



Repaso Matemático Minería de Datos y Machine Learning

Felipe Bravo

Álgebra Lineal

En ML usamos las siguientes representaciones algebraicas para nuestros objetos y sus atributos: **escalares**, **vectores** y **matrices**.

- **Escalares**: un escalar es simplemente un número como 7.56. El valor de un atributo numérico para un objeto se representa por un escalar.
- **Vectores**: un vector es un arreglo ordenado de escalares $\mathbf{x} = [x_1, x_2, \dots, x_n]$ donde x_i es el i -ésimo elemento de \mathbf{x} .
 - a. En ML un objeto de n **atributos** numéricos puede ser representado por un vector de n dimensiones.
 - b. El **producto punto** entre dos vectores $\mathbf{a} \cdot \mathbf{b}$ es la suma de la multiplicación de todos sus elementos:

$$(a_1, a_2, a_3, \dots, a_n) \cdot (b_1, b_2, b_3, \dots, b_n) = a_1 b_1 + a_2 b_2 + \dots + a_n b_n = \sum a_i \cdot b_i$$

Álgebra Lineal

Norma y distancia euclidiana:

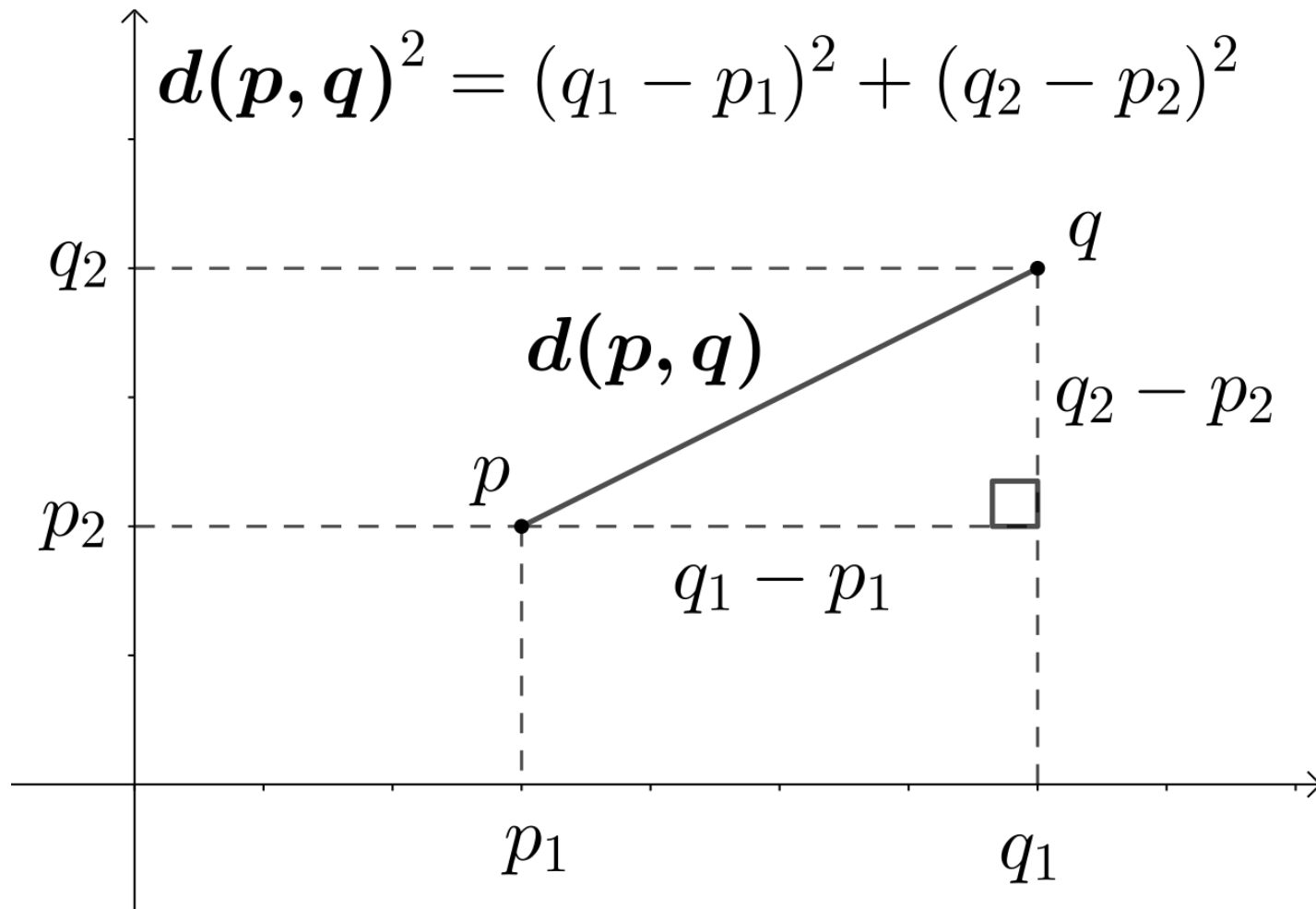
- La norma euclidiana de un vector $\|\mathbf{v}\|_2$ es el largo del vector en un espacio euclidiano (piensen en pitágoras).

