Redes Neuronales Artificiales
ANN

Mateo Tobón Henao
mtobonh@unal.edu.co

Juan Carlos Aguirre Arango
jucaguirrear@unal.edu.co

Universidad Nacional de Colombia - Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica, Electrónica y Computación

18 de noviembre de 2022

Tabla de Contenido

Tabla de Contenido

Definición del problema

Queremos encontrar una función $g : \mathbb{R}^p \rightarrow \mathbb{R}^q$, tal que esta mapeé el conjunto $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^p\}_{n=1}^N$ al conjunto $\mathbf{Y} = \{\mathbf{y}_n \in \mathbb{R}^q\}_{n=1}^N$

Definición del problema

Queremos encontrar una función $g : \mathbb{R}^p \to \mathbb{R}^q$, tal que esta mapeé el conjunto $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^p\}_{n=1}^N$ al conjunto $\mathbf{Y} = \{\mathbf{y}_n \in \mathbb{R}^q\}_{n=1}^N$
Ya que estamos usando aprendizaje profundo, esta g está construida en forma de red neuronal.

$$g(\mathbf{x}) = (f_L \circ f_{L-1} \circ \cdots \circ f_1)(\mathbf{x}) = \mathbf{y}$$

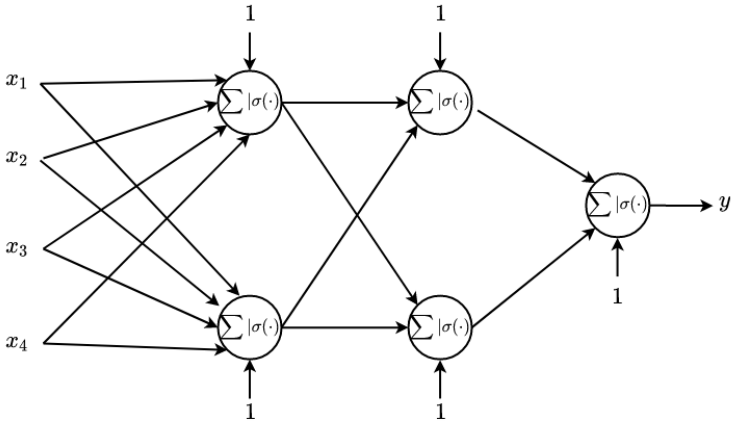
Para la primera capa

$$f_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \in \mathbb{R}^o$$

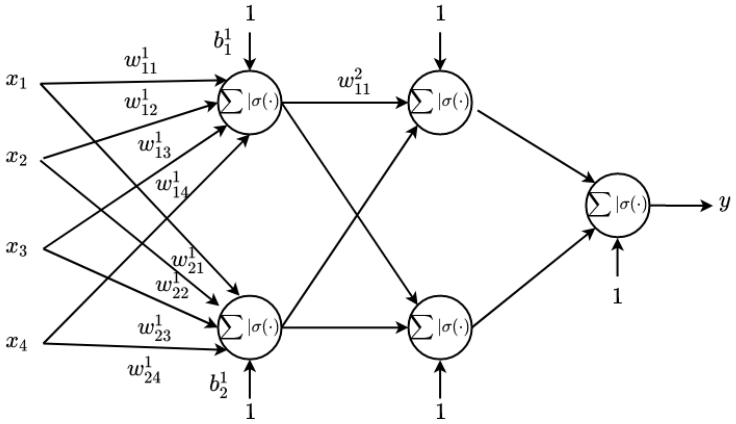
Donde σ es la función de activación, $\mathbf{W}_1 \in \mathbb{R}^{o \times p}$, y $\mathbf{b}_1 \in \mathbb{R}^o$. Por ende, el conjunto total de parámetros de nuestra función g (red neuronal) es $\Theta = \{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^L$

