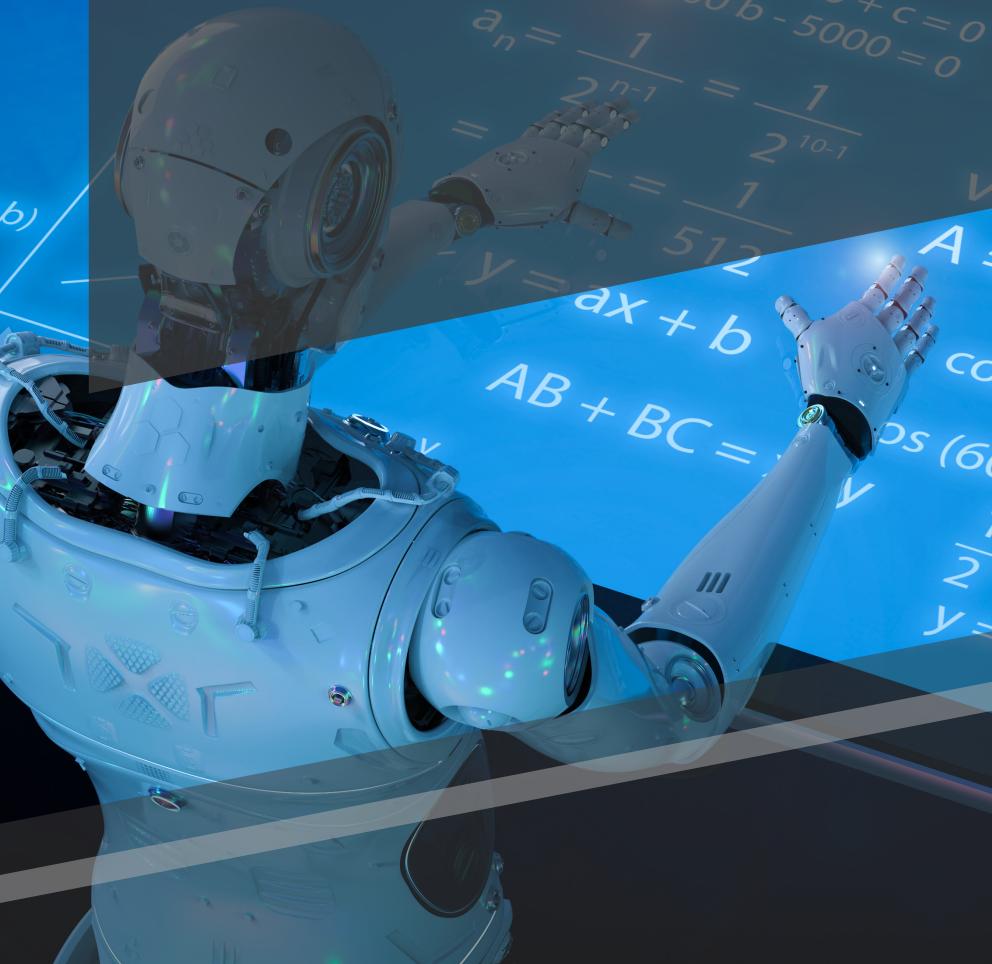


MATEMÁTICAS PARA LA INTELIGENCIA ARTIFICIAL (IA)



Dr. Gherardo Varando

MÁSTER EN INTELIGENCIA ARTIFICIAL

Módulo II. Fundamentos matemáticos

viu

**Universidad
Internacional
de Valencia**

Este material es de uso exclusivo para los alumnos de la Universidad Internacional de Valencia. No está permitida la reproducción total o parcial de su contenido ni su tratamiento por cualquier método por aquellas personas que no acrediten su relación con la Universidad Internacional de Valencia, sin autorización expresa de la misma.

Edita

Universidad Internacional de Valencia

Máster en
Inteligencia Artificial

Matemáticas para la inteligencia artificial (IA)

Módulo II. Fundamentos matemáticos

6 ECTS

Dr. Gherardo Varando

Leyendas



Enlace de interés



Ejemplo



Importante



abc Los términos resaltados a lo largo del contenido en color **naranja** se recogen en el apartado **GLOSARIO**.

Índice

CAPÍTULO 1. ÁLGEBRA LINEAL	7
1.1. Vectores y matrices	8
1.1.1. Vectores	8
1.1.2. Producto escalar	9
1.1.3. Matrices	10
1.1.4. Producto de matrices.....	11
1.1.5. Matrices particulares	12
1.2. Determinante, matriz inversa y autovalores	14
1.2.1. Determinante de una matriz	14
1.2.2. Menores y rango	17
1.2.3. Matriz inversa	18
1.2.4. Autovalores y autovectores	19
1.3. Algunas factorizaciones de matrices	20
1.3.1. Factorización LU.....	20
1.3.2. Factorización de Cholesky	20
1.3.3. Factorización QR.....	21
1.3.4. Diagonalización de una matriz	21
1.3.5. Descomposición en valores singulares	21
1.4. Sistemas de ecuaciones lineales	22
1.4.1. Forma matricial	23
1.4.2. Sistemas triangulares y método de Gauss.....	24
CAPÍTULO 2. ANÁLISIS	25
2.1. Funciones reales y límites.....	25
2.1.1. Funciones.....	25
2.1.2. Gráficas de funciones	27
2.1.3. Límites de funciones reales y continuidad.....	28
2.2. Derivadas de funciones	28
2.2.1. La derivada como límite	28
2.2.2. Calcular la derivada de una función.....	30
2.2.3. Funciones monótonas	31
2.3. Máximos y mínimos de funciones reales	32
2.3.1. Máximos y mínimos absolutos.....	33
2.3.2. Máximos y mínimos relativos y puntos críticos.....	34
2.3.3. Derivadas de órdenes superiores	35

2.4. Funciones de diversas variables	37
2.4.1. Derivadas parciales.....	38
2.4.2. Gradiente y derivadas direccionales	39
2.4.3. Segundas derivadas parciales y matriz hessiana	40
2.4.4. Máximos, mínimos y puntos críticos	40
2.5. Integrales	41
2.5.1. Integrales de funciones	42
2.5.2. Primitivas de funciones y teorema fundamental del cálculo	43
2.5.3. Propiedades de la integral y algunas primitivas de funciones básicas.....	44
 CAPÍTULO 3. PROBABILIDAD Y ESTADÍSTICA	 46
3.1. Probabilidad básica.....	47
3.1.1. Algunos ejemplos	47
3.1.2. Espacio muestral y probabilidad	47
3.1.3. Eventos independientes	49
3.1.4. Probabilidad condicionada y teorema de Bayes.....	50
3.1.5. Espacios muestrales infinitos y continuos.....	52
3.2. Variables aleatorias	52
3.2.1. Variables aleatorias discretas	53
3.2.2. Variables aleatorias continuas	53
3.2.3. Vectores aleatorios	54
3.2.4. Valor esperado y varianza	55
3.2.5. Covarianza y matriz de covarianza	58
3.3. Algunas distribuciones	59
3.3.1. Distribuciones discretas	59
3.3.2. Distribuciones continuas.....	60
3.4. Estimación de parámetros	63
3.4.1. Límites de variables aleatorias	65
3.4.2. Método de los momentos	65
3.4.3. Método de máxima verosimilitud	66
 GLOSARIO	 69
 ENLACES DE INTERÉS	 78
 BIBLIOGRAFÍA	 79



Capítulo 1

Álgebra lineal

En este capítulo repasaremos los conceptos fundamentales del álgebra lineal: vectores, matrices, sus propiedades y las posibles operaciones entre ellos. Además, veremos cómo calcular **autovalores** y **autovectores** de matrices, así como resolver sistemas de ecuaciones lineales.



El álgebra lineal forma parte de los fundamentos matemáticos de la inteligencia artificial y del aprendizaje automático en particular. En el nivel más básico, podemos simplemente observar como los conjuntos de datos se proporcionan (o transforman) muchas veces en matrices numéricas: las tablas de valores numéricos o las imágenes son los ejemplos clásicos.

Además, muchos algoritmos estadísticos y de aprendizaje automático se resuelven con fórmulas y técnicas propias del álgebra lineal.

El ejemplo más llamativo es la técnica de análisis de componentes principales (*principal component analysis*, PCA) (Pearson, 1901), utilizada para reducir la dimensión de algunos problemas de predicción. El PCA se calcula a través de la descomposición en autovalores de la matriz XX^t , donde X es la matriz de datos.

1.1. Vectores y matrices

1.1.1. Vectores

Un **vector** v de dimensión n es una lista ordenada de números reales $v = (v_1, v_2, \dots, v_n) \in R^n$. Denotamos los vectores con letras minúsculas, y con la misma letra y el correspondiente subíndice denotamos el valor del vector en una posición específica. El siguiente ejemplo debería aclarar cualquier duda.



Ejemplo

Si $w = (10; 23; 56; 4.5; 7.8)$ es un vector de dimensión 5, entonces el primer componente del vector w que denotamos con el símbolo w_1 es igual a 10. De forma similar, el tercer componente es $w_3 = 56$.

Los vectores de dimensión fijada n se pueden interpretar geométricamente como las coordenadas de puntos en el espacio n -dimensional. Por esta razón utilizaremos las palabras *vectores* o *puntos* para indicar los mismos objetos.

