Modelos de Deep Learning

Multi-Layer Perceptron

Universidad ORT Uruguay

1 de Setiembre, 2025

Modelo de una neurona

- Modelo paramétrico:

$$\widehat{y} = f\left(\mathbf{x}; \boldsymbol{\theta}\right)$$

  donde la función f está parametrizada por $\boldsymbol{\theta}$

- Regresión lineal - Modelo paramétrico básico de regresión

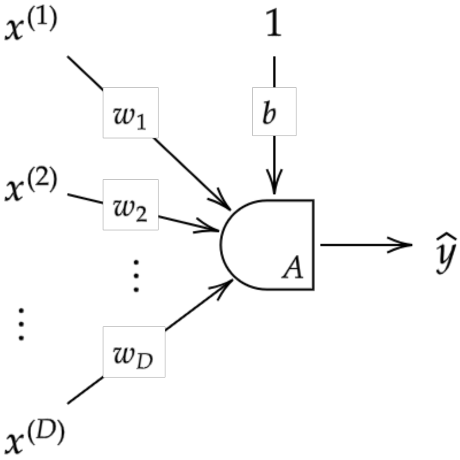
$$\widehat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_D x_D + b = \mathbf{x}^\top \mathbf{w} + b$$

- Regresión logística - Modelo paramétrico básico de clasificación

$$\widehat{y} = \text{Sigmoid}\left(w_1 x_1 + w_2 x_2 + \cdots + w_D x_D + b\right) = \text{Sigmoid}\left(\mathbf{x}^\top \mathbf{w} + b\right)$$

- En ambos modelos los parámetros son $\boldsymbol{\theta} = (\mathbf{w}, b)$

Perceptron: modelo de una neurona



Usando la notación matricial:

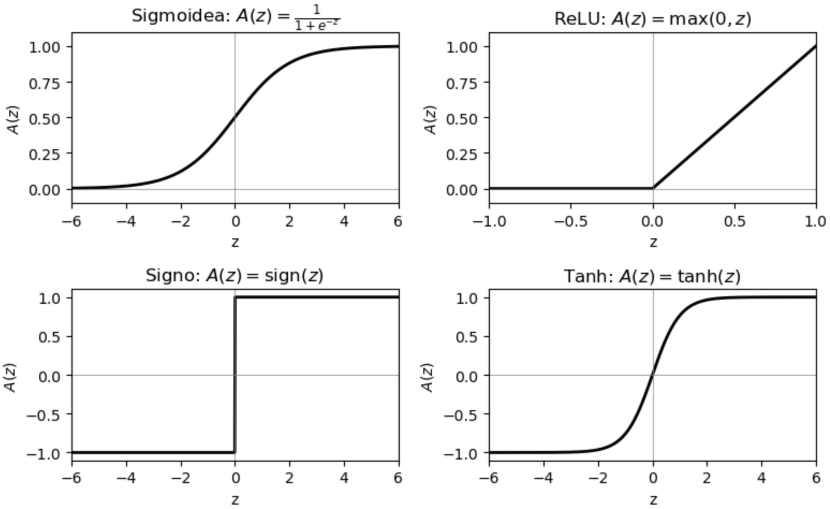
$$\widehat{y} = A\left(\mathbf{x}^\top \mathbf{w} + b\right)$$

en donde

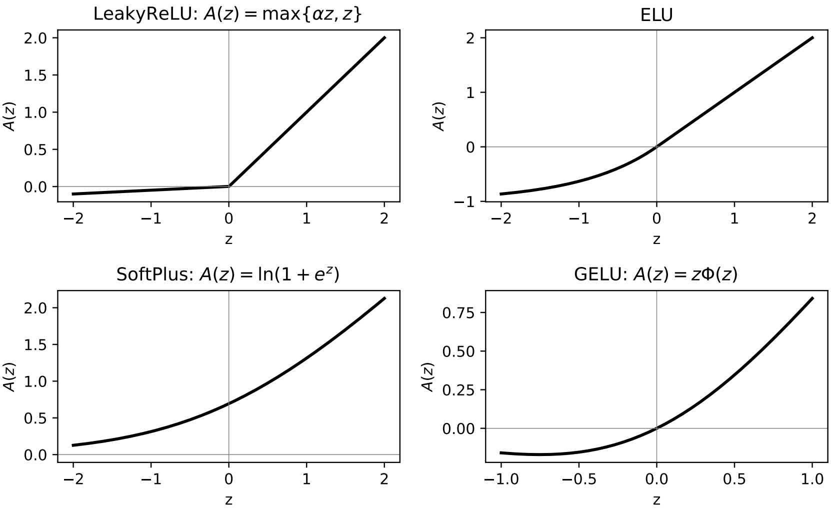
$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$$

$A : \mathbb{R} \to \mathbb{R}$ activación

Funciones de activación



Funciones de activación



Componente básica: Densa/Lineal

La **componente básica** de un MLP es una capa (layer) densa/lineal:

