

Análisis matemático

Clase 5

Análisis multivariado

Motivación

Muchos algoritmos de ML se basan en optimizar alguna función de costo, y encontrar, de acuerdo a este criterio, los mejores parámetros.

Para ello, nos va a interesar particularmente la idea de funciones, y el cálculo de derivadas.

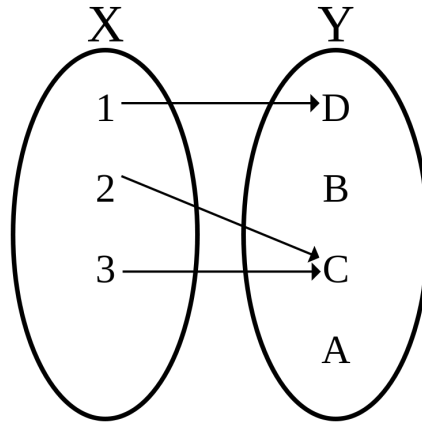
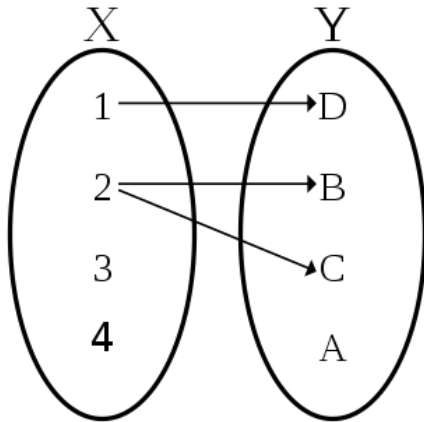
Función

Def: Una **función** es f es un elemento que vincula dos conjuntos, de forma tal que a cada elemento del primer conjunto se le asigna exactamente un valor del segundo.

Al primer conjunto se lo llama **dominio** de la función, mientras que el segundo conjunto es el **codominio**.

Funciones

Cada función queda unívocamente representada por el conjunto de todos los pares $(x, f(x))$. Notación: $f : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$, tal que $x \mapsto f(x)$.



Recuerdo la clase 2:
pido que se **inyectiva**

Tipos de funciones

Dada $f : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$:

Si $m = 1$ diremos que es una función **escalar**. Si además $n > 1$ se lo llama **campo escalar**. ([applet](#) campo escalar)

Si $m > 1$ diremos que es una función **vectorial**. Si además $n > 1$ se lo llama **campo vectorial**.

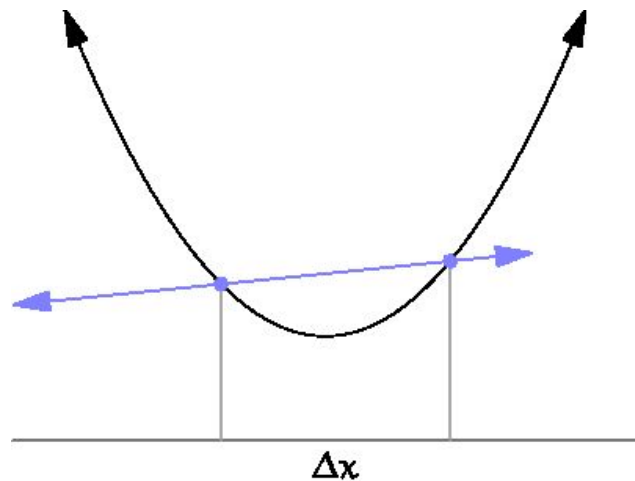
Funciones escalares univariadas

Repaso: Funciones escalares univariadas

Def: Sea f una función tal que $x \mapsto f(x)$, para $h > 0$, la **derivada** de f en x se define como

$$\frac{df}{dx} := \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Recuerdo: la derivada de f es la tangente de la curva.