Podemos definir algunas operaciones entre vectores:

- Suma $v = u + w$ para $u, w \in R^n$, cuyo resultado es $v = (v_1, \dots, v_n)$, donde $v_i = u_i + w_i$, es decir, los dos vectores se suman componente a componente.
- Producto por un escalar, $v = aw$ para $a \in R, w \in R^n$, cuyo vector resultante se obtiene multiplicando cada componente w_i por el número a .

Estas dos operaciones definen lo que se llama una estructura de **espacio vectorial** para los puntos en R^n .

Dado un vector $v \in R^n$, definimos la norma euclídea o norma 2:

$$\|v\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$$

Y, de forma similar, la norma 1:

$$\|v\|_1 = \sum_{i=1}^n |v_i|$$

Cada norma define una distancia a través de la fórmula $d(v, w) = \|v - w\|$.

La norma euclídea corresponde a la distancia usual en el plano o en el espacio, mientras que la norma 1 define la que se llama distancia del taxista o distancia Manhattan. En la Figura 1 se puede apreciar la diferencia entre la distancia euclídea y la distancia Manhattan en el plano bidimensional.

Se pueden definir muchas otras normas (y respectivas distancias). Una familia muy utilizada son las normas p , definidas como sigue para cada $p \geq 1$:

$$\|v\|_p = \left(\sum_{i=1}^n v_i^p \right)^{1/p}$$

Además, se puede también definir la norma p para $p = +\infty$:

$$\|v\|_\infty = \max_{i \leq i \leq n} |v_i|$$

En la Figura 1, las líneas roja, azul y amarilla tienen la misma longitud (12) en la distancia euclídea y en la distancia Manhattan, y son todos caminos de mínima longitud entre los dos puntos. Mientras que en la geometría euclídea el único camino de mínima longitud es el verde, en la geometría Manhattan el camino verde también tiene longitud 12, así que no es más corto que los demás.

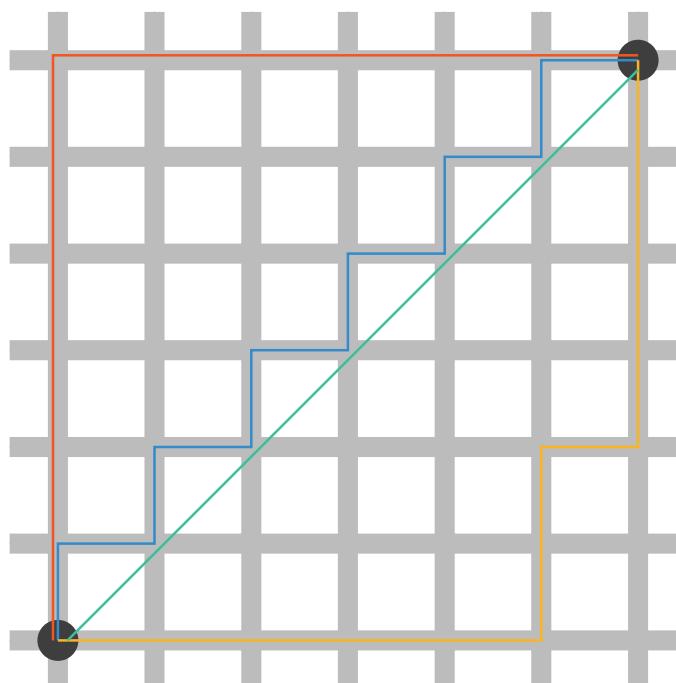


Figura 1. Diferencia entre la distancia Manhattan y la distancia euclídea. Dominio público. Recuperado de https://commons.wikimedia.org/wiki/File:Manhattan_distance.svg

1.1.2. Producto escalar

Es posible también definir una operación que a partir de dos vectores devuelva un número real, el **producto escalar**.

Para cada dos vectores de las mismas dimensiones $u, v \in R^n$, definimos el producto escalar de u y v de la siguiente forma:

$$u \cdot v = \sum_{i=1}^n u_i v_i$$

En particular, el producto escalar de un vector por sí mismo devuelve el cuadrado de la norma 2, como es fácil de comprobar:

$$v \cdot v = \sum_{i=1}^n v_i v_i = \sum_{i=1}^n v_i^2 = \|v\|_2^2$$



Se dice que dos vectores $u, v \in R^n$ cuyo producto escalar es 0 son **ortogonales** (esta definición generaliza a una dimensión cualquiera el concepto de líneas ortogonales en dos o tres dimensiones).

En general, el producto escalar de dos vectores permite calcular la proyección de un vector en otro. Si $u, v \in R^n$ son dos vectores de la misma dimensión, entonces podemos definir la proyección de u en el vector v como se muestra a continuación:

$$\text{Proj}_v(u) = \frac{(u \cdot v)v}{\|v\|_2^2}$$

La proyección tiene las siguientes propiedades:

- $\text{Proj}_v(\lambda v) = \lambda \text{Proj}_v(v)$ para cada $\lambda \in R$.
- $\text{Proj}_v(\text{Proj}_v(u)) = \text{Proj}_v(u)$
- $\text{Proj}_v(u) \cdot v = u \cdot v$

Así pues, gracias a la última propiedad, podemos definir el complemento ortogonal de la siguiente forma:

$$\text{Ort}_v(u) = u - \text{Proj}_v(u)$$

Este vector es el vector ortogonal que completa la proyección:

$$u = \text{Proj}_v(u) + \text{Ort}_v(u)$$

1.1.3. Matrices

Una **matriz** A de dimensiones $n \times m$ es una tabla de números reales:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \dots & \dots & \dots & \dots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix}$$

Cada elemento de una matriz se identifica con dos coordenadas (i, j) . Aquí utilizaremos letras mayúsculas para identificar las matrices y las correspondientes letras minúsculas con doble subíndice para denotar los elementos de las matrices.



Ejemplo

Consideremos la siguiente matriz de dimensión 3×2 :

$$A = \begin{pmatrix} 2 & 4 & 9 \\ 1 & -2 & -5 \\ 9,5 & -1 & 3 \end{pmatrix}$$

El elemento $(1,2)$ de la matriz A, es decir, el elemento de la primera fila y la segunda columna, es $a_{1,2} = 4$.

De forma similar a los vectores, es posible definir la suma de matrices (de las mismas dimensiones) y el producto de una matriz con un número real (componente a componente).

El conjunto de matrices de dimensiones $n \times m$ y elementos reales se denota como $R^{n \times m}$, donde las dimensiones de las matrices se pueden leer en el exponente.



Ejemplo

Sean A y B las matrices:

$$A = \begin{pmatrix} -1 & 3 \\ 3 & 1 \\ 0 & 5 \end{pmatrix} B = \begin{pmatrix} 1 & 0 \\ -2 & 6 \\ 1 & -1 \end{pmatrix}$$

Así pues, la matriz suma $A + B$ se obtiene sumando cada componente de A con el correspondiente componente de B :

$$A + B = \begin{pmatrix} -1 & 3 \\ 3 & 1 \\ 0 & 5 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ -2 & 6 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 3 \\ 1 & 7 \\ 1 & 4 \end{pmatrix}$$

1.1.4. Producto de matrices

Definimos el producto de dos matrices compatibles $A \in R^{m \times n}$ y $B \in R^{n \times p}$ como la matriz $AB = D \in R^{m \times p}$ con componentes obtenidos de la siguiente forma:

$$d_{i,j} = \sum_{k=1}^n a_{i,k} b_{k,j} = a_i \cdot b_j$$

Denotamos con a_i el vector de la i -ésima fila de A y, de forma similar, con b_j el vector de la j -ésima columna de B . Este tipo de producto de matrices se llama también producto fila-columna.



Solo las parejas de matrices con dimensiones compatibles se pueden multiplicar entre sí. En particular, el producto AB se puede definir si el número de columnas de la matriz A es igual al número de filas de la matriz B .

Una forma sencilla de comprobar si dos matrices son compatibles para la multiplicación (y también de calcular la dimensión de la matriz resultante) es escribir las dos dimensiones de las matrices, $(m \times n)(n \times p)$, y comprobar que los números centrales son iguales (en este caso, n). La matriz resultante tendrá dimensiones iguales a los números externos (en este caso, $m \times p$).

El producto de matrices satisface la propiedad distributiva con respecto a la suma:

$$A(B + D) = AB + AD$$

Pero no satisface la propiedad conmutativa y, en general, $AB \neq BA$, también en los casos donde es posible multiplicar las matrices de las dos formas.



Ejemplo

Sean A y B las siguientes matrices:

$$A = \begin{pmatrix} 3 & 2 & 4 \\ 8 & 1 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 10 & 10 \\ -2 & 3 \\ 4 & -5 \end{pmatrix}$$

Las dimensiones de las matrices son 2×3 y 3×2 , de modo que las dos matrices son compatibles para el producto fila-columna. El resultado es una matriz 2×2 :

$$AB = \begin{pmatrix} 30 - 4 + 16 & 30 + 6 - 20 \\ 80 - 2 + 4 & 80 + 3 - 5 \end{pmatrix} = \begin{pmatrix} 42 & 16 \\ 82 & 78 \end{pmatrix}$$

Los vectores se pueden ver como casos particulares de matrices donde una de las dos dimensiones es 1. Es decir, podemos llamar **vector columna** a una matriz de dimensiones $n \times 1$:

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in R^{n \times 1}$$

Del mismo modo, podemos llamar **vector fila** a una matriz de dimensiones $1 \times n$:

$$w = (w_1, w_2, \dots, w_n) \in R^{1 \times n}$$

De ahora en adelante, si no se especifica, todos los vectores se consideran vectores columna.