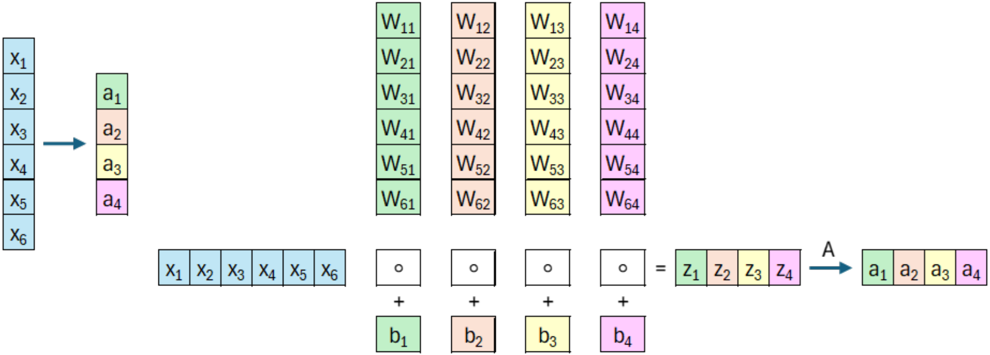
$$\mathbf{a} = D(\mathbf{x}; \mathbf{W}, \mathbf{b}) = A\left(\mathbf{x}^\top \mathbf{W} + \mathbf{b}^\top\right) \text{ con } \mathbf{x} \sim (I),\ \mathbf{a} \sim (O), \mathbf{W} \sim (I, O),\ \mathbf{b} \sim (O)$$

- la operación $\mathbf{z} = \mathbf{x}^\top \mathbf{W} + \mathbf{b}^\top$ es la unidad lineal
- A es la función de activación con $\mathbf{a} = A(\mathbf{z})$ point-wise

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_I \end{bmatrix} \mapsto \mathbf{z} = \begin{bmatrix} x_1 & \cdots & x_I \end{bmatrix} \begin{bmatrix} W_{11} & \cdots & W_{1O} \\ \vdots & \ddots & \vdots \\ W_{I1} & \cdots & W_{IO} \end{bmatrix} + \begin{bmatrix} b_1 & \cdots & b_O \end{bmatrix}$$

$$\mapsto \mathbf{a} = \begin{bmatrix} a_1 & \cdots & a_O \end{bmatrix}$$

Ejemplo Capa Densa



La cantidad de parámetros de D es $I \times O + O = (I+1) \times O$.

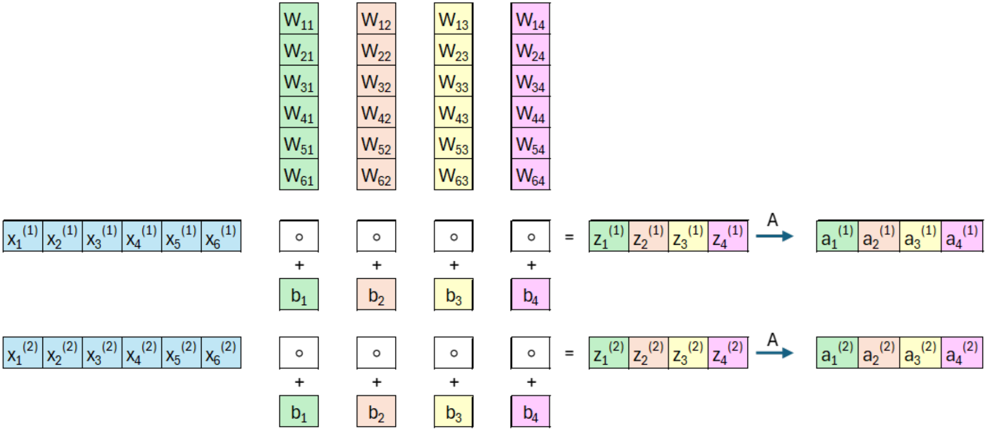
Operando en batch

Usando un batch de N inputs:

- $\mathbf{X} \sim (N, I)$, $\mathbf{a} \sim (N, O)$, $\mathbf{W} \sim (I, O)$, $\mathbf{b} \sim (O)$
- Aquí cada fila de $\mathbf{X}$ es un input de shape (I).
- A $\mathbf{b}$ se aplica broadcasting para tener shape (N, O)
- La salida de la capa es

$$\mathbf{a} = D(\mathbf{X}; \mathbf{W}, \mathbf{b}) = A\left(\mathbf{X}\mathbf{W} + \mathbf{b}\right) \sim (N, O)$$

Ejemplo Capa Densa Operación en Batch



Multi-layer Perceptron

- Un MLP de $d > 1$ *capas* se obtiene componiendo d capas *densas*:

$$\widehat{\mathbf{y}} = f(\mathbf{x}; \mathbf{W}_1, \ldots, \mathbf{W}_d, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_d) = D_d(D_{d-1}(\ldots, \mathbf{W}_{d-1}, \boldsymbol{b}_{d-1}); \mathbf{W}_d, \boldsymbol{b}_d)$$

- La dimensión de salida de O_i tiene que ser igual a la de entrada de I_{i+1}
- La cantidad de parámetros de $f(\mathbf{x})$ es:

$$\text{params}(f) = \sum_{i=1}^{d} \text{params}(D_i) = \sum_{i=1}^{d} (I_i + 1) O_i$$

con $\mathbf{W}_i \sim (I_i, O_i), \mathbf{b}_i \sim (O_i)$.