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{\sum_{i=1}^n v_i^2}$$

- Luego, la distancia euclidiana nos permite calcular qué tan lejos están dos vectores \mathbf{x} e \mathbf{y} .

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2},$$

Álgebra Lineal



Distancia Euclidiana en \mathbb{R}^2 . Fuente: Wikipedia

Álgebra Lineal

Matrices

- Una matriz es un arreglo de dos dimensiones, entonces cada elemento se identifica por dos índices en vez de uno.
- Ejemplo: sea A una matriz de dos filas y dos columnas ($A \in \mathbb{R}^{2 \times 2}$)


$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- Un dataset de n objetos y m atributos numéricos se puede representar por una matriz de $n \times m$.
- El i -ésimo ejemplo de un dataset se representa como el vector de la i -ésima fila de su matriz correspondiente.
- Un vector puede ser visto como una matriz de una sola columna.

Álgebra Lineal

Matrices

- La transpuesta de una matriz A^T , corresponde a una copia de la matriz donde se intercambian las filas por las columnas.


$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

- Podemos sumar dos matrices, siempre y cuando éstas tengan las mismas dimensiones, sumando sus elementos correspondientes: $C = A + B$ donde $C_{i,j} = A_{i,j} + B_{i,j}$. Ejemplo:

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 \\ 1+7 & 0+5 \\ 1+2 & 2+1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 8 & 5 \\ 3 & 3 \end{bmatrix}$$

Álgebra Lineal

Matrices

- También podemos sumar un escalar a una matriz o multiplicar una matriz por un escalar, simplemente realizando esa operación en cada elemento de una matriz: $D = a*B + c$ donde $D_{i,j} = a*B_{i,j} + c$
- Podemos multiplicar dos matrices A y B ($C=A*B$) siempre y cuando el número de columnas de A sea igual al número de filas de B. El valor de C_{ij} es igual al producto punto de la i-ésima fila de A por la j-ésima columna de B ($A_{i,:} \cdot B_{:,j}$).

Fuente:

<https://www.javabrahman.com/wp-content/uploads/MatrixMultiplicationExample.png>

$$\begin{array}{ccc} & M & \\ \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} & \times & \begin{array}{ccc} & N & \\ \begin{bmatrix} 3 & 6 & 9 \\ 4 & 7 & 10 \\ 5 & 8 & 11 \end{bmatrix} & = & \begin{array}{ccc} & M \times N & \\ \begin{bmatrix} 26 & 44 & 62 \\ 62 & 107 & 152 \end{bmatrix} & \end{array} \end{array}$$

[[1X3)+(2X4)+(3X5)] [(1X6)+(2X7)+(3X8)] [(1X9)+(2X10)+(3X11)]
[[4X3)+(5X4)+(6X5)] [(4X6)+(5X7)+(6X8)] [(4X9)+(5X10)+(6X11)]

Matrix multiplication example

Copyright © JavaBrahman.com, all rights reserved.