Así pues, el producto escalar se puede ver como un caso particular de producto de matrices. En particular, es el producto de un vector fila y un vector columna, y para dos vectores columnas $u, v \in R^n$ podemos escribir $u \cdot v = u^t v$.

De forma similar, podemos ver las matrices como aplicaciones entre espacios vectoriales. En particular, para cada matriz $A \in R^{n \times m}$ podemos definir una aplicación lineal a partir del espacio vectorial R^m en R^n , considerando el producto entre la matriz A y los vectores columna $v \in R^{m \times 1}$:

$$v \in R^m \rightarrow Av \in R^n$$

1.1.5. Matrices particulares

Una matriz con dimensiones iguales, es decir $A \in R^{n \times n}$, se llama **matriz cuadrada**. El vector de elementos $(a_{1,1}, a_{2,2}, a_{3,3}, \dots, a_{n,n})$ se llama **diagonal de la matriz** A . La suma de los valores en la diagonal de la matriz se llama **traza** de la matriz y se indica con el símbolo $\text{tr}(A)$. Si $A \in R^{n \times n}$ es una matriz cuadrada, entonces la **matriz traspuesta** se define como la matriz con las mismas dimensiones que A y elementos iguales, pero permutando las filas y las columnas entre ellas.

Así pues, el elemento (i, j) de la matriz A traspuesta es igual al elemento (j, i) de la matriz A . La matriz traspuesta de A se denota como A^t . Si una matriz $A \in R^{n \times n}$ es tal que $A = A^t$, entonces la matriz A es una **matriz simétrica**.

Veamos ahora las principales propiedades de la traspuesta si $A \in R^{n \times n}$ es una matriz cuadrada:

- $(A)^t = A$
- $(sA)^t = sA^t$ para cada número real s

Para cada dos matrices cuadradas $A, B \in R^{n \times n}$:

- $(A + B)^t = A^t + B^t$
- $(AB)^t = B^t A^t$

Además, para cada matriz $A \in R^{n \times n}$, las matrices AA^t , $A^t A$ y $(A + A^t)$ son todas simétricas, como es fácil de comprobar.



La siguiente matriz es una matriz simétrica de dimensión 3:

$$\begin{pmatrix} 1 & 4 & -1 \\ 4 & -3 & 7 \\ -1 & 7 & 9 \end{pmatrix}$$

La traza de A es igual a $\text{tr}(A) = 1 - 3 + 9 = 7$.

Para cada matriz cuadrada $A \in R^{n \times n}$, es posible definir una forma bilineal sobre las parejas de vectores $(v, w) \in R^n \times R^n$:

$$v^t Aw : R^n \times R^n \rightarrow R$$

Esta aplicación asocia a cada pareja de vectores de dimensión n un número real, obtenido con la operación $v^t Aw$. Se puede observar que las dimensiones de las matrices en la operación son $(1 \times n)(n \times n)(n \times 1)$, así que el resultado es una matriz de dimensiones 1×1 , es decir, un número real.

Si la matriz es simétrica, entonces la forma bilineal asociada es simétrica en los dos argumentos.

Una matriz es **definida positiva** si $x^t Ax > 0$ para cada vector x distinto del vector 0, y es **definida negativa** si $x^t Ax < 0$. Además, la matriz A es semidefinida positiva si vale $x^t Ax \geq 0$, y es semidefinida negativa si vale $x^t Ax \leq 0$.

Una matriz cuadrada particular es la **matriz identidad** de dimensión n , que denotamos con $I_n \cdot I_n$ es la matriz cuadrada con elementos diagonales 1, y 0 en los demás elementos.

$$I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Claramente, para cada n , I_n es una matriz simétrica definida positiva, y la forma bilineal asociada es exactamente el producto escalar estándar.

Si consideramos la aplicación lineal definida por la matriz $A \in R^{n \times n}$, entonces la forma bilineal asociada a la matriz $A^t A$ indica cómo se modifica el producto escalar, dado que $(Av) \cdot (Aw) = (Av)^t (Aw) = v^t A^t Aw$. Además, está claro que la matriz $A^t A$ es semidefinida positiva, dado que $x^t A^t Ax = (Ax)^t (Ax) = \|Ax\|_2^2 \geq 0$.

Una **matriz ortogonal** es una matriz cuadrada $A \in R^n$ tal que $AA^t = A^t A = I_n$, es decir, el producto de A y su traspuesta es igual a la matriz identidad. Si $A \in R^{n \times n}$ es una matriz ortogonal, entonces sus columnas (y sus filas) son vectores ortogonales entre sí y además su norma 2 es igual a 1. Dado que $A^t A = I_n$, está claro que la aplicación lineal asociada a una matriz ortogonal no modifica el producto escalar entre vectores, por lo que se trata de una transformación rígida que no modifica las distancias entre los puntos.

Se dice que una matriz cuadrada es **triangular inferior** si todos los elementos que hay encima de la diagonal son iguales a 0. De forma similar, una matriz es **triangular superior** si todos los elementos por debajo de la diagonal son iguales a 0.

1.2. Determinante, matriz inversa y autovalores

1.2.1. Determinante de una matriz

El **determinante** de una matriz cuadrada es un número real que puede ser visto como el factor de escala de la trasformación lineal asociada a la matriz. Como veremos más adelante, es una herramienta fundamental para calcular la inversa de una matriz y para resolver sistemas de ecuaciones lineales.

Empezamos a definir el determinante para matrices cuadradas 2×2 :

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

El determinante de A , indicado como $\det(A)$, es el número:

$$\det(A) = ad - bc$$

Es decir, el determinante para matrices 2×2 es la diferencia entre el producto de los elementos en la diagonal y los elementos en la antidiagonal.

Para las matrices 3×3 hay una fórmula parecida. En particular, sea ahora A la siguiente matriz:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix}$$

El determinante de A se puede calcular como la suma de los productos de los elementos en todas las diagonales menos la suma de los productos de los elementos en las antidiagonales. Esta regla para memorizar el cálculo del determinante para matrices 3×3 se llama regla de Sarrus (Cohn, 1994) (véase Figura 2).

$$\det(A) = (a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{2,1}a_{3,2}a_{1,3}) - (a_{1,3}a_{2,2}a_{3,1} + a_{1,2}a_{2,1}a_{3,3} + a_{1,1}a_{3,2}a_{2,3})$$

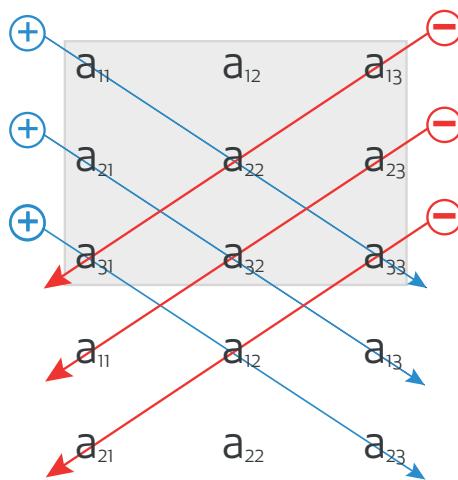


Figura 2. Representación gráfica de la regla de Sarrus para el cálculo de determinantes de matrices 3×3 .



Ejemplo

Consideremos la siguiente matriz:

$$A = \begin{pmatrix} 2 & 0 & 4 \\ -2 & 3 & 1 \\ -1 & -1 & 3 \end{pmatrix}$$

Utilizando la regla de Sarrus podemos calcular el determinante:

$$\det(A) = (2 \cdot 3 \cdot 3 + 0 \cdot 1 \cdot (-1) + (-2) \cdot (-1) \cdot 4) - (4 \cdot 3 \cdot (-1) + 0 \cdot (-2) \cdot 3 + 2 \cdot (-1) \cdot 1) = 26 + 14 = 40$$

Para las matrices cuadradas de dimensiones cualesquiera, se puede utilizar la regla de Laplace, una fórmula recursiva que reduce el cómputo del determinante de una matriz al cómputo de determinantes de submatrices de dimensiones inferiores.

En particular, para una matriz $A \in R^{n \times n}$ definimos el **menor complementario** del elemento a_{ij} como el determinante de la submatriz obtenida a partir de A eliminando la fila i y la columna j . Denotamos el menor complementario del elemento con el símbolo α_{ij} .



