación lineal  $\{a_1 g_1, a_2 g_2, \dots, a_K g_K\}$

Definimos  $\bar{g}$  como el promedio de hipótesis en  $H$ , entonces:

$$\bar{g}(x) = \frac{1}{K} \sum_{k=1}^K g_k(x)$$

Y mostramos que  $\bar{g} \in H$ :

$\{g_1, g_2, \dots, g_K\} \rightarrow$  cualquier conjunto finito de hipótesis en  $H$ . Dado que  $H$  es cerrado bajo combinaciones lineales, la combinación lineal

$$g = a_1 g_1 + a_2 g_2 + \dots + a_K g_K \text{ para cualquier conjunto } \{a_1, a_2, \dots, a_K\} \in H$$

$$\hookrightarrow \bar{g} = \frac{1}{K} \sum_{k=1}^K g_k = \frac{1}{K} (g_1 + g_2 + \dots + g_K) \rightarrow \text{y si hacemos lo siguiente en la combinación lineal para } g_1 \text{ obtenemos:}$$

$$\left[ \begin{array}{l} g = \frac{1}{K} g_1 + \frac{1}{K} g_2 + \dots + \frac{1}{K} g_K \in H \end{array} \right]$$

Lo que implica

$$\hookrightarrow \bar{g} = \frac{1}{K} g_1 + \frac{1}{K} g_2 + \dots + \frac{1}{K} g_K \in H \text{ porque } H \text{ es cerrado bajo combinaciones lineales.}$$

Por lo tanto, se prueba que si  $H$  es cerrado bajo combinaciones lineales, entonces el promedio de hipótesis en  $H$ ,  $\bar{g}$  también está en  $H$  ( $\bar{g} \in H$ )

2) Modelo en donde la función promedio  $\bar{g}$  no está en el conjunto de hipótesis del modelo

Partimos de un conjunto de hipótesis  $H$  compuesto por funciones lineales de la forma  $h(x) = mx+b$ , donde  $m, b \in \mathbb{R}$ .  
 $H$  = conjunto de rectas en el plano.

Tomamos:

$$g_1(x) = x \text{ (recta que pasa por } (0,0) \text{ y } (1,1))$$

$$g_2(x) = x+1 \text{ (pasa por } (0,1) \text{ y } (1,2))$$

Su función promedio es:

$$\bar{g}(x) = \frac{(g_1(x) + g_2(x))}{2} = x + \frac{1}{2}$$

Por lo que  $\bar{g}(x)$  no está en  $H$  dado que no es una recta.

Pregunta 4. Para funciones binarias, muestre que  $P[h(x) \neq f(x)]$  puede ser escrita como el valor esperado de la métrica de error cuadrático medio en los siguientes casos:

1. La convención usada para la función binaria es 0 o 1 (en lugar de -1 o +1).

$$\begin{aligned}
 P[h(x) \neq f(x)] &= P[h(x) \neq f(x)] \cdot 1 + P[h(x) = f(x)] \cdot 0 \\
 &= P[h(x) \neq f(x)] (h(x) - f(x))^2 + P[h(x) = f(x)] (h(x) - f(x))^2 \\
 &= E[(h(x) - f(x))^2]
 \end{aligned}$$

### Error cuadrático

$$\frac{1}{N} \sum ((h(x) - f(x))^2)$$

$$\frac{1}{N} \sum (0 - 0)^2 = \frac{0}{N} = 0$$

$$\frac{1}{N} \sum (0 - 1)^2 = \frac{(-1)}{N} = \frac{1}{N}$$

$$\frac{1}{N} \sum (1 - 0)^2 = \frac{1^2}{N}$$

$$\frac{1}{N} \sum (1 - 1)^2 = \frac{0^2}{N}$$

Usando  $0 \circ 1$

$$\frac{1}{N} \sum_{n=1}^N [h(x_n) \neq f(x_n)]$$

$$f(x_n) \in \{0, 1\}$$

2. La convención usada para la función binaria es -1 o +1.

$$\begin{aligned} P[h(x) \neq f(x)] &= \frac{1}{4} P[h(x) \neq f(x)] \cdot 4 + \frac{1}{4} P[h(x) = f(x)] \cdot 0 \\ &= \frac{1}{4} P[h(x) \neq f(x)] (h(x) - f(x))^2 + \frac{1}{4} P[h(x) = f(x)] (h(x) - f(x))^2 \\ &= \frac{1}{4} E[(h(x) - f(x))^2] \end{aligned}$$

$$P(0 \neq 0)$$

$$P(1 \neq 0)$$

$$P(1 \neq 1)$$

$$P(0 \neq 0)$$

⑤ Para el error de regresión logística.

EQUIPO 1

$$E_{in}(\omega) = \frac{1}{N} \sum_{n=1}^N \ln (1 + e^{-y_n \omega^T x_n})$$

muestre que el gradiente está dado por

$$\nabla_{\omega} E_{in}(\omega) = -\frac{1}{N} \sum_{n=1}^N \frac{y_n x_n}{1 + e^{y_n \omega^T x_n}}$$

① Derivando de  $E_{in}(\omega) = \frac{1}{N} \sum_{n=1}^N \ln (1 + e^{-y_n \omega^T x_n})$  con respecto de  $\omega$ .

$$\begin{aligned} \frac{d E_{in}(\omega)}{d \omega} &= -\frac{1}{N} \sum_{n=1}^N \left( \frac{d \ln (1 + e^{-y_n \omega^T x_n})}{d \omega} \right) = \frac{1}{N} \sum_{n=1}^N \frac{1}{1 + e^{-y_n \omega^T x_n}} \cdot \frac{d (-y_n \omega^T x_n)}{d \omega} \\ &= \frac{1}{N} \sum_{n=1}^N \frac{1}{1 + e^{-y_n \omega^T x_n}} \cdot e^{-y_n \omega^T x_n} \cdot \frac{d (-y_n \omega^T x_n)}{d \omega} = -\frac{y_n x_n}{1 + e^{-y_n \omega^T x_n}} \end{aligned}$$

↑  
Se ve que  
negativo

② Por lo tanto, obtenemos:

$$\begin{aligned} \nabla E_{in}(\omega) &= -\frac{1}{N} \sum_{n=1}^N \frac{y_n x_n e^{-y_n \omega^T x_n}}{1 + e^{-y_n \omega^T x_n}} \Rightarrow = -\frac{1}{N} \sum_{n=1}^N \frac{y_n x_n}{1 + e^{y_n \omega^T x_n}} \\ &\quad \text{↓ El signo del exponente es invertido.} \quad \text{↓ Se demuestra //} \\ &= \frac{1}{N} \sum_{n=1}^N -y_n x_n \theta(-y_n \omega^T x_n). \end{aligned}$$

- ✓ Cuando una muestra está mal clasificada  $\Rightarrow y_n \omega^T x_n < 0$
- ✓ Si la muestra está bien clasificada  $\Rightarrow \theta(-y_n \omega^T x_n) \leq 0.5$

Por lo tanto, la contribución mal clasificada implementa de mejor manera el gradiente, que la 1<sup>a</sup> era clasificación.