Tabla de Contenido

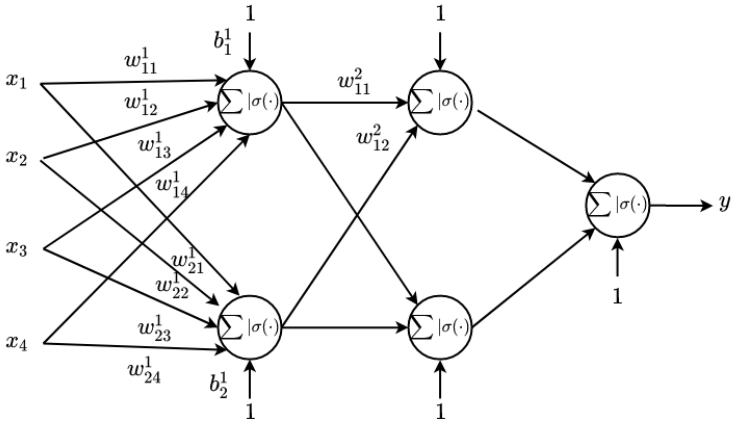
Propagación hacia adelante I



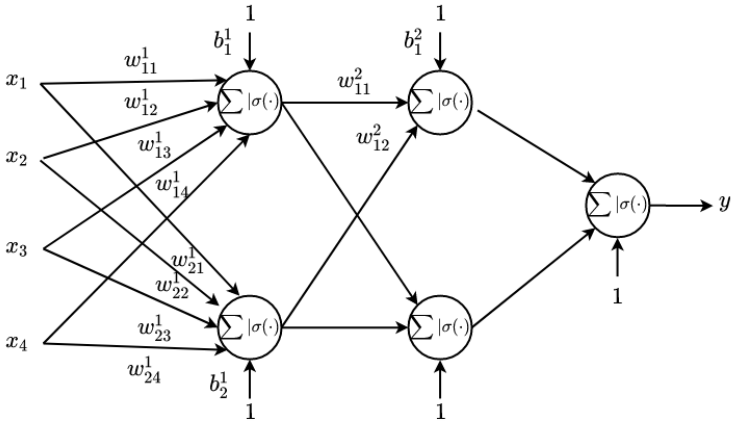
Propagación hacia adelante II



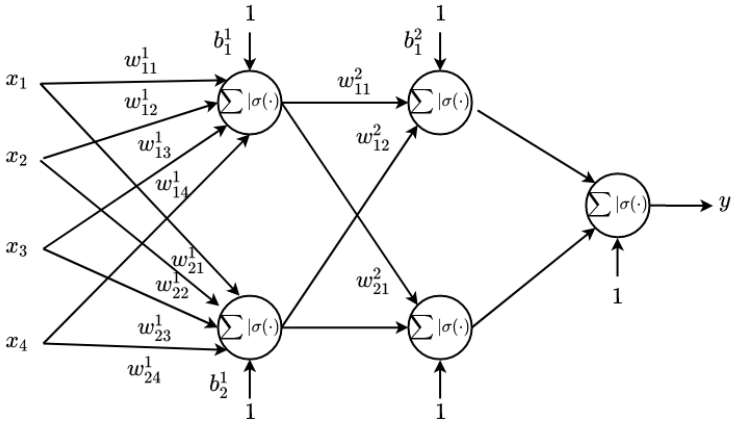
Propagación hacia adelante III



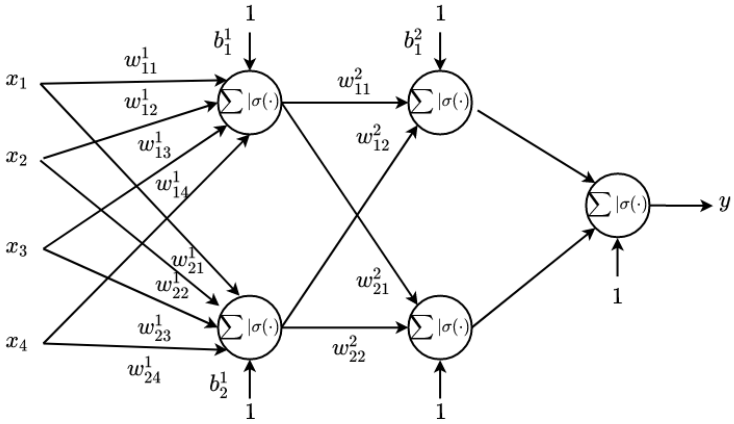
Propagación hacia adelante IV



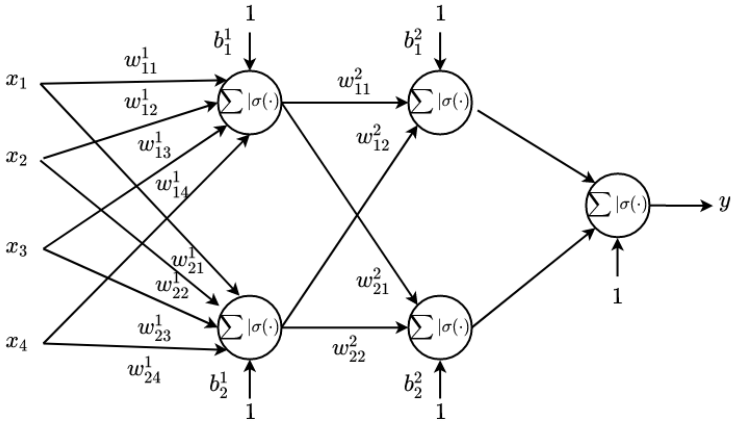
Propagación hacia adelante V



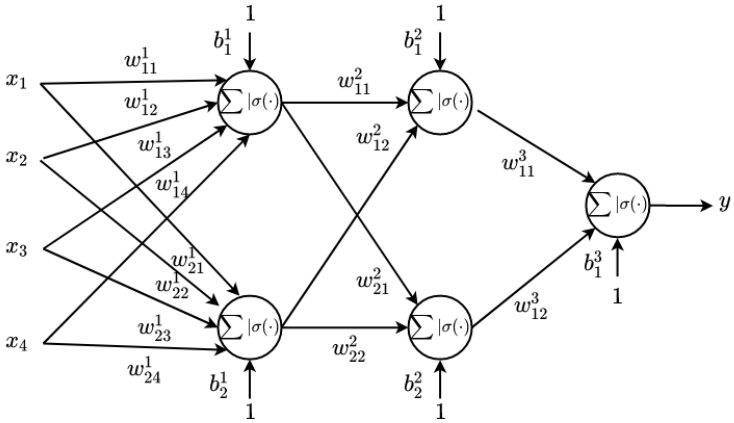
Propagación hacia adelante VI



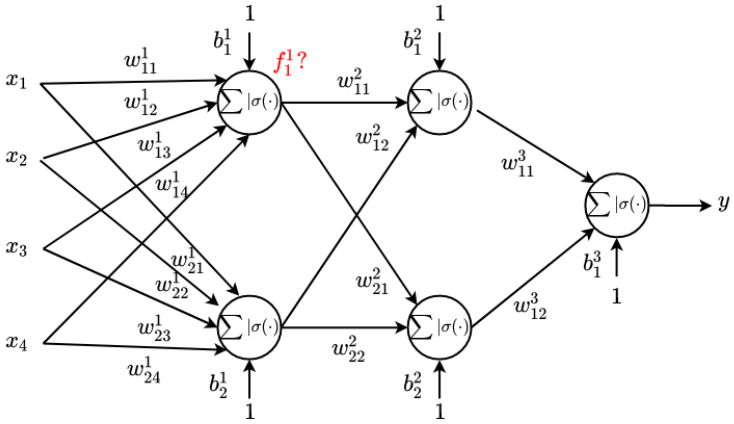
Propagación hacia adelante VII



Propagación hacia adelante VIII



Propagación hacia adelante IX

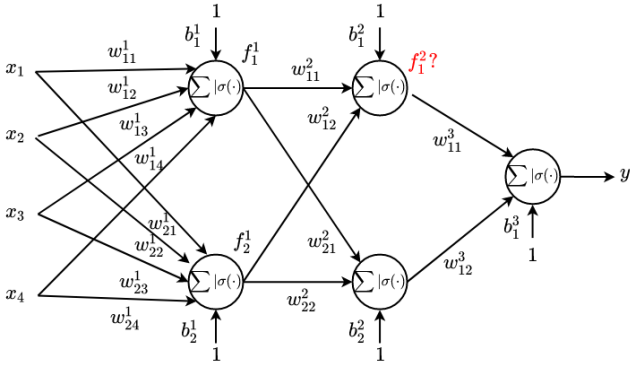


Propagación hacia adelante X

$$s = w_{11}^1 x_1 + w_{12}^1 x_2 + w_{13}^1 x_3 + w_{14}^1 x_4 + b_1^1$$

$$f_1^1 = \sigma(s)$$

Propagación hacia adelante XI



$$s = w_{11}^2 f_1^1 + w_{12}^2 f_2^1 + b_1^2$$

Propagación hacia adelante XII

$$f_1^2 = \sigma(s)$$

Propagación hacia adelante XIII

Tarea

Escribir en forma matricial la red neuronal anterior

$$y = \sigma(\mathbf{W}^3\mathbf{f}^2 + b^3) \text{ Donde } \mathbf{W}^3 \in \mathbb{R}^{1\times 2}, \mathbf{f}^2 \in \mathbb{R}^2, b^3 \in \mathbb{R}$$

Tabla de Contenido

Gradiente descendiente

¿Cómo encontramos el conjunto de parámetros Θ?

$$\Theta^* = \arg\min_{\Theta} C(\mathbf{Y}, \hat{\mathbf{Y}})$$

$$C(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \sum_{n=1}^{N} H(\mathbf{y}_n, g(\mathbf{x}_n))$$

Gradiente descendiente

Gradiente descendiente es un algoritmo de optimización usado pata minimizar alguna función. El funcionamiento básico de este algoritmo es moverse iterativamente en la dirección definida por el negativo del gradiente.