Álgebra Lineal

Matrices

- Una matriz cuadrada es una matriz que tiene el mismo número de filas y columnas.
- Una matriz cuadrada muy particular es la matriz identidad I que tiene 1s en la diagonal y ceros en todas las otras celdas.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

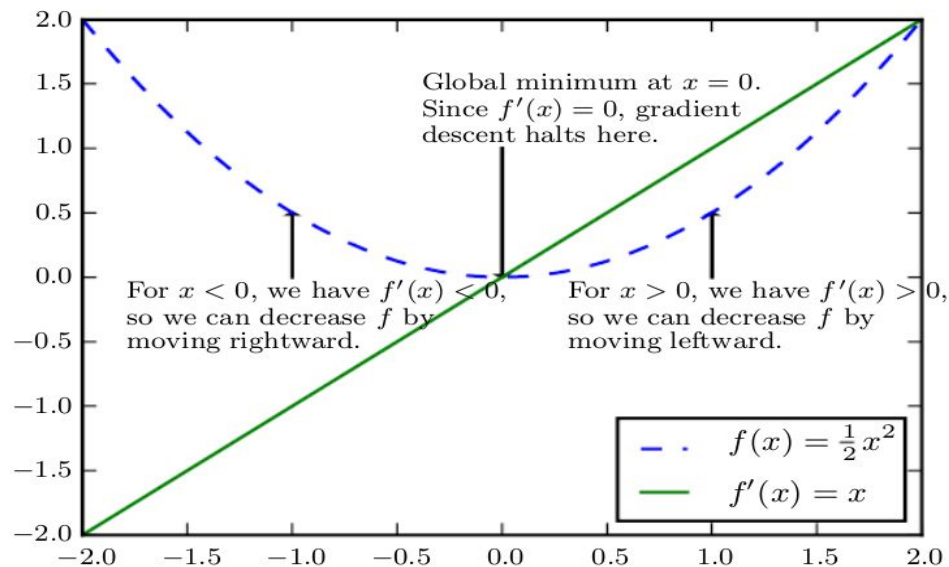
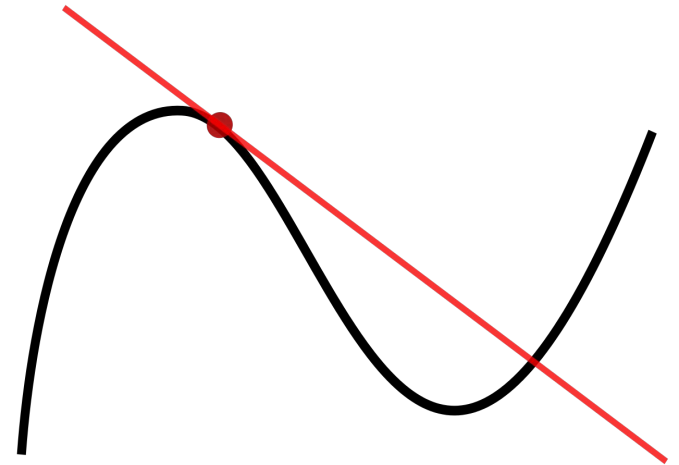
- La inversa de una matriz A se denota como A^{-1} y cumple con la propiedad que $A^* A^{-1} = I$

Optimización

La **optimización** es una metodología para encontrar el valor **máximo** o **mínimo** de una función matemática.

Muchos métodos de ML se plantean como problemas de optimización.

