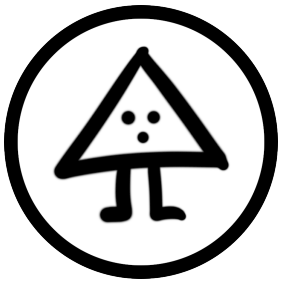
Matrices y Machine Learning

Apuntes y Ejercicios

Benja Vera



$$f(x) = \sigma(Wx + b)$$

Versión actualizada al
8 de octubre de 2023

Contenidos

Capítulo 1

Introducción

Este pequeño curso tiene el objetivo de explorar los conceptos principales de la matemática de primeros años de universidad (esto es, cálculo en una variable, álgebra lineal y elementos del cálculo en varias variables) guiado por el ejemplo/objetivo de construir con ellos código que permita entrenar una red neuronal. Así, a nivel transversal, con este curso buscamos lograr varias cosas distintas.

lalalal

1.1. ¿Qué es una red neuronal?

1.2. Aprendizaje y minimización

1.3. Hoja de ruta

Capítulo 2

Cálculo en una Variable

En este primer capítulo, se resumen los elementos principales del cálculo en una variable que son necesarios para desarrollar la teoría de la optimización. Los dos algoritmos principales de minimización que se exploran son el **método de descenso** y el **método de Newton**. Siendo el primero sin duda el más utilizado en el mundo del machine learning hoy en día, aunque algunos autores han sugerido utilizar variaciones del otro (ver por ejemplo [Le+11]). En la sección 2.1 se desarrolla el concepto intuitivo de la derivada de una función, junto con un resumen de las propiedades que esperaríamos que esta cumpliera. Luego, en la sección 2.2 hablamos más bien de cómo realmente calcular la derivada de una función, estas ideas se enlazan con la sección anterior a través de un ejemplo de cómo las primeras dos derivadas de una función se pueden utilizar para esbozar su gráfico completo, el capítulo termina con una discusión sobre la **regla de la cadena**, concepto fundamental para lo que sigue. Después de esto, la sección 2.3 aplica lo discutido sobre la regla de la cadena para construir los llamados **polinomios de Taylor** asociados a una función, los cuales vamos interpretar como buenas aproximaciones de la función cerca de un cierto punto x_0. El capítulo termina en la sección 2.4, la cual expone los dos algoritmos principales de optimización mencionados al inicio.

Las ideas mencionadas en este capítulo se encuentran más detalladas, por ejemplo, en [GOM23]. Pero para una introducción completamente rigurosa al cálculo de una variable, se recomienda ver lo expuesto en [SM88]

2.1. La derivada y sus propiedades

2.2. Las reglas de derivación

2.3. Polinomios de Taylor

2.4. Algoritmos de optimización

2.5. Ejercicios

2.5.1. Preliminares

1. Analice, mediante una tabla, los signos de las siguientes expresiones ya factorizadas

   a) $\frac{x+1}{x-1}$
   b) $(x+2)(x-3)$
   c) $(x+4)(1-x)$
   d) $\frac{(x+1)(x-1)}{x-3}$

2. Factorice las siguientes expresiones y luego analice sus signos como en el item anterior

   a) $\frac{x^2-4}{x+1}$
   b) x^2-4x+3
   c) x^2-3x

2.5.2. Cálculo de derivadas por definición

Entre las dos definiciones de derivada que hemos visto:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}, \qquad f'(x) = \lim_{w\to x} \frac{f(w)-f(x)}{w-x}$$

Utilice la que más le convenga para calcular las derivadas de las siguientes funciones

1. $f(x) = \frac{1}{x}$
2. $f(x) = \sqrt{x}$
3. $f(x) = x^3$

2.5.3. Reglas de derivación sin regla de la cadena

Utilizando las reglas de derivación, calcule las derivadas de las siguientes funciones

1. $f(x) = x^4$
2. $f(x) = 3x^5 - x^3$
3. $f(x) = 4x^2 - 3\pi^2$
4. $f(x) = (x+1)(x-1)$
5. $f(x) = \frac{x+1}{x-1}$
6. $f(x) = \frac{1}{\sqrt{x}}$

2.5.4. La regla de la cadena

Utilizando las reglas de derivación conocidas, además de la regla de la cadena, obtenga las derivadas de las siguientes funciones.