Ejemplo

El menor complementario del elemento en posición (1, 2) de la matriz $\begin{pmatrix} 1 & -1 & 2 \\ 2 & -1 & 6 \\ 1 & 5 & 0 \end{pmatrix}$ es el siguiente determinante:

$$\alpha_{1,2} = \det \begin{pmatrix} 2 & 6 \\ 1 & 0 \end{pmatrix} = 2 \cdot 0 - 6 \cdot 1 = -6$$

Definimos ahora el adjunto (o cofactor) de un elemento a_{ij} de la siguiente forma:

$$c_{ij} = (-1)^{i+j} \alpha_{ij}$$

La regla de Laplace para el cálculo del determinante de una matriz dice que el determinante de una matriz es igual a la suma de los productos de los elementos de una columna (o una fila) por los adjuntos correspondientes. Es decir, si fijamos una columna j :

$$\det(A) = \sum_{i=1}^n a_{ij} c_{ij} = \sum_{i=1}^n a_{ij} (-1)^{i+j} \alpha_{ij}$$

Si fijamos una fila i :

$$\det(A) = \sum_{j=1}^n a_{ij} c_{ij} = \sum_{j=1}^n a_{ij} (-1)^{i+j} \alpha_{ij}$$



Ejemplo

Sea $A \in R^{4 \times 4}$ la siguiente matriz cuadrada:

$$A = \begin{pmatrix} 1 & -1 & 2 & 4 \\ 0 & 2 & 4 & 0 \\ 3 & -2 & 5 & 1 \\ 1 & 1 & 3 & -1 \end{pmatrix}$$

Utilizamos la regla de Laplace para el cálculo del determinante. Para reducir al mínimo el número de menores que tenemos que calcular, elegimos calcular el determinante fijando la segunda fila, dado que es la fila con el mayor número de ceros.

Así pues, el determinante de A se obtiene de la siguiente manera:

$$\det(A) = \sum_{j=1}^4 a_{2,j} (-1)^{2+j} \alpha_{2+j} = -a_{2,1} \alpha_{2,1} + a_{2,2} \alpha_{2,2} - a_{2,3} \alpha_{2,3} + a_{2,4} \alpha_{2,4}$$

No hay que calcular los menores ni $\alpha_{2,4}$, dado que $a_{2,1} = a_{2,4} = 0$ y los correspondientes términos desaparecen de la fórmula.

Necesitamos, entonces, calcular los menores $\alpha_{2,2}$ y $\alpha_{2,3}$. Dado que los menores son determinantes de matrices 3×3 , podemos utilizar la regla de Sarrus:

$$\alpha_{2,2} = \det \begin{pmatrix} 1 & 2 & 4 \\ 3 & 5 & 1 \\ 1 & 3 & -1 \end{pmatrix} = (-5 + 2 + 36) - (20 - 6 + 3) = 33 - 17 = 16$$

De forma similar, obtenemos que $\alpha_{2,3} = 17$.

Sustituyendo ahora en la fórmula del determinante, obtenemos que:

$$\det(A) = a_{2,2} \alpha_{2,2} - a_{2,3} \alpha_{2,3} = 2 \cdot 16 - 4 \cdot 17 = -36$$

Las propiedades fundamentales del determinante de una matriz son las siguientes:

- $\det(AB) = \det(A)\det(B)$ para cada $A, B \in R^{n \times n}$
- $\det(I_n) = 1$ para cada n
- $\det(A^t) = \det(A)$ para cada matriz $A \in R^{n \times n}$
- Si A es una matriz triangular (superior o inferior), entonces su determinante es el producto de los elementos en la diagonal, $\det(A) = a_{1,1}a_{2,2} \cdots a_{n,n}$
- Si Q es una matriz ortogonal, entonces $\det(Q) = 1$

1.2.2. Menores y rango

El concepto de menor complementario de una matriz cuadrada es un caso particular del concepto más general de **menor de una matriz**.

Un menor de una matriz cualquiera es el determinante de una submatriz cuadrada obtenida eliminando algunas filas y/o columnas. Decimos que el menor de una matriz es un menor $k \times k$ o un menor de orden k si es el determinante de una matriz de dimensiones $k \times k$.

El **rango de una matriz** A , indicado con la expresión $\text{rank}(A)$ o $\text{rk}(A)$, es el orden del menor más grande distinto de 0 de la matriz A . En particular, cada matriz distinta de la matriz de ceros tiene rango por lo menos 1, y para cada matriz $m \times n$ el rango es siempre menor o igual a la dimensión mínima, es decir $\text{rank}(A) \leq \min(n,m)$. Además las matrices cuadradas $n \times n$ con determinante distinto de 0 tienen rango máximo, es decir, $\text{rank}(A) = n$.



Ejemplo

Sea $B \in R^{2 \times 3}$ la siguiente matriz:

$$B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

Existen tres menores 2×2 :

$$\det \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} = -3$$

$$\det \begin{pmatrix} 1 & 3 \\ 4 & 6 \end{pmatrix} = -6$$

$$\det \begin{pmatrix} 2 & 3 \\ 5 & 6 \end{pmatrix} = -3$$

Dado que por lo menos uno de los menores es distinto de 0, la matriz tiene rango máximo $\text{rank}(B) = 2$.



Ejemplo

Sea $D \in R^{3 \times 2}$:

$$D = \begin{pmatrix} 1 & 3 \\ 2 & 6 \\ -4 & -12 \end{pmatrix}$$

El rango de D es igual a 1, como veremos ahora. Sabemos que $1 \leq \text{rank}(D) \leq 2$, y hay que comprobar los menores de orden 2 para saber si por lo menos uno de ellos es distinto de 0. Hay tres menores de orden 2, obtenidos eliminando una de las filas:

$$\det \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix} = 6 - 6 = 0$$

$$\det \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix} = 6 - 6 = 0$$

$$\det \begin{pmatrix} 2 & 6 \\ -4 & -12 \end{pmatrix} = -24 + 24 = 0$$

Dado que todos los menores de orden 2 son iguales a 0, entonces el rango de D es igual a 1.

1.2.3. Matriz inversa

Podemos ahora introducir el concepto de matriz inversa (con respecto al producto fila-columna). En la matriz cuadrada $A \in R^{n \times n}$, decimos que A es invertible, o que admite una inversa, si existe una matriz de la misma dimensión que A , que denotamos como A^{-1} , tal que:

$$AA^{-1} = A^{-1}A = I_n$$

A partir de la propiedad del determinante $\det(AB) = \det(A)\det(B)$, obtenemos que $1 = \det(AA^{-1}) = \det(A)\det(A^{-1})$. Entonces:

$$\det(A^{-1}) = 1/\det(A)$$

Y podemos escribir la igualdad solo si $\det(A) \neq 0$.

De hecho, se puede demostrar que las matrices invertibles coinciden con el conjunto de las matrices con determinante no nulo.



Es decir, una matriz es invertible si y solo si su determinante es distinto de 0. Cuando una matriz es tal que su determinante es igual a 0, decimos que es una **matriz singular**.

Para calcular la matriz inversa, podemos utilizar la matriz de cofactores C , donde los elementos son $c_{ij} = (-1)^{i+j} \alpha_{ij}$. Así pues, la matriz inversa A^{-1} se obtiene con la siguiente fórmula:

$$A^{-1} = \frac{1}{\det(A)} C^t$$

La matriz C^t es la traspuesta de la matriz de cofactores. A continuación, presentamos las propiedades principales de la matriz inversa:

- $(A^{-1})^{-1} = A$
- $(aB)^{-1} = \frac{1}{a}B^{-1}$
- $(A^t)^{-1} = (A^{-1})^t$, es decir, inversión y traspuesta comutan
- $\det(A^{-1}) = 1/\det(A)$
- $(AB)^{-1} = B^{-1}A^{-1}$

Además, podemos observar que, para las matrices ortogonales, la traspuesta y la inversa coinciden, dado que la matriz inversa es única.

1.2.4. Autovalores y autovectores

Sea $A \in R^{n \times n}$ una matriz cuadrada y sea $x \in R^n$ un vector (distinto del vector nulo). En ese caso, decimos que x es el autovector (o **vector propio**) asociado al autovalor (o **valor propio**) $\lambda \in R$ si:

$$Ax = \lambda x$$

Es decir, el resultado de multiplicar la matriz A y el vector x es igual a multiplicar el vector x por el escalar λ .

A partir de la ecuación anterior, se puede deducir que el conjunto de los autovalores es igual al conjunto de soluciones del siguiente polinomio:

$$p(\lambda) = \det(A - \lambda I_n)$$

El polinomio $p(\lambda)$ se llama **polinomio característico**.

Para cada raíz del polinomio característico λ_i , es posible encontrar el espacio de autovectores asociado resolviendo la siguiente ecuación en $v \in R^n$:

$$(A - \lambda_i I_n)v = 0$$



Ejemplo

Consideremos la siguiente matriz:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

El polinomio característico asociado es el siguiente:

$$p(\lambda) = \det(A - \lambda I_2) = \det\begin{pmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{pmatrix} = (1-\lambda)(4-\lambda) - 6$$

>>>

>>>

Entonces, para encontrar los autovalores, tenemos que encontrar las raíces del polinomio característico:

$$p(\lambda) = \lambda^2 - 5\lambda - 2 = 0$$

Resolviendo una simple ecuación de segundo grado, obtenemos los dos autovalores de la matriz A:

$$\lambda_1 = (5 + \sqrt{33})/2$$

$$\lambda_2 = (5 - \sqrt{33})/2$$

Una matriz es definida positiva si y solo si sus autovalores son todos positivos, y es definida negativa si sus autovalores son todos negativos. Además, el determinante de una matriz es igual al producto de sus autovalores.