Repaso: propiedades de la diferenciación

Product rule: $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$

Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$

Sum rule: $(f(x) + g(x))' = f'(x) + g'(x)$

Chain rule: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$

Campos escalares

Algunos ejemplos

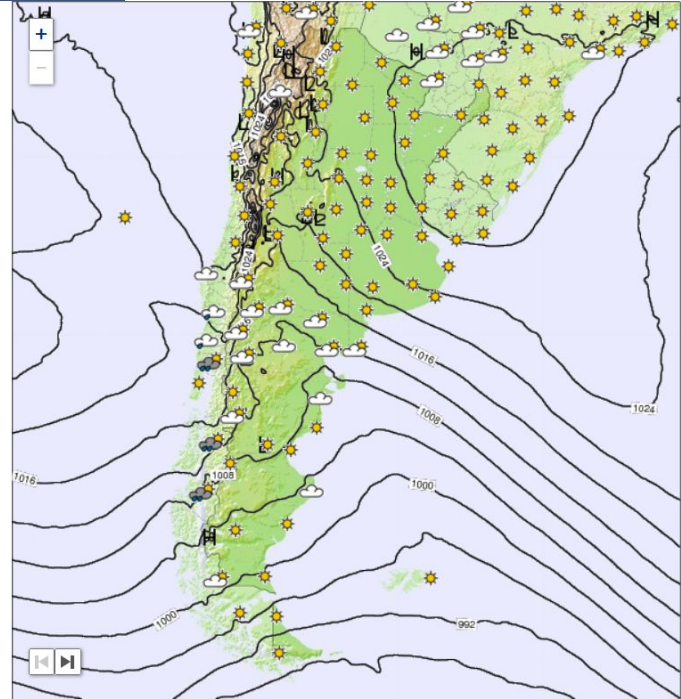
- Campos de temperatura, humedad y presión
- Campos potencial eléctrico, campo gravitacional
- Funciones de costo

Conjuntos de nivel

Fuente

Dada $f : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ un conjunto de nivel k de f es L_k definido por:

$$L_k := \{X \in \mathbb{R}^n / X \in D \wedge f(X) = k\}.$$



Funciones escalares multivariadas

Def: Sea $f : \mathbb{R}^n \rightarrow \mathbb{R}$ una función, $x \mapsto f(x)$, con $x = [x_1, \dots, x_n]$. Se definen las **derivadas parciales** como

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \rightarrow 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \rightarrow 0} \frac{f(x_1, x_2, \dots, x_n + h) - f(x_1, x_2, \dots, x_n)}{h} \end{aligned}$$

Gradiente

Def: Definimos el **gradiente** como

$$\nabla f = \text{grad}(f) = \frac{d f}{d x} = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right]$$

Obs: Para calcular las derivadas parciales podemos usar las propiedades de diferenciación para escalares

Obs: El gradiente apunta en la dirección de máximo crecimiento.

Propiedades de las derivadas parciales

Regla del producto: $\frac{\partial}{\partial \mathbf{x}} (f(\mathbf{x})g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} g(\mathbf{x}) + f(\mathbf{x}) \frac{\partial g}{\partial \mathbf{x}}$

Regla de la suma: $\frac{\partial}{\partial \mathbf{x}} (f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}}$

Regla de la cadena: $\frac{\partial}{\partial \mathbf{x}} (g \circ f)(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} (g(f(\mathbf{x}))) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial \mathbf{x}}$

Regla de la cadena en forma matricial

Sea $f(x_1, x_2)$ una función de x_1 y x_2 , donde $x_1 = x_1(s, t)$ y $x_2 = x_2(s, t)$. De acuerdo con la regla de la cadena

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial s}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t}$$

Luego:

$$\frac{df}{d(s,t)} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial (s,t)} = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \right] \overbrace{\left[\begin{array}{cc} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{array} \right]}^{\frac{\partial \mathbf{x}}{\partial (s,t)} = \begin{bmatrix} \nabla_{(s,t)} x_1 \\ \nabla_{(s,t)} x_2 \end{bmatrix}}.$$

Campos vectoriales

Campos vectoriales

Sea una función $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m > 1$. f se puede escribir como

$$f(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix},$$

donde cada $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ es un campo escalar, y por lo tanto las reglas de diferenciación para cada f_i son las vistas en las diapositivas anteriores.

Jacobiano

$$\frac{\partial f}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \rightarrow 0} \frac{f_1(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f_1(\mathbf{x})}{h} \\ \vdots \\ \lim_{h \rightarrow 0} \frac{f_m(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f_m(\mathbf{x})}{h} \end{bmatrix}.$$

Def: Se define el **Jacobiano** como la derivada de f respecto de \mathbf{x} , que resulta

$$J = \nabla_{\mathbf{x}} f = \frac{d f}{d \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Interpretación

Recordemos que $\det(A)$ nos servía para saber como se escalan las áreas al aplicar una transformación lineal.

Sea f función tal que $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \mathbf{x} \mapsto \mathbf{y} = f(\mathbf{x})$. El jacobiano de f nos dice cómo se modifica \mathbf{y} ante una perturbación en \mathbf{x} . Si la transformación es lineal, J nos da exactamente la transformación que estamos buscando. Si no, nos aproxima localmente de manera lineal a la transformación. El determinante de J nos da el factor de escala de áreas en la transformación.