1. $f(x) = (1 + \sqrt{x})^2$
2. $f(x) = \sqrt[3]{2x}$
3. $f(x) = (4 + 2x)^{2023}$
4. $f(x) = \sqrt{1 - x^2}$
5. $f(x) = (\frac{1}{x} + x)^3$
6. $f(x) = (1 - x)^5$

2.5.5. Problemas varios sobre rectas tangentes

Para esta sección, recuerde que la recta tangente al gráfico de una función f en el punto $(x_0, f(x_0))$ viene dada por la ecuación

$$y = f(x_0) + f'(x_0)(x - x_0)$$

Además, recordemos también el hecho de que si L_1 y L_2 son dos rectas dadas respectivamente por

$$y = m_1 x + n_1, \qquad y = m_2 x + n_2$$

Entonces estas rectas son perpendiculares cuando $m_1 \cdot m_2 = -1$

1. Considere el gráfico de la función $f(x) = \frac{1}{x}$ y sea $x_0 > 0$ fijo.
   a) Encuentre la ecuación de la recta tangente al gráfico de f en el punto $(x_0, 1/x_0)$. Debería obtener una recta cuyos parámetros vienen escritos en términos de x_0.
   b) Obtenga los puntos de intersección entre la recta obtenida anteriormente y los ejes de coordenadas. Recuerde que esto se puede hacer imponiendo $x = 0$ y $y = 0$ según corresponda. Debería obtener dos puntos cuyas coordenadas vienen dadas en términos de x_0.
   c) Al dibujar lo que está sucediendo, notará que se forma un triángulo (dado por la recta tangente y los ejes de coordenadas). Calcule su área ¿Qué puede decir con respecto a este área?
2. En este problema, vamos a probar el hecho de que las rectas tangentes a los círculos son siempre perpendiculares al radio. Hecho que ya se conoce de la geometría euclideana. Recordemos para esto que la ecuación que define a una circunferencia de radio 1 centrada en el punto $(0, 0)$ viene dada por

   $$x^2 + y^2 = 1$$

   De modo que depejando y, la función que define al semicírculo superior de esta circunferencia viene dada por

   $$f(x) = \sqrt{1 - x^2}$$

   Dado $x_0 \in]-1, 1[$ fijo, proceda como sigue:
   a) Considere el radio que une el origen con el punto $(x_0, f(x_0))$. Conociendo dos puntos, calcule la pendiente de este radio.
   Nota: Es posible calcular también la ecuación de la recta que define a este radio, pero no la necesitamos realmente.
   b) Calcule además la derivada de la función f en el punto x_0.
   c) Interpretando sus resultados anteriores, concluya lo pedido.

2.5.6. Máximos y mínimos

1. Pruebe, utilizando lo expuesto en este capítulo, el hecho ya conocido de que si $f(x) = ax^2 + bx + c$ es una función cuadrática cualquiera con $a \neq 0$, entonces el punto $x_0 = \frac{-b}{2a}$ corresponde a un mínimo si $a > 0$ y un máximo si $a < 0$.

2. Se desea encontrar el punto $P = \left(x, \frac{1}{x}\right)$ con $x > 0$ (construido así de modo que P pertenece al gráfico de la función $f(x) = 1/x$) cuya distancia euclideana al origen de coordenadas sea mínima. Para esto, y teniendo en cuenta la siguiente figura, proceda como sigue:

   [FIGURA]

   a) Pruebe mediante el teorema de pitágoras que para $x > 0$ fijo, la distancia viene dada por

   $$d(x) = \sqrt{x^2 + \frac{1}{x^2}}$$

   b) Obtenga la derivada de esta función $d'(x)$ y resuelva la ecuación $d'(x) = 0$. Llamemos x_0 al valor obtenido.

   c) Con este valor en mano, calcule $d''(x_0)$ y concluya.

2.5.7. Gráficos de funciones

2.5.8. Polinomios de Taylor

Capítulo 3

Álgebra Lineal

En el capítulo anterior dimos una visita a todas las herramientas de optimización que son comunes en el cálculo de una variable. Ahora haremos un salto a un tema completamente nuevo y aparentemente no relacionado que es el Álgebra Lineal[1], exploraremos brevemente la teoría de matrices y transformaciones lineales y el siguiente capítulo, sobre el cálculo y los métodos de optimización en varias variables, se encargará de enlazar estos dos círculos de ideas. Por el momento, volveremos sobre nuestro problema original para entender un poco más sobre las operaciones involucradas.