1.3. Algunas factorizaciones de matrices

En este apartado veremos algunas factorizaciones conocidas para algunos tipos de matrices. Estas factorizaciones pueden ser muy útiles para resolver determinados problemas o para simplificar algunas operaciones entre matrices.

1.3.1. Factorización LU

Esta factorización se puede aplicar a todas las matrices cuadradas no singulares y permite escribir la matriz como producto de dos matrices (triangular inferior y triangular superior) de forma única. En las fórmulas, sea $A \in R^{n \times n}$ una matriz cuadrada no singular, es decir $\det(A) \neq 0$. Así pues, existen dos únicas matrices $L, U \in R^{n \times n}$, L triangular inferior y U triangular superior, tales que $A = LU$.

Si se conoce la factorización LU de una matriz, entonces es muy sencillo calcular su determinante, dado que el determinante de un producto de dos matrices es el producto de los determinantes de los factores. Además, al ser los factores matrices triangulares, sus determinantes son simplemente el producto de los elementos de las diagonales.

1.3.2. Factorización de Cholesky

La factorización de Cholesky es la descomposición de una matriz simétrica y definida positiva como producto de una matriz triangular inferior y su traspuesta:

$$S = LL^t \text{ para cada } S \in R^{n \times n} \text{ simétrica y definida positiva}$$

Esta factorización es única, es decir, la matriz L es única y es un caso particular de la factorización LU para matrices simétricas definidas positivas.

1.3.3. Factorización QR

La factorización QR descompone cada matriz cuadrada A en el producto de una matriz ortogonal Q y una matriz triangular superior R :

$$A = QR$$

La forma más sencilla de calcular la factorización QR es utilizar el método de Gram-Schmidt para ortogonalizar las columnas de la matriz A . El método de Gram-Schmidt es un algoritmo iterativo que transforma un conjunto de vectores en un conjunto de vectores ortogonales entre sí. A cada paso, el algoritmo modifica uno de los vectores restándole las proyecciones de los otros vectores ya modificados (calculando entonces los complementos ortogonales).

1.3.4. Diagonalización de una matriz

En la matriz cuadrada $A \in R^{n \times n}$, A es diagonalizable si existe una matriz invertible $P \in R^{n \times n}$ y una matriz diagonal $D \in R^{n \times n}$, tales que:

$$A = PDP^{-1}$$

Además, los elementos de la diagonal de D son los autovalores de la matriz A y las columnas de P son los autovectores correspondientes.

No todas las matrices son diagonalizables; una condición suficiente es que la matriz sea simétrica. Además, las matrices simétricas son diagonalizables ortogonalmente, es decir, que la matriz P es una matriz ortogonal y podemos escribir la descomposición siguiente:

$$A = PDP^t$$

1.3.5. Descomposición en valores singulares

Esta descomposición se puede aplicar a todas las matrices $A \in R^{m \times n}$, también a las no cuadradas. Existen dos matrices ortogonales $U \in R^{m \times m}$ y $V \in R^{n \times n}$ tales que:

$$A = U\Sigma V$$

Donde $\Sigma \in R^{m \times n}$ es una matriz de ceros con solo algunos valores distintos de 0 en la diagonal principal. Estos valores son los **valores singulares** de A y son iguales a la raíz cuadrada de los autovalores de la matriz $A^t A \in R^{n \times n}$, que, al ser una matriz simétrica, es siempre diagonalizable.

1.4. Sistemas de ecuaciones lineales

Una ecuación lineal en una variable es una ecuación como la siguiente:

$$ax - b = 0$$

Donde $a, b \in \mathbb{R}$ son coeficientes reales y x es la variable o incógnita. Resolver la ecuación significa encontrar el valor numérico que, al sustituir la x , haga verdadera la igualdad. La solución de una ecuación lineal en una variable existe siempre si $a \neq 0$ y se obtiene con la siguiente fórmula:

$$x = \frac{b}{a}$$

Consideremos ahora múltiples ecuaciones lineales en múltiples incógnitas x_1, x_2, \dots, x_n :

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,n}x_n = b_2 \\ \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \dots + a_{m,n}x_n = b_m \end{cases}$$

El conjunto de las m ecuaciones en n incógnitas es un sistema lineal $m \times n$, y la fórmula precedente es la representación del sistema en forma normal.

Cuando un sistema tiene solución, es decir, cuando existe una asignación de las variables x_1, x_2, \dots, x_n que satisface todas las ecuaciones, entonces se habla de **sistema compatible**. Si no existe ninguna solución, se habla de **sistema incompatible**. Los sistemas compatibles se pueden distinguir por el número de soluciones que admiten:

- Se llama **sistema compatible determinado** cuando existe una única solución.
- Se llama **sistema compatible indeterminado** cuando admite un conjunto infinito de soluciones.



Para resolver un sistema lineal, se puede utilizar el método de la sustitución, que consiste en resolver una de las ecuaciones respecto a una de las incógnitas, sustituir el resultado en las distintas ecuaciones y seguir así hasta obtener una ecuación con una sola incógnita.



Ejemplo

Consideremos el sistema lineal de dos ecuaciones con dos variables:

$$\begin{cases} 2x + 3y = 0 \\ x + 4y = 5 \end{cases}$$

Podemos resolver el sistema por sustitución, y a partir de la primera ecuación obtenemos que $x = \frac{-3}{2}y$. Al sustituir este resultado en la segunda ecuación, obtenemos la ecuación con la única incógnita y :

$$-3y + 8y = 10$$

>>>

>>>

Podemos simplificar esta ecuación y resolverla respecto a y calculando que $y = 2$. Después se puede sustituir el valor de y en $x = \frac{-3}{2}y$ obtener la solución del sistema: $x = -3, y = 2$. Está claro que esta es la única solución posible, así que este sistema es compatible y determinado.



Ejemplo

Consideremos ahora un sistema de tres ecuaciones con tres incógnitas:

$$\begin{cases} 3x_1 + x_2 = 0 \\ x_3 + x_2 + x_1 = -1 \\ -x_1 + \frac{x_3}{2} = 5 \end{cases}$$

Utilizamos el método de la sustitución. Desde la primera ecuación obtenemos $x_2 = -3x_1$. Al sustituirlo en la segunda ecuación, obtenemos:

$$\begin{cases} x_2 = -3x_1 \\ x_3 + -2x_1 = -1 \\ -x_1 + \frac{x_3}{2} = 5 \end{cases}$$

A continuación, a partir de la segunda ecuación obtenemos $x_3 = 2x_1 - 1$ y lo sustituimos en la tercera ecuación:

$$-x_1 + \frac{2x_1 - 1}{2} = 5$$

Simplificando esta ecuación, obtenemos que $-1 = 10$, lo cual es absurdo, de modo que el sistema no admite ninguna solución y es incompatible.

1.4.1. Forma matricial

Este es un sistema de ecuaciones lineales en forma normal:

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,n}x_n = b_2 \\ \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \dots + a_{m,n}x_n = b_m \end{cases}$$

El mismo sistema de ecuaciones se puede escribir en forma matricial:

$$Ax = b$$

Donde $R^n \ni A = (a_{i,j})$ es la matriz de coeficientes, $x = (x_1, x_2, \dots, x_n)^t$ es el vector columna de las variables y $b = (b_1, b_2, \dots, b_m)^t$ es el vector de los términos independientes. De ahora en adelante consideraremos solo los sistemas cuadrados, es decir, los sistemas de ecuaciones lineales en los que el número de incógnitas es el mismo que el número de ecuaciones.



Para los sistemas de ecuaciones lineales $n \times n$, la condición necesaria y suficiente para que el sistema sea compatible y determinado es que la matriz de coeficientes sea invertible, es decir, no singular ($\det(A) \neq 0$).

Si la matriz es invertible, entonces podemos multiplicar a la izquierda la ecuación del sistema (en forma matricial) por A^{-1} y obtener:

$$A^{-1}Ax = A^{-1}b$$

Y, dado que $A^{-1}A = I_n$ e $I_nx = x$, obtenemos la solución del sistema:

$$x = A^{-1}b$$

En general, resolver sistemas lineales es equivalente a invertir la matriz de coeficientes.

Por otro lado, en general, la condición necesaria y suficiente para que un sistema sea compatible (determinado o indeterminado) es que el rango de la matriz de los coeficientes sea igual al rango de la matriz de los coeficientes aumentada (o ampliada) $A|b$.

La matriz ampliada $A|b$ es igual a la matriz de los coeficientes más una columna igual a la matriz de los términos independientes.