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \alpha \nabla f(\mathbf{x}_{t-1})$$

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right]^T$$

El gradiente nos dice cuanto crece una función en un punto determinado. Por ende, usamos el gradiente para saber en que dirección movernos. α es un **hiperpárametro** que nos ayuda a controlar que cuanto nos movemos, es conocido como la tasa de aprendizaje o learning rate en inglés.

Gradiente descendiente

Ejemplo:

$$f(x) = (x-4)^2 + 3$$

Gradiente descendiente

Ejemplo:

$$f(x) = (x-4)^2 + 3$$

$$x^* = \arg\min_x (x-4)^2 + 3$$

Gradiente descendiente

$$\frac{d((x-4)^2+3)}{dx} = 2(x-4)$$

$\alpha = 0{,}6, x_0 = 7$

Iteración 1:

$$x_1 = x_0 - 0{,}1(2(x_0 - 4))$$

$$x_1 = 7 - 0{,}1(2(7-4)) = 3{,}4$$

Iteración 2:

$$x_2 = 3{,}4 - 0{,}1(2(3{,}4-4)) = 4{,}12$$

Iteración 3:

$$x_3 = 4{,}12 - 0{,}1(2(4{,}12-4)) = 3{,}976$$

Iteración 4:

$$x_4 = 3{,}976 - 0{,}1(2(3{,}976-4)) = 4{,}0048$$

Gradiente descendiente

$$\frac{d((x-4)^2+3)}{dx} = 2(x-4) = 0$$

$$x^* = 4$$

Cuaderno Colab

Optimizadores

El gradiente descendiente es un optimizador (formas de encontrar el mínimo de una función con respecto a un parámetro), pero existen muchos más, cada uno intentando mejorar la eficiencia y evitar inconvenientes con los mínimos locales y puntos de silla. Visualización optmizadores

Tabla de Contenido

Regla de la cadena

Como hemos visto las redes neuronales son funciones compuestas, por ende para el cálculo de los gradientes recurrimos a la **regla de la cadena**.

$$h(x) = (f \circ g)(x) = f(g(x))$$

$$\frac{dh(x)}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}$$

Ejemplo: $h(x) = f(g(x))$ donde $f(x) = x^2$ y $g(x) = in(x) + 4$, encuentre $\frac{dh(x)}{dx}$

Regla de la cadena

Como hemos visto las redes neuronales son funciones compuestas, por ende para el cálculo de los gradientes recurrimos a la **regla de la cadena**.

$$h(x) = (f \circ g)(x) = f(g(x))$$

$$\frac{dh(x)}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}$$

Ejemplo: $h(x) = f(g(x))$ donde $f(x) = x^2$ y $g(x) = in(x) + 4$, encuentre $\frac{dh(x)}{dx}$
Solución:
$\frac{df(x)}{dx} = 2x$

Regla de la cadena

Como hemos visto las redes neuronales son funciones compuestas, por ende para el cálculo de los gradientes recurrimos a la **regla de la cadena**.

$$h(x) = (f \circ g)(x) = f(g(x))$$

$$\frac{dh(x)}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}$$

Ejemplo: $h(x) = f(g(x))$ donde $f(x) = x^2$ y $g(x) = in(x) + 4$, encuentre $\frac{dh(x)}{dx}$
Solución:
$\frac{df(x)}{dx} = 2x$, $\frac{dg(x)}{dx} = \frac{1}{x}$

Regla de la cadena

Como hemos visto las redes neuronales son funciones compuestas, por ende para el cálculo de los gradientes recurrimos a la **regla de la cadena**.