3.1. La necesidad de las matrices

Recordemos el esquema básico de cómo actúa una red neuronal. En la figura [LAL] se muestra cómo, a partir de una primera columna de valores percibidos por la red, estos valores se pueden combinar en forma de sumas y productos para obtener la así llamada *activación* de cada neurona de la capa siguiente. En la introducción vimos que este proceso en realidad tiene más pasos, ya que hay más operaciones que se hacen a partir de esta activación obtenida. Pero ya que este proceso de multiplicar y sumar para cada neurona parece ser el más complejo, es bueno aislarlo y analizarlo más de cerca por un momento.

[FIGURA]

Supongamos que tenemos n neuronas de entrada en la capa L_0 y la siguiente capa L_1 tiene m neuronas. Para $i \in \{1, \dots, m\}$ y $j \in \{1, \dots n\}$, sea a_j la activación de la neurona j-ésima de la capa L_0 y b_i la activación de la neurona i-ésima de la capa L_1. Así, si w_{ij} denota el peso por el que se pondera la activación a_j en el cálculo de b_i, tenemos la fórmula

$$b_i = w_{i1}a_1 + w_{i2}a_2 + \dots w_{in}a_n = \sum_{j=1}^{n} w_{ij}a_j$$

Escribamos esta fórmula de manera un poco más extendida como sigue

$$\begin{aligned} b_1 &= w_{11}a_1 + w_{12}a_2 + \dots w_{1n}a_n \\ b_2 &= w_{21}a_1 + w_{22}a_2 + \dots w_{2n}a_n \\ &\vdots \\ b_m &= w_{m1}a_1 + w_{m2}a_2 + \dots w_{mn}a_n \end{aligned}$$

[1] Existen muchos textos clásicos sobre álgebra lineal, pero el que se recomienda más seguido es [HK04]

Seguramente a primera vista, este proceso se ve bastante complejo e intimidante. Pero para entenderlo desde un mejor panorama, nos gustaría empaquetar a todas las activaciones de las capas en vectores como sigue

$$\vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

Y preguntarnos por cuál es la operación que nos permite obtener el vector $\vec{b}$ a partir del vector $\vec{a}$. El sistema de ecuaciones de arriba es, por supuesto, el que describe esta operación. Pero a partir de mirarlo, vemos que la operación está completamente determinada por los w_{ij}. En otras palabras, si definimos el objeto

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & & w_{2n} \\ \vdots & & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{pmatrix}$$

Junto con la operación $W\vec{v}$ definida para vectores $\vec{v}$ de largo n y que entrega vectores de largo m dada por

$$(W\vec{v})_j = w_{j1}v_1 + w_{j2}v_2 + \dots w_{jn}v_n = \sum_{i=1}^{n} w_{ji}v_i$$

Entonces la fórmula para el vector $\vec{b}$ se puede escribir como

$$\vec{b} = W\vec{a}$$

Forma que sin duda suena mucho más simple de manejar que el sistema de ecuaciones que teníamos anteriormente. Formalicemos lo que acabamos de hacer con una pila de definiciones:

Definiciones (Espacio $\mathbb{R}^n$, matriz, multiplicación vector-matriz)

Los vectores de largo n con componentes reales (que son con los que casi siemore trabajamos), forman el conjunto $\mathbb{R}^n$ que suele recibir el nombre de *espacio euclideano de n dimensiones*. El objeto W que acabamos de definir recibe el nombre de *matriz de* $m \times n$, el conjunto de todas las cuales se denota $\mathbb{R}^{m\times n}$. La operación de la que vino acompañado este objeto (y que como vimos, fundamenta su existencia) se llama *multiplicación vector-matriz* y constituye una función $\mathbb{R}^n \to \mathbb{R}^m$. En el caso en que m o n sean 1, identificamos las matrices de 1×1 o los vectores de largo 1 con simplemente números reales. De modo que si $A \in \mathbb{R}^{1\times n}$ y $\vec{v} \in \mathbb{R}^n$, podemos libremente entender que $A\vec{v}$ es un número real. También en adelante nos olvidaremos de las flechas al denotar un vector, ya que resulta engorroso de escribir y muchas veces es claro a partir del contexto cuándo el símbolo v se refiere a un vector y cuándo se refiere a un escalar.

[LOS EJEMPLOS]

[LOS EJERCICIOS]

Comentario

Hasta este momento, quien esté leyendo esto tiene todo el derecho del mundo a sentir que en realidad no hemos hecho nada para simplificar nuestro problema de la red neuronal. Si bien la ecuación quedó más sencilla de escribir, esto se hizo solamente a partir de definir una operación que a primera impresión podría parecer bastante compleja. Desde ese punto de vista, no hemos hecho mucho más que enmascarar el problema. El lado positivo de esto es que la máscara que hemos puesto (es decir, la multiplicación vector-matriz) es una operación muy firmemente entendida por la matemática, y lo que haremos a continuación para entenderla es estudiar sus propiedades. Antes de esto, hegamos una breve incursión en un área cercana de la matemática.