1.4.2. Sistemas triangulares y método de Gauss

Consideremos ahora un caso particular de sistemas lineales en los que la matriz de coeficientes A es triangular. En este caso, el sistema es de la forma:

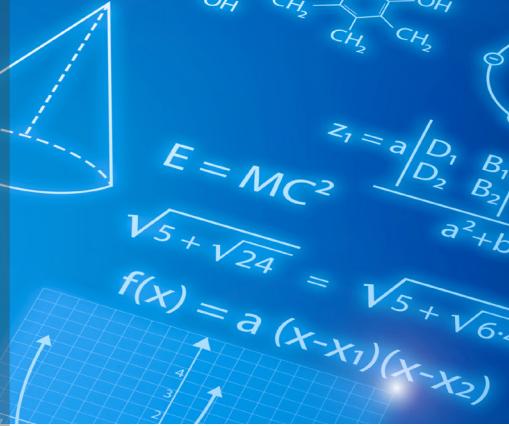
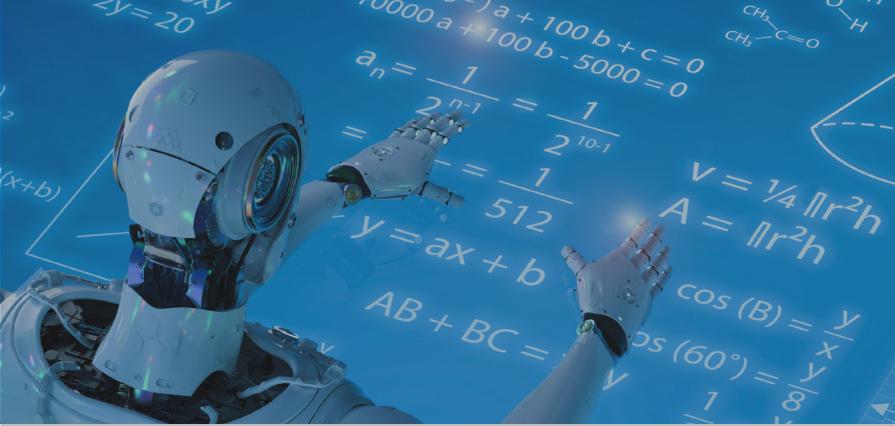
$$\left\{ \begin{array}{l} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n = b_1 \\ a_{2,2}x_2 + \dots + a_{2,n}x_n = b_2 \\ a_{3,3} + \dots + a_{3,n}x_n = b_3 \\ \vdots \\ a_{n,n}x_n = b_n \end{array} \right.$$

Así pues, empezando por la última ecuación, es muy sencillo aplicar el método de la sustitución y encontrar la solución del sistema lineal (si el sistema es compatible).

El método de Gauss (o método de eliminación gaussiana) es un algoritmo que, partiendo de un sistema lineal, utiliza algunas operaciones para transformar el sistema en un sistema triangular y así encontrar una solución. La factorización LU de la matriz de coeficientes es, a la práctica, una forma modificada del método de Gauss. Veamos ahora cómo es posible resolver un sistema lineal usando la factorización LU .

Sea $Ax = b$ un sistema lineal de n ecuaciones y n incógnitas, y sea $A = LU$ la factorización LU de la matriz de coeficientes ($L, U \in R^{n \times n}$ son dos matrices triangulares). Podemos resolver el sistema $LUx = b$ en dos pasos:

1. Resolvemos el sistema triangular $Lv = b$ para el vector de incógnitas $v \in R^n$.
2. Resolvemos el segundo sistema triangular $Ux = v_0$, donde v_0 es la solución del sistema triangular en el punto 1.



Capítulo 2

Análisis

En este capítulo veremos de forma rápida los conceptos básicos del análisis real —funciones reales de variables reales y límites de funciones— para definir el concepto de derivada de función. Veremos también que las derivadas se utilizan para encontrar máximos y mínimos de funciones reales.

En aprendizaje automático (*machine learning*) e inteligencia artificial es muy común tener que encontrar mínimos y máximos de funciones reales. Por ejemplo, al entrenar cualquier modelo predictivo, el objetivo es casi siempre minimizar algún tipo de error de predicción (o de coste) o maximizar una función objetivo que representa el problema.

2.1. Funciones reales y límites

2.1.1. Funciones

Una **función** (o **mapa**) $f : A \rightarrow B$ es una regla que asigna a cada elemento de A un único elemento del segundo conjunto B . A es el **dominio** de la función f o conjunto de definición. Por otra parte, el conjunto B de posibles valores que toma la función se llama **codominio** de f . Para cada valor $x \in A$, la salida o valor de retorno (*output*) de la función se denota con $f(x)$. La **imagen inversa** o **preimagen** de un subconjunto del codominio $C \subseteq B$, $f^{-1}(C)$ es el conjunto de puntos del dominio cuya imagen está en C :

$$f^{-1}(C) = \{x \in A \text{ tal que } f(x) \in C\}$$



Ejemplo

Sea f la función desde los números $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ en el conjunto de los meses del año. La función asigna a cada punto x su respectivo mes en el calendario, es decir, $f(1) = \text{enero}, f(2) = \text{febrero} \dots$ El dominio de f es el conjunto de números naturales entre 1 y 12 (incluido los extremos) y el codominio es el conjunto de los meses del año.

Hablando formalmente, una función está definida no solo por la regla o las operaciones matemáticas que permiten calcular la salida (*output*) de la función para cada entrada (*input*), sino también por el dominio y el codominio. Una función con la misma regla del ejemplo precedente, pero con los números naturales entre 3 y 8 como dominio, no sería la misma función que la del ejemplo.

El subconjunto de puntos del codominio, que efectivamente son la salida de algunos valores del dominio, se llama **Imagen de la función** y se denota con $\text{Img}(f)$. Está claro que $\text{Img}(f) \subseteq B$, donde B es el codominio de la función.

En este capítulo nos interesan las funciones de variable real a valores reales, es decir, cuando tanto el dominio como el codominio son subconjuntos de los números reales.



Ejemplo

Sea f la función desde $[1, 4]$ en $[1, 16]$, que asigna a cada punto x su cuadrado x^2 , es decir, $f(x) = x^2$. El dominio de f es el conjunto de puntos entre 1 y 4 (incluidos los extremos) y el codominio es el conjunto de puntos entre 1 y 16 (incluidos los extremos). Dado que tanto el dominio como el codominio son subconjuntos de los números reales, la función f es una función real de variable real.

Recordamos aquí la notación para los intervalos en los números reales:

- El intervalo cerrado $[a, b]$ indica el conjunto de números reales entre los valores a y b , extremos incluidos, es decir, $[a, b] = \{x \in R : a \leq x \leq b\}$
- Intervalo abierto: $(a, b) = \{x \in R : a < x < b\}$
- Intervalo abierto a la izquierda y cerrado a la derecha: $(a, b] = \{x \in R : a < x \leq b\}$
- Intervalo cerrado a la izquierda y abierto a la derecha: $[a, b) = \{x \in R : a \leq x < b\}$

En las funciones reales, muchas veces el dominio y el codominio no se especifican. En este caso, se toma como dominio de la función el subconjunto de los números reales, maximal con respecto a la inclusión, tales que la función está bien definida. Por esta razón, el dominio de una función también se llama **conjunto de definición**. El codominio de la función siempre puede ser considerado el conjunto de todos los números reales R .



Ejemplo

Sea $f(x) = \ln(x)$ la función logaritmo natural. El logaritmo natural solo es definido para números positivos y, entonces, podemos considerar como dominio de la función f el conjunto $\{x \in \mathbb{R} : x > 0\}$ o, en notación de intervalos $(0, +\infty)$.



Ejemplo

Sea $g(x) = 1/(x - 5)$. Dicha función se puede definir para todos $x \neq 5$, dado que para $x = 5$ el denominador de la fracción es 0. Entonces, como conjunto de definición de g , se puede considerar el conjunto de números reales $(-\infty, 5) \cup (5, +\infty)$.

2.1.2. Gráficas de funciones

Cada función real de variable real puede ser representada gráficamente mediante la **gráfica de función** asociada. Formalmente, dada una función real f , podemos definir la gráfica de la función como el subconjunto del plano cartesiano identificado por los puntos $(x, f(x))$ al variar $x \in A$ en el dominio de f .



Enlace de interés

Una forma muy rápida y sencilla de dibujar gráficas de funciones reales se puede encontrar en el buscador de Google. Basta con escribir la función que se quiere dibujar en el buscador y aparecerá la gráfica de la función. En la Figura 3 es posible ver el resultado de escribir la función $x^2 + x$ en el buscador.

www.google.com

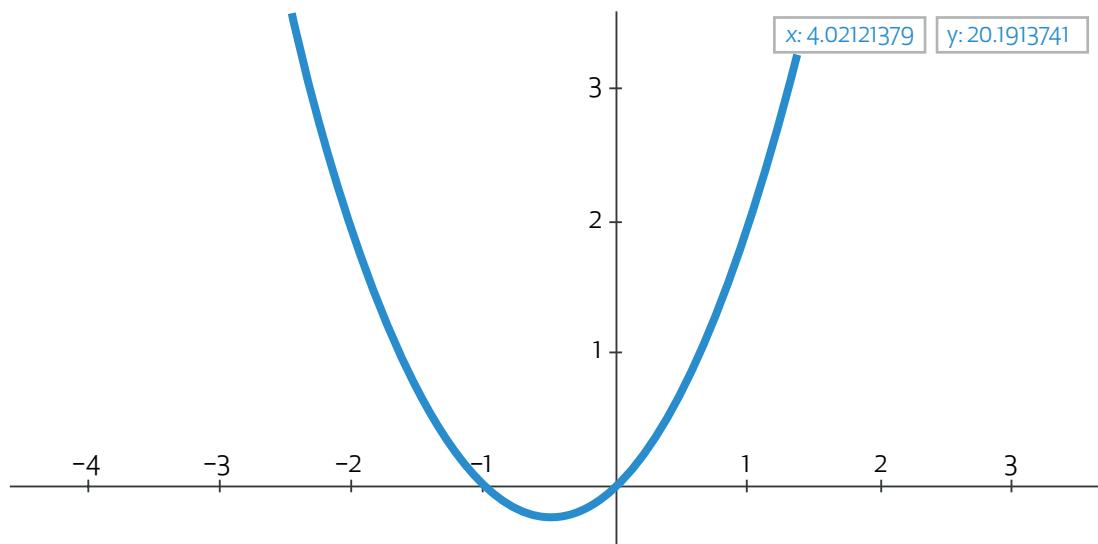


Figura 3. Gráfica de la función $f(x) = x^2 + x$.