Ejemplo que ya conocemos

Tomemos por ejemplo el método del Jacobiano que vimos en Probabilidad, que decía que, dado el vector aleatorio $[X, Y]^T$ dos con función de densidad conjunta $f_{X,Y}(x, y)$, y dadas dos funciones uno a uno $u = g_1(x, y)$ y $v = g_2(x, y)$, la función de densidad conjunta del vector aleatorio

$[U, V]^T = [g_1(X, Y), g_2(X, Y)]^T$ resulta

$$f_{U,V}(u, v) = \frac{f_{X,Y}(x,y)}{|J|} \Big|_{(x,y)=(g_1^{-1}(u,v), g_2^{-1}(u,v))}.$$

Ejemplo que ya conocemos

Se puede demostrar utilizando los conceptos de diferenciación.

Partamos de que

$$f_Y(y) = \frac{dF_Y}{dy}, \text{ pero } y = g(x).$$

Además, $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$. Si la transformación g es estrictamente creciente,

$$F_Y(y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Finalmente, podemos aplicar la regla de la cadena para obtener


$$f_Y(y) = \frac{dF_Y}{dy} = \frac{dF_Y}{dg^{-1}(x)} \frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}.$$

Si g hubiera sido decreciente, nos quedaban los signos cambiados.

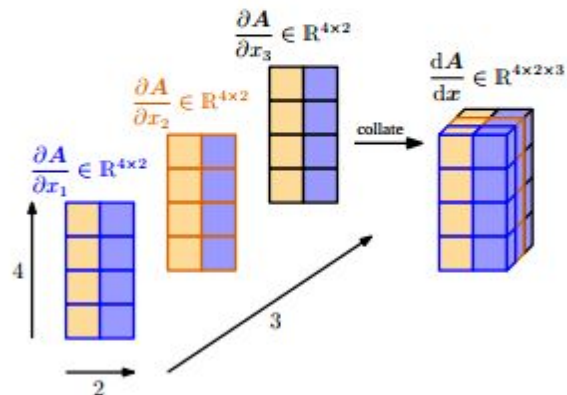
Para poder escribir una expresión genérica es que se incorpora el valor absoluto de la derivada.

Derivando matrices

Derivación de matrices

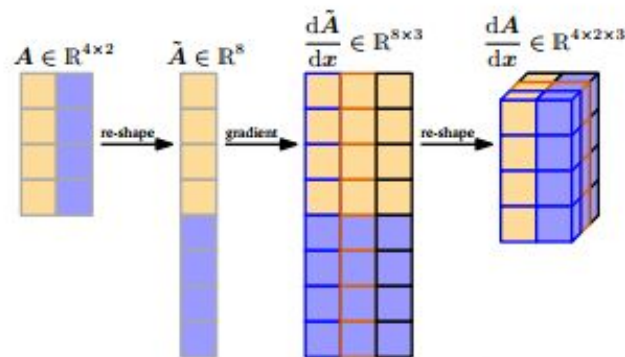
$$A \in \mathbb{R}^{4 \times 2} \quad x \in \mathbb{R}^3$$


Partial derivatives:



(a) Approach 1: We compute the partial derivative $\frac{\partial A}{\partial x_1}$, $\frac{\partial A}{\partial x_2}$, $\frac{\partial A}{\partial x_3}$, each of which is a 4×2 matrix, and collate them in a $4 \times 2 \times 3$ tensor.

$$A \in \mathbb{R}^{4 \times 2} \quad x \in \mathbb{R}^3$$

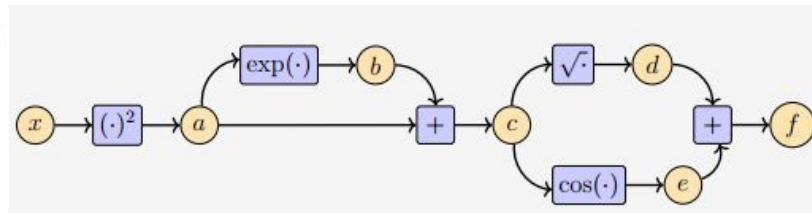



(b) Approach 2: We re-shape (flatten) $A \in \mathbb{R}^{4 \times 2}$ into a vector $\tilde{A} \in \mathbb{R}^8$. Then, we compute the gradient $\frac{d\tilde{A}}{dx} \in \mathbb{R}^{8 \times 3}$. We obtain the gradient tensor by re-shaping this gradient as illustrated above.