3.1.1. Un breve tour por el álgebra abstracta

Se le llama *álgebra abstracta*[2] a un grupo de teorías que encapsulan el estudio de los grupos, anillos, cuerpos, entre otros. No necesitamos entender lo que cada uno de esos términos significan, pero lo que todas esas áreas tienen en común es el estudio de *operaciones actuando sobre conjuntos*, conjuntos de elementos que no necesitan ser números reales, sino que pueden ser otros objetos matemáticos como matrices, funciones, entre otros. Teniendo un conjunto de objetos, una operación le entrega a este conjunto una cierta noción de *estructura*, en el sentido de que si combinamos dos elementos de este conjunto, obtenemos un tercero. Veamos en ese sentido cuál es la estructura de los espacios de vectores $\mathbb{R}^n$.

Hasta este momento, hemos tratado a los vectores solamente como paquetes de números, pero sería bueno en este punto recordar que sí tenemos definidas operaciones sobre ellos. Ya que así podríamos escribir la multiplicación vector-matriz en términos de las operaciones que ya conocemos. Las dos principales que tenemos son

1. **La suma:** Si $v, w \in \mathbb{R}^n$, entonces definimos $v + w$ como $(v + w)_i = v_i + w_i$. En otras palabras:

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_n \end{pmatrix}$$

2. **La ponderación:** (también conocida como multiplicación por escalares) Si $v \in \mathbb{R}^n$ y $\lambda \in \mathbb{R}$, entonces λv se define como $(\lambda v)_i = \lambda v_i$. En otras palabras:

$$\lambda \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \lambda v_1 \\ \lambda v_2 \\ \vdots \\ \lambda v_n \end{pmatrix}$$

Notablemente, lo que no tenemos es una noción de *multiplicación* entre dos vectores[3]. Sin embargo, utilizando solo estas dos operaciones, vemos que la multiplicación vector-matriz se puede escribir como sigue:

$$Wa = \begin{pmatrix} w_{11}a_1 + w_{12}a_2 + \dots w_{1n}a_n \\ w_{21}a_1 + w_{22}a_2 + \dots w_{2n}a_n \\ \vdots \\ w_{m1}a_1 + w_{m2}a_2 + \dots w_{mn}a_n \end{pmatrix} = a_1 \begin{pmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{m1} \end{pmatrix} + a_2 \begin{pmatrix} w_{12} \\ w_{22} \\ \vdots \\ w_{m2} \end{pmatrix} + \cdots + a_n \begin{pmatrix} w_{1n} \\ w_{2n} \\ \vdots \\ w_{mn} \end{pmatrix} = \sum_{i=1}^{n} W_{\bullet i} a_i$$

En que hemos definido $W_{\bullet i}$ como la i-ésima columna de W. Esta escritura del producto vector-matriz nos permite entenderlo como una suma entre diferentes ponderaciones de las columnas de la matriz, en que los ponderadores están dados por las componentes del vector a.

Ya que estamos aquí, hablemos un poco más sobre las operaciones de suma y producto escalar en $\mathbb{R}^n$. Es fácil verificar que estas operaciones satisfacen las siguientes propiedades:

[2] La referencia clásica que se suele recomendar para una primera lectura sobre álgebra abstracta es [DF03]

[3] Podría pensarse que eso es por flojera de que no nos hemos dado el trabajo de definirla, y podría pensarse que $(v * w)_i = v_i \cdot w_i$ es por ejemplo una buena noción de multiplicación entre vectores, y si bien nada nos impide de definir eso, hay ciertas propiedades que le pedimos a la multiplicación, y esta operación no las cumple. De hecho, el problema es más grande de lo que parece: se puede demostrar que **no existe en $\mathbb{R}^n$ una operación que cumpla esas propiedades** salvo que n sea una potencia de 2, un encuentro con este hecho sirve para transmitir lo profunda que puede ser el álgebra abstracta