2.1.3. Límites de funciones reales y continuidad

Definimos ahora el concepto de **límite de una función** real. Dada una función real f y un número $L \in \mathbb{R}$, decimos que el límite de $f(x)$ es L cuando x tiende a $c \in \mathbb{R}$, si podemos encontrar un valor x cerca de c tal que $f(x)$ esté tan cerca de L como se desee. Escribimos:

$$\lim_{x \rightarrow c} f(x) = L$$

Formalmente, el límite de una función se puede definir con la definición épsilon-delta. Decimos que $\lim_{x \rightarrow c} f(x) = L$ si y solo si para cualquier número $\varepsilon > 0$ existe un número $\delta > 0$ tal que, si $|x - c| < \delta$, entonces $|f(x) - L| < \varepsilon$.

Es importante recordar que el límite de una función en un punto dado puede no existir. En otras palabras, no existe ningún número L tal que el límite de la función sea igual a L .

Se dice que una función real es una **función continua** en el punto p si el límite de la función es igual al valor de la función $f(p)$:

$$\lim_{x \rightarrow p} f(x) = f(p)$$

Entonces, para que una función sea continua en un punto p , tienen que cumplirse las siguientes condiciones:

- $f(p)$ existe, es decir, p pertenece al dominio de definición de f .
- El límite $\lim_{x \rightarrow p} f(x)$ existe.
- El límite es igual al valor de la función en el punto $\lim_{x \rightarrow p} f(x) = f(p)$.

Se dice que una función es continua si es continua en todos los puntos de su dominio. El conjunto de las funciones reales continuas se denota con $C(\mathbb{R})$.

2.2. Derivadas de funciones

La **derivada de una función** en un punto x es un valor numérico que representa cómo dicha función cambia al variar el valor de x . Geométricamente, se puede interpretar como el coeficiente de la recta tangente a la gráfica de la función en el punto $(x, f(x))$.

2.2.1. La derivada como límite

La derivada de una función f en x se define con el siguiente límite (si este existe):

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Si la derivada existe, es decir, si el límite es definido en el punto x , decimos que la función f es derivable en x y denotamos con $f'(x)$ el valor de la derivada en el punto x . Existe otra notación para la operación de derivación:

$$f'(x) = \frac{df(x)}{dx}$$

Esta notación será útil más adelante, cuando consideraremos las funciones con más de una variable. En ese momento, tendremos que especificar respecto a qué variable se deriva la función.

Podemos observar que la derivada de una función es igual al límite de las pendientes de las rectas secantes de la gráfica de la función en los puntos $(x, f(x))$ y $(x + h, f(x + h))$. Como se puede ver en la Figura 4, el límite de estas rectas es exactamente la recta tangente en el punto $(x, f(x))$. Tras esta observación podemos concluir la primera regla para calcular derivadas, dado que la recta tangente a una línea recta es ella misma y que las líneas rectas son las gráficas de funciones del tipo $f(x) = ax + b$. Para estas funciones, $f'(x) = a$, es decir, la derivada de una función lineal es una constante igual al coeficiente de la función. En el caso particular $f(x) = b$, obtenemos que la derivada de una función constante ($a = 0$) es 0.

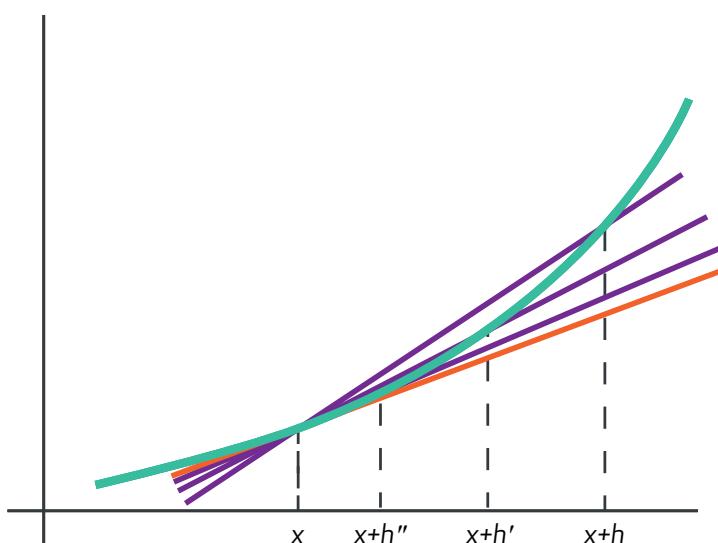


Figura 4. La recta tangente como límite de rectas secantes. Por Pbroks13 bajo licencia CC BY-SA 3.0. Disponible en <https://en.wikipedia.org/wiki/Derivative#/media/File:Lim-secant.svg>

Si la función es derivable para cada punto en un conjunto A , entonces se dice que la función es derivable en A . Está claro que la derivada f' es una función definida en el conjunto de puntos donde f es derivable. Además, es sencillo observar que, si una función es derivable en x , entonces tiene que ser continua en x . El hecho de que una función sea continua no es condición suficiente para su derivabilidad, sino que hay funciones que son continuas pero no derivables en algunos puntos.



Ejemplo

El ejemplo clásico de una función continua pero no derivable es la función valor absoluto. Sea $f(x) = |x|$. Para cada $x > 0$, la función es igual a la función $f(x) = x$, cuya derivada es 1. De forma similar, para $x < 0$, la derivada es constante e igual a -1. En el punto $x = 0$, el límite $\lim_{h \rightarrow 0} (f(h) - f(0))/h$ no existe, dado que para los valores de $h < 0$, el cociente es -1 y, para $h > 0$, el cociente es +1.

2.2.2. Calcular la derivada de una función

Veamos ahora las reglas básicas para calcular las derivadas de una función. En particular, veremos cuáles son las derivadas de algunas funciones básicas (polinomios, $\ln(x)$, e^x , $\sin(x)$, $\cos(x)$...) y cómo se calculan las derivadas de sumas, productos, fracciones y composiciones de funciones. De esta forma, es posible calcular las derivadas de todas las funciones que son combinaciones de funciones básicas.

A continuación se encuentran las derivadas de la mayoría de las funciones:

- Monomios: si $f(x) = x^n$, entonces $f'(x) = nx^{n-1}$
- Polinomios: si $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$, entonces $f'(x) = a_1 + 2a_2x^1 + 3a_3x^2 + \dots + na_nx^{n-1}$
- Exponenciales: si $f(x) = e^x$, entonces $f'(x) = e^x$, es decir, la derivada de la función exponencial es la propia función exponencial
- Logaritmos: si $f(x) = \ln(x)$, entonces $f'(x) = 1/x$ para $x > 0$
- Funciones trigonométricas:
 - si $f(x) = \sin(x)$, entonces $f'(x) = \cos(x)$
 - si $f(x) = \cos(x)$, entonces $f'(x) = -\sin(x)$
 - si $f(x) = \tan(x)$, entonces $f'(x) = 1/\cos(x)^2 = 1 + \tan(x)^2$
 - si $f(x) = \cot(x)$, entonces $f'(x) = -1/\sin(x)^2 = -(1 + \cot(x)^2)$

Las derivadas precedentes se pueden combinar entre ellas de las siguientes formas:

- Suma de funciones: si $f(x) = g(x) + h(x)$, entonces $f'(x) = g'(x) + h'(x)$
- Producto por escalar: si $f(x) = ag(x)$ para $a \in \mathbb{R}$, entonces $f'(x) = ag'(x)$
- Producto de dos funciones: si $f(x) = g(x)h(x)$, entonces $f'(x) = g'(x)h(x) + g(x)h'(x)$
- Cociente de dos funciones: si $f(x) = g(x)/h(x)$, entonces $f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$
- Composición de funciones: si $f(x) = g(h(x))$, entonces $f'(x) = g'(h(x))h'(x)$



Ejemplo

Sea $f(x) = \exp(5x^3 - x + x^2 - 1)$. Aquí la derivada se puede calcular utilizando la regla de derivación para composición de funciones y las derivadas básicas para la función exponencial y para los polinomios. En particular, podemos escribir $f(x) = g(h(x))$, donde $h(x) = 5x^3 - x + x^2 - 1$ y $g(t) = \exp(t)$. Gracias a las reglas de derivación para polinomios y función exponencial, sabemos que $h'(x) = 15x^2 - 1 + 2x$ y que $g'(t) = e^t$. Ahora, utilizando la fórmula para derivar las funciones compuestas, obtenemos que:

$$f'(x) = \exp(5x^3 - x + x^2 - 1)(15x^2 - 1 + 2x)$$



Ejemplo

Sea $g(x) = 2x^4 + \sin(3x)$. Para calcular la derivada de g , tenemos que sumar las derivadas de los dos términos de la suma. Entonces obtenemos $g'(x) = (2x^4)' + \cos(3x)(3x)' = 8x^3 + 3\cos(3x)$, donde la derivada de $\sin(3x)$ se obtiene aplicando la regla de derivación para la composición de las dos funciones $\sin(t)$ y $3x$.