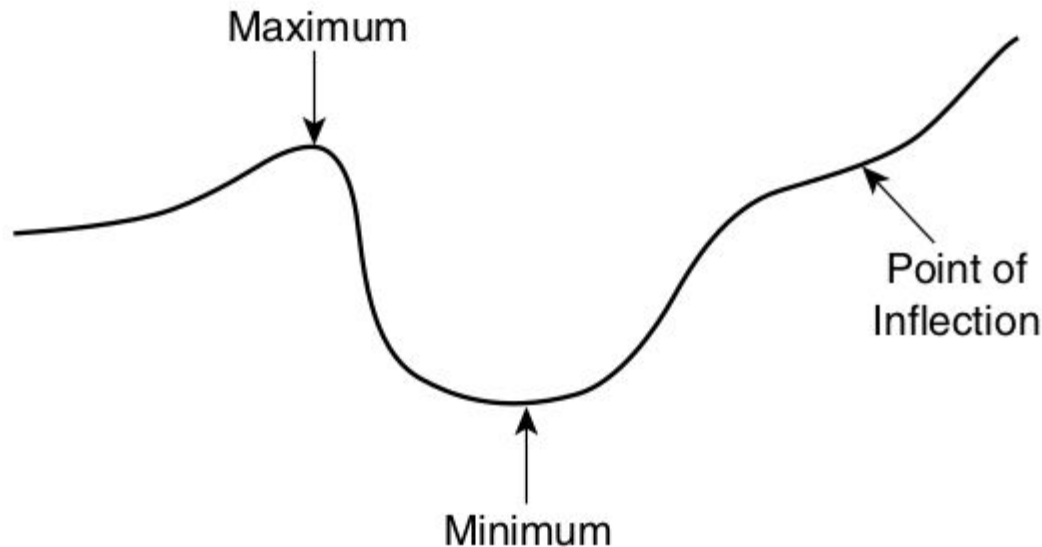
- Supongamos que tenemos una función $y = f(x)$, donde tanto x como y son números reales.
- La derivada de esta función se escribe así $f'(x)$ o así: $\frac{dy}{dx}$
- La derivada $f'(x)$ nos entrega el valor de la pendiente de $f(x)$ en el punto x .



Optimización

La derivada de una función es muy útil para encontrar valores mínimos o máximos.

- Los puntos en que la derivada de una función vale cero ($f'(x)=0$) se conocen como puntos críticos: **máximo**, **mínimo** o punto de **inflexión** (o punto silla).

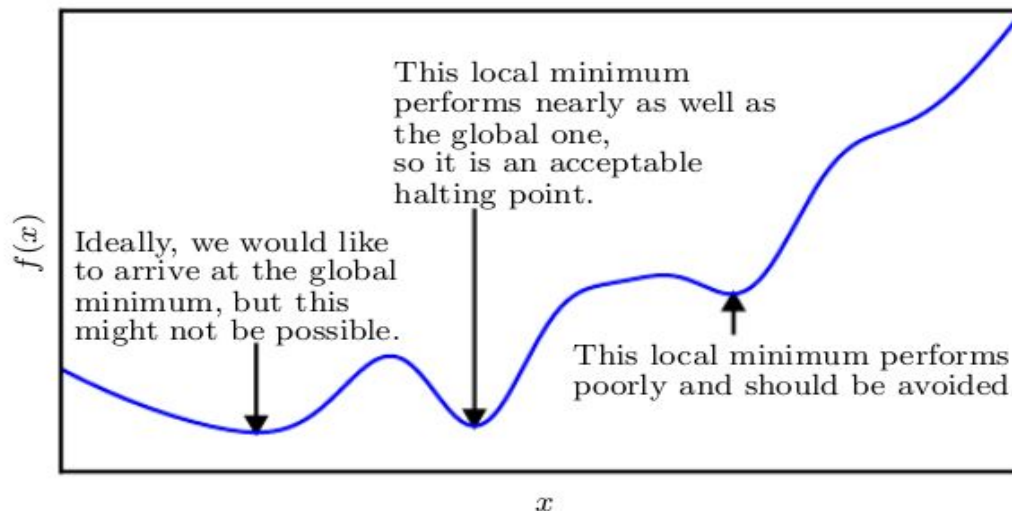


Optimización

- Para distinguir entre estos tres tipos de puntos críticos es necesario analizar la **segunda derivada** de la función $f''(x)$ (la derivada de la derivada) que nos da información sobre la curvatura de la función.

$$\frac{d^2 f}{dx^2}$$

- Otra gran dificultad al optimizar funciones es que a veces los puntos críticos pueden corresponder a mínimos o máximos **locales**.