- **Conmutatividad de +:** $\vec{u} + \vec{v} = \vec{v} + \vec{u}$
- **Asociatividad de +:** $(vecu + \vec{v}) + \vec{w} = vecu + (\vec{v} + \vec{w})$
- **Neutro de +:** Existe un vector $\vec{0}$ tal que para todo $v \in \mathbb{R}^n$ se cumple $\vec{0} + \vec{v} = \vec{v}$
- **Inverso para +:** Para todo $\vec{v} \in \mathbb{R}^n$, existe un correspondiente $-\vec{v} \in \mathbb{R}^n$ tal que $\vec{v} + (-\vec{v}) = \vec{0}$
- **Compatibilidad de ·:** $\lambda(\mu\vec{v}) = (\lambda\mu)\vec{v}$
- **Distributividad de escalares:** $\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w}$
- **Distributividad de vectores:** $(\lambda + \mu)\vec{v} = \lambda\vec{v} + \mu\vec{v}$

Razonablemente, resulta que el vector $\vec{0}$ es aquel vector con solo ceros en cada componente. Y el inverso $-\vec{v}$ viene dado por $(-\vec{v})_i = -v_i$.

Por mucho que lo pueda parecer, estas propiedades no están elegidas al azar, decimos que un **espacio vectorial** es un conjunto V en el que consideramos dos operaciones que denotamos + y · (a veces denotamos esto como $(V, +, \cdot)$) y en el que estas operaciones satisfacen la lista de propiedades anterior. Es decir, en demostrar que esas propiedades se cumplen para $\mathbb{R}^n$ con la suma y producto escalar definidas, estamos demostrando que $(\mathbb{R}^n, +, \cdot)$ posee la estructura de un espacio vectorial.

3.1.2. El teorema de representación matricial

Una buena actitud de vida en la matemática suele ser estudiar los objetos no tanto por lo que son, sino más bien a través de estudiar las *buenas* funciones que los conectan. Es decir, una vez que tenemos una estructura algebraica (como la de un espacio vectorial por ejemplo), la buena cosa que podemos hacer es preguntarnos cuáles son las funciones que conectan nuestro espacio vectorial con otro *preservando la estructura*. Ilustremos a qué nos referimos con preservar la estructura en este caso.

Imaginemos que en un espacio vectorial V tenemos dos elementos x, y, podemos sumar estos dos elementos para obtener uno nuevo que llamamos $z = x + y$. Esta situación está ilustrada en la figura [FIG]. Ahora, si además tenemos una función $f : V \to W$ en que W es otro espacio vectorial, una función que preserva la estructura sería una en que esta misma situación se repite en W. En otras palabras, que si $x + y = z$, entonces $f(x) + f(y) = f(z)$. Esto se puede resumir solamente en la ecuación $f(x + y) = f(x) + f(y)$, cosa que pedimos que se cumpla para cualquier par $x, y \in V$. Pero en un espacio vectorial no solamente tenemos la estructura de la suma, también tenemos un producto por escalares. Entonces el mismo tipo de razonamiento nos diría que nuestra *buena función* debería también satisfacer que para cualquier $x \in V, \lambda \in \mathbb{R}$, debería cumplirse que $f(\lambda x) = \lambda f(x)$. Estas dos condiciones juntas significan que la función es una *buena función*. En álgebra abstracta se utiliza el término *homomorfismo* para describir a este tipo de funciones, pero en el contexto específico del álgebra lineal, les llamamos *funciones o transformaciones lineales*.

Tenemos entonces una familia de espacios vectoriales que son los $\mathbb{R}^n$, y tenemos una noción de lo que significa ser una función lineal entre ellos. También tenemos estos objetos que son las matrices de $m \times n$ y vimos cómo una matriz $A \in \mathbb{R}^{m \times n}$ induce una función $L_A : \mathbb{R}^n \to \mathbb{R}^m$, en particular la función $L_A(x) = Ax$. Queda por supuesto preguntarnos si esta función es lineal.

Bibliografía

[DF03] D.S. Dummit y R.M. Foote. *Abstract Algebra*. Wiley, 2003. ISBN: 9780471433347. URL: https://books.google.cl/books?id=KJDBQgAACAAJ.

[GOM23] Alejandro González, Harold Ojeda e Iván Morales. *Apunte de matemáticas 1*. Mar. de 2023.

[HK04] Kenneth Hoffman y Ray A. Kunze. *Linear Algebra*. Second. PHI Learning, 2004. ISBN: 8120302702. URL: http://www.worldcat.org/isbn/8120302702.

[Le+11] Quoc V. Le et al. «On Optimization Methods for Deep Learning». En: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, págs. 265-272. ISBN: 9781450306195.

[SM88] M. Spivak y B.F. Marqués. *Cálculo Infinitesimal*. Reverté, 1988. ISBN: 9788429151367. URL: https://books.google.cl/books?id=mjdXY8rshREC.