Diferenciación automática

Diferenciación automática

Sean x_1, \dots, x_d las variables de entrada a una función, x_{d+1}, \dots, x_{D-1} son variables intermedias y x_D es la variable de salida.



donde g_i son funciones elementales y $x_{Pa(x_i)}$ son los nodos padres de x_i . Esto define un **grafo de cómputo**. Recordando que $f = D$, tenemos que $\frac{\partial f}{\partial x_D} = 1$. Para las otras variables x_i aplicamos la regla de la cadena:

$$\frac{\partial f}{\partial x_i} = \sum_{x_j: x_i \in Pa(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial x_j}{\partial x_i} = \sum_{x_j: x_i \in Pa(x_j)} \frac{\partial f}{\partial g_j} \frac{\partial g_j}{\partial x_i}$$

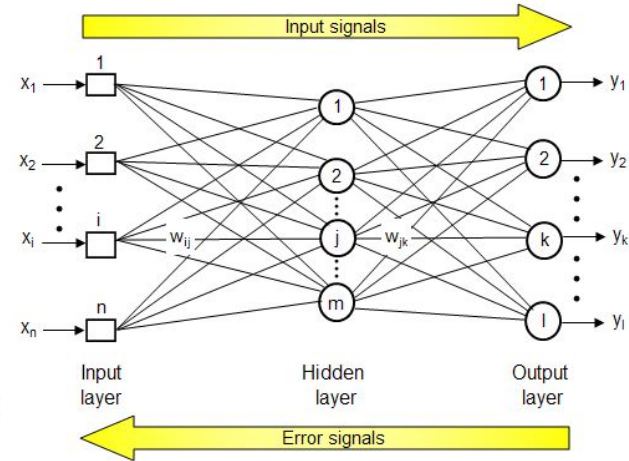
La **diferenciación automática** sirve siempre que la función pueda representarse como un grafo de cómputo.

Backpropagation

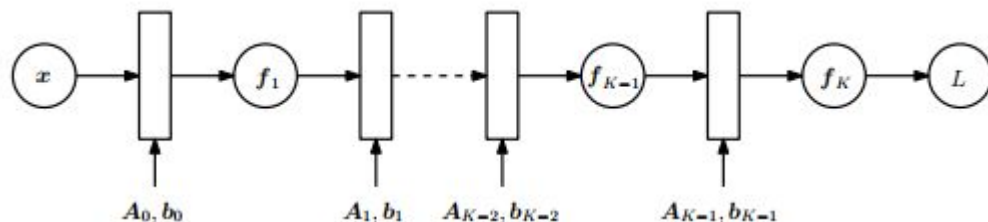
¿Dónde se aplica la diferenciación automática?

Backpropagation

Se utiliza para computar los gradientes necesarios para actualizar los parámetros a entrenar de las redes neuronales profundas. El objetivo es calcular las derivadas de la función costo respecto de los parámetros de la red neuronal. En este caso, las variables intermedias serían las activaciones de las distintas capas de la red.



Backpropagation



$$f_0 := x$$

$$f_i := \sigma_i(A_{i-1}f_{i-1} + b_{i-1}), \quad i = 1, \dots, K,$$

$$L(\theta) = \|y - f_K(\theta, x)\|^2 \quad \theta_j = \{A_j, b_j\}$$

$$\frac{\partial L}{\partial \theta_{K-1}} = \frac{\partial L}{\partial f_K} \frac{\partial f_K}{\partial \theta_{K-1}}$$

$$\frac{\partial L}{\partial \theta_{K-2}} = \frac{\partial L}{\partial f_K} \boxed{\frac{\partial f_K}{\partial f_{K-1}} \frac{\partial f_{K-1}}{\partial \theta_{K-2}}}$$

$$\frac{\partial L}{\partial \theta_{K-3}} = \frac{\partial L}{\partial f_K} \frac{\partial f_K}{\partial f_{K-1}} \boxed{\frac{\partial f_{K-1}}{\partial f_{K-2}} \frac{\partial f_{K-2}}{\partial \theta_{K-3}}}$$

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial f_K} \frac{\partial f_K}{\partial f_{K-1}} \dots \boxed{\frac{\partial f_{i+2}}{\partial f_{i+1}} \frac{\partial f_{i+1}}{\partial \theta_i}}$$

Bibliografía

Bibliografía

- Apunte AMII - FIUBA
- "Deep Learning", de Ian Goodfellow, Yoshua Bengio y Aaron Courville. Sección 6.4