Optimización

- ¿Cómo optimizamos funciones con múltiples inputs?

$$f(x_1, x_2, x_3) = 2 * x_1 + x_2^2 - 5 * x_3$$

- La **derivada parcial** $\frac{\partial}{\partial x_i} f(\mathbf{x})$ mide cómo cambia f sólo cuando hacemos un cambio en x_i .

$$\frac{\partial f}{\partial x_1} = 2, \frac{\partial f}{\partial x_2} = 2 * x_2, \frac{\partial f}{\partial x_3} = -5$$

- El **gradiente** ∇_f es un vector con todas las derivadas parciales de una función.

$$\nabla_f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right]$$

- Para encontrar puntos críticos en funciones de varios inputs tenemos que encontrar los valores de x donde el gradiente vale **cero**.
- Para distinguir entre máximos, mínimos y puntos silla tenemos que recurrir al **Hessiano**, que es una matriz con todas las segundas derivadas.

Optimización

- ¿Qué pasa cuando le agregamos restricciones al problema?
- Existen dos tipos de restricciones: 1) restricciones de igualdad y 2) restricciones de desigualdad.

Restricciones de igualdad

- Supongamos que queremos encontrar el mínimo de $f(x_1, x_2, \dots, x_d)$ sujeto a las siguientes p restricciones de igualdad:

$$g_i(\mathbf{x}) = 0, \quad i = 1, 2, \dots, p.$$

- Cada restricción de igualdad es una función del tipo $g(x) = 2x_1 + 3x_2 - 3 = 0$
- Esto se puede resolver usando un método llamado **multiplicadores de Lagrange**.

Optimización

Multiplicadores de Lagrange

1. Definimos el Lagrangiano $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i g_i(\mathbf{x})$
donde λ_i es una variable adicional llamada **multiplicador de Lagrange**.
2. Derivamos el Lagrangiano respecto a \mathbf{x} y λ , igualamos a cero y despejamos el sistema de ecuaciones.

$$\frac{\partial L}{\partial x_i} = 0, \quad \forall i = 1, 2, \dots, d$$

$$\frac{\partial L}{\partial \lambda_i} = 0, \quad \forall i = 1, 2, \dots, p.$$

La solución óptima del Lagrangiano corresponde al óptimo de $f(\mathbf{x})$ que satisface las restricciones de igualdad.

Optimización

Restricciones de desigualdad:

- ¿Qué hacemos cuando tenemos restricciones de desigualdad del tipo $h_i(\mathbf{x}) \leq 0$?
Por ejemplo $x_1 + x_2 \leq 0$.
- Se formula entonces un problema optimización con restricciones como minimizar $f(x_1, x_2, \dots, x_n)$ sujeto a las siguientes q restricciones de desigualdad:

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, q.$$

- El método para resolver este problema es bastante similar al método de Lagrange descrito anteriormente.
- Sin embargo, las restricciones de la desigualdad plantean condiciones adicionales al problema de optimización.

Optimización

- El problema de optimización con restricciones de desigualdad se formula con el siguiente Lagrangiano:

$$L = f(\mathbf{x}) + \sum_{i=1}^q \lambda_i h_i(\mathbf{x})$$

- Una solución óptima de este problema debe satisfacer las condiciones Karush-Kuhn-Tucker (KKT):

$$\begin{aligned}\frac{\partial L}{\partial x_i} &= 0, \quad \forall i = 1, 2, \dots, d \\ h_i(\mathbf{x}) &\leq 0, \quad \forall i = 1, 2, \dots, q \\ \lambda_i &\geq 0, \quad \forall i = 1, 2, \dots, q \\ \lambda_i h_i(\mathbf{x}) &= 0, \quad \forall i = 1, 2, \dots, q.\end{aligned}$$

Optimización

- Notemos ahora que los multiplicadores de Lagrange no pueden ser negativos.
- Las restricciones KKT nos obligan verificar que un óptimo satisfaga todas las condiciones.
- Eso puede ser muy difícil de lograr analíticamente, especialmente si se tienen muchas restricciones.
- Por lo general recurrimos a métodos numéricos como la programación lineal y la programación cuadrática.



dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

f @ in  / DCCUCHILE