$$h(x) = (f \circ g)(x) = f(g(x))$$

$$\frac{dh(x)}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}$$

Ejemplo: $h(x) = f(g(x))$ donde $f(x) = x^2$ y $g(x) = in(x) + 4$, encuentre $\frac{dh(x)}{dx}$
Solución:
$\frac{df(x)}{dx} = 2x$, $\frac{dg(x)}{dx} = \frac{1}{x}$ Según la regla de la cadena

$$\frac{dh(x)}{dx} = \frac{df(in(x)+4)}{dg}(\frac{1}{x})$$

Regla de la cadena

Como hemos visto las redes neuronales son funciones compuestas, por ende para el cálculo de los gradientes recurrimos a la **regla de la cadena**.

$$h(x) = (f \circ g)(x) = f(g(x))$$

$$\frac{dh(x)}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}$$

Ejemplo: $h(x) = f(g(x))$ donde $f(x) = x^2$ y $g(x) = in(x) + 4$, encuentre $\frac{dh(x)}{dx}$
Solución:
$\frac{df(x)}{dx} = 2x$, $\frac{dg(x)}{dx} = \frac{1}{x}$ Según la regla de la cadena

$$\frac{dh(x)}{dx} = \frac{df(in(x)+4)}{dg}(\frac{1}{x})$$

$$\frac{dh(x)}{dx} = 2(in(x)+4)(\frac{1}{x})$$

Regla de la cadena

Escriba la derivada de $h(x) = (f \circ g \circ u \circ m)(x)$

Regla de la cadena

Escriba la derivada de $h(x) = (f \circ g \circ u \circ m)(x)$

Recordemos que nuestra red neuronal se define matemáticamente como una **función compuesta**, donde cada función es una capa de nuestra red.

$$g(\mathbf{x}) = (f_L \circ f_{L-1} \circ \cdots \circ f_1)(\mathbf{x}) = y$$

Para la primera capa

$$f_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \in \mathbb{R}^o$$

Regla de la cadena

Escriba la derivada de $h(x) = (f \circ g \circ u \circ m)(x)$

Recordemos que nuestra red neuronal se define matemáticamente como una **función compuesta**, donde cada función es una capa de nuestra red.

$$g(\mathbf{x}) = (f_L \circ f_{L-1} \circ \cdots \circ f_1)(\mathbf{x}) = y$$

Para la primera capa

$$f_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \in \mathbb{R}^o$$

función que queremos minimizar:

$$C(\mathbf{X}, \mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N} H(\mathbf{y}_n, g(\mathbf{x}_n))$$

¿Con respecto a que variables debemos derivar?

Regla de la cadena

Escriba la derivada de $h(x) = (f \circ g \circ u \circ m)(x)$

Recordemos que nuestra red neuronal se define matemáticamente como una **función compuesta**, donde cada función es una capa de nuestra red.

$$g(\mathbf{x}) = (f_L \circ f_{L-1} \circ \cdots \circ f_1)(\mathbf{x}) = y$$

Para la primera capa

$$f_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \in \mathbb{R}^o$$

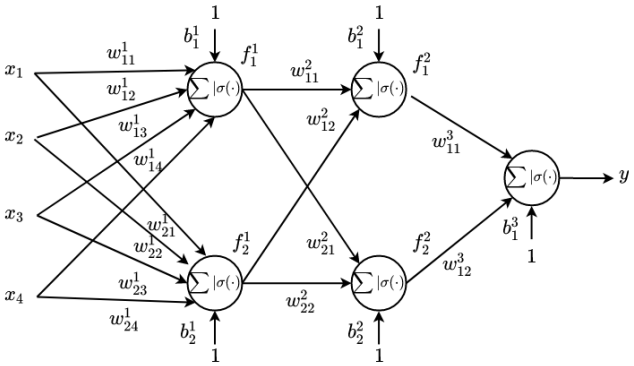
función que queremos minimizar:

$$C(\mathbf{X}, \mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N} H(\mathbf{y}_n, g(\mathbf{x}_n))$$

¿Con respecto a que variables debemos derivar? $\Theta = \{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^{L}$

Propagación hacia atrás I

Recordando el ejemplo inicial



Propagación hacia atrás II

El modelo matricial sería:

$$\mathbf{f}^1 = \sigma(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \in \mathbb{R}^2 \quad \mathbf{W}^1 \in \mathbb{R}^{2\times 4}, \mathbf{b}^1 \in \mathbb{R}^2$$

$$\mathbf{f}^2 = \sigma(\mathbf{W}^2\mathbf{f}^1 + \mathbf{b}^2) \in \mathbb{R}^2 \quad \mathbf{W}^2 \in \mathbb{R}^{2\times 2}, \mathbf{b}^2 \in \mathbb{R}^2$$

$$\mathbf{f}^3 = y = \sigma(\mathbf{W}^1\mathbf{f}^2 + \mathbf{b}^1) \in \mathbb{R} \quad \mathbf{W}^3 \in \mathbb{R}^{1\times 2}, \mathbf{b}^3 \in \mathbb{R}$$

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{W}^1$?

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{W}^1$?

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{W}^1}$$

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{W}^1$?

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3} \frac{\partial \mathbf{f}^3}{\partial \mathbf{W}^1}$$

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{W}^1$?

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3} \frac{\partial \mathbf{f}^3}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3} \frac{\partial \mathbf{f}^3}{\partial \mathbf{f}^2} \frac{\partial \mathbf{f}^2}{\partial \mathbf{W}^1}$$

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{W}^1$?

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3}\frac{\partial \mathbf{f}^3}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3}\frac{\partial \mathbf{f}^3}{\partial \mathbf{f}^2}\frac{\partial \mathbf{f}^2}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3}\frac{\partial \mathbf{f}^3}{\partial \mathbf{f}^2}\frac{\partial \mathbf{f}^2}{\partial \mathbf{f}^1}\frac{\partial \mathbf{f}^1}{\partial \mathbf{W}^1}$$

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{W}^1$?

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3}\frac{\partial \mathbf{f}^3}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3}\frac{\partial \mathbf{f}^3}{\partial \mathbf{f}^2}\frac{\partial \mathbf{f}^2}{\partial \mathbf{W}^1}$$

$$\frac{\partial C}{\partial \mathbf{W}^1} = \frac{\partial H}{\partial \mathbf{f}^3}\frac{\partial \mathbf{f}^3}{\partial \mathbf{f}^2}\frac{\partial \mathbf{f}^2}{\partial \mathbf{f}^1}\frac{\partial \mathbf{f}^1}{\partial \mathbf{W}^1}$$

Cálculo matricial

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{b}^1$?

Propagación hacia atrás

¿Cuál es la derivada de la función de costo con respecto a $\mathbf{b}^1$?

El nombre propagación hacia atrás (backporpagation en inglés) lleva este nombre porque se propaga la derivada del error por la red, cómo se vio en los ejemplos anteriores. Visualización de la propagación hacia atrás

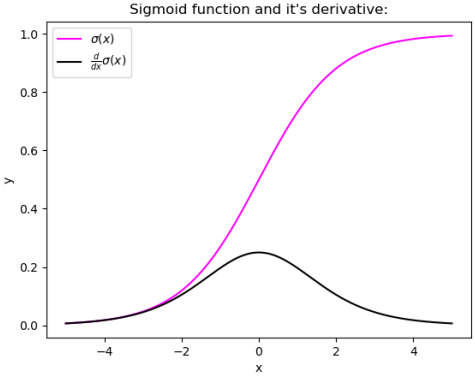
Tabla de Contenido

Desvanecimiento y Explosión del gradiente I

Para calcular los gradientes de los parámetros de la red es necesario usar la regla de la cadena, la cual es la multiplicación sucesiva de valores (gradientes).

- **Desvanecimiento**: El desvanecimiento del gradiente se presenta cuando los gradientes, a medida que se propagan, tiende a ser más y más pequeños, convirtiéndose en casi cero en las primeras capas, y por ende **los parámetros dejan de cambiar y el aprendizaje se detiene**. Por ello, en los modelos profundo se presenta este fenómeno. ¿Qué pasa cuando tenemos multiplicación de números sucesivos menores a cero?.

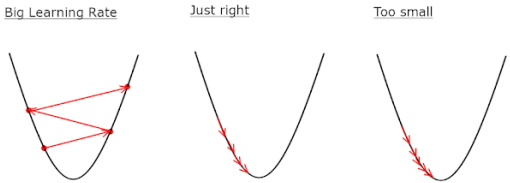
Desvanecimiento y Explosión del gradiente II



Este problema, esta relacionado con las funciones de activación. (¿por que es tan popular la ReLU?).

Desvanecimiento y Explosión del gradiente III

- **Explosión**: La explosión del gradiente, es un fenómeno contrario al anterior, aquí los gradientes tienden a tener valores muy elevados, y esto hace que el modelo diverja (en vez de acercase el mínimo se aleje). ¿Qué pasa cuando tenemos multiplicación de números sucesivos mayores a cero?.



La explosión del gradiente está relacionado con la tasa de aprendizaje.