2.2.3. Funciones monótonas

Se dice que una función f definida en un conjunto (a, b) es una **función monótona creciente** si:

$$f(x) > f(y) \text{ para cada } x > y, x, y \in (a, b)$$

Se dice que una función f es una **función monótona decreciente** si:

$$f(x) < f(y) \text{ para cada } x > y, x, y \in (a, b)$$

Si una función es monótona creciente, entonces su gráfica será una línea creciente (Figura 5). Y si, por el contrario, una función es monótona decreciente, su gráfica será una línea decreciente al aumentar los valores de la variable x (Figura 6).

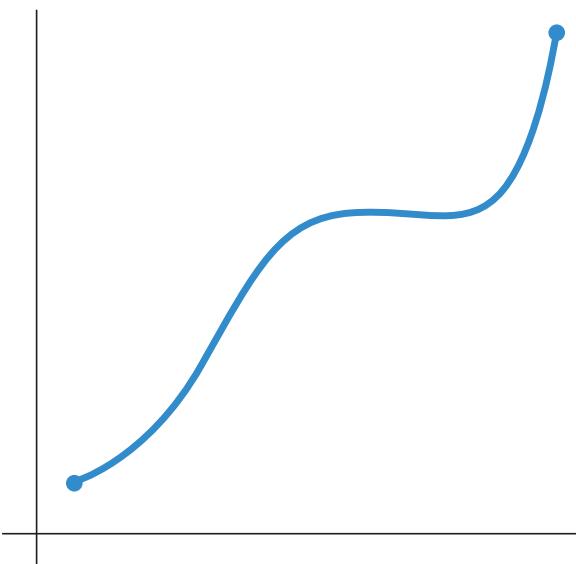


Figura 5. Gráfica de una función monótona creciente.

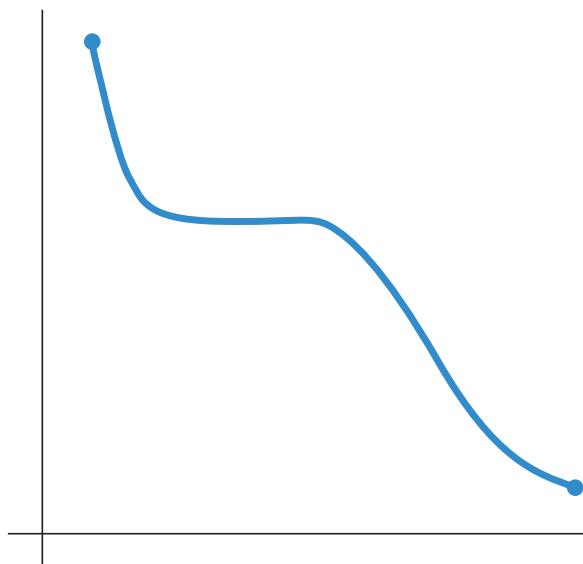


Figura 6. Gráfica de una función monótona decreciente.

Si consideramos las funciones derivables, entonces podemos relacionar el signo de la derivada con el comportamiento de las gráficas de las funciones.



En particular, si la derivada es positiva en un punto x , entonces la recta tangente tiene pendiente positiva y la función es creciente alrededor del punto x . Por otro lado, si la derivada es negativa, entonces la función es decreciente alrededor del punto x .

De esta forma, comprobando el signo de la derivada $f'(x)$ de una función $f(x)$, es posible entender cómo la función crece o decrece al variar la variable x y dibujar, por lo tanto, una gráfica aproximada de $f(x)$.



Ejemplo

Sea $g(x) = \ln(x) + x^2$ para valores positivos. Su derivada es igual a $g'(x) = 1/x + 2x = (1 + 2x^2)/x$ y es positiva para todos los valores en el dominio de la función. Así pues, podemos saber, sin tener que dibujar la gráfica con un ordenador y de forma más rigurosa, que la función g es monótona creciente para todos los valores en su dominio.



Ejemplo

Sea $h(x) = 4x - x^2$. Su derivada es claramente $h'(x) = 4 - 2x$ y, si observamos la desigualdad $h'(x) > 0$, vemos que la derivada es positiva para $x < 2$ y negativa para $x > 2$. Como se puede ver en la gráfica de la función en la Figura 7, $h(x)$ es creciente para $x < 2$ y decreciente para $x > 2$. El punto $x = 2$ (donde la derivada es 0) es el máximo de la función. Esta observación se puede extender a casos generales, como veremos en el próximo apartado.

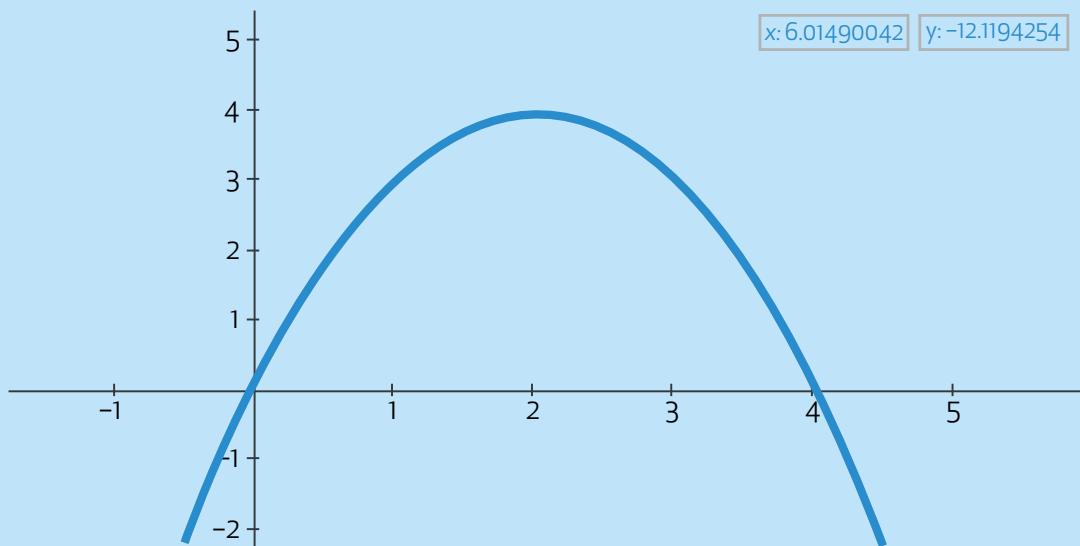


Figura 7. Gráfica de la función $h(x) = 4x - x^2$.

2.3. Máximos y mínimos de funciones reales

En este apartado veremos que, gracias al estudio de la derivada de una función, es posible encontrar máximos y mínimos de dicha función. Primero definimos con precisión qué entendemos por máximos y mínimos de una función.

2.3.1. Máximos y mínimos absolutos

Un punto de máximo, o más precisamente **punto de máximo absoluto** (o **punto de máximo global**) de una función f , definida en el dominio A , es cada punto $x_0 \in A$ tal que:

$$f(x_0) \geq f(x) \text{ para cada } x \in A$$

De forma opuesta, un **punto de mínimo absoluto** (o **punto de mínimo global**) es cada punto $x_0 \in A$ tal que:

$$f(x_0) \leq f(x) \text{ para cada } x \in A$$

Una función puede no tener máximos y/o mínimos absolutos. Si la función f no admite ningún máximo absoluto, entonces, desde la definición, obtenemos que para cada $x_0 \in A$ existe otro punto en el dominio $x \in A$ tal que $f(x) \geq f(x_0)$ y, entonces, eligiendo varios puntos en el dominio de la función podemos alcanzar valores siempre mayores. De forma similar, si la función no admite un mínimo absoluto, entonces podemos alcanzar valores siempre menores.



Ejemplo

Sea $f(x) = -x^2$. Está claro que la función f es menor o igual que 0 para cada $x \in R$ y, además, $f(0) = 0$. Así pues, $x_0 = 0$ es el (único) punto de máximo absoluto de f . Se puede también observar que la función f no admite ningún mínimo absoluto en su dominio. De hecho, si hacemos crecer los valores de x , podemos alcanzar cualquier valor negativo.

Una función puede no admitir máximos y/o mínimos absolutos cuando estos valores están en el borde del dominio (que no es cerrado). Estos casos suelen ocurrir cuando el dominio de la función está restringido por razones ajenas a la propia función.



Ejemplo

Sea $h(x) = x$ definida en el intervalo abierto $(0,2)$. Está claro que la gráfica de dicha función es una recta entre los puntos $(0,0)$ y $(2,2)$, pero estos dos puntos son excluidos por la definición del dominio de h como intervalo abierto. Entonces, h no admite ni máximo ni mínimo absoluto. Podemos comprobar este hecho para el mínimo de la siguiente manera: si $x_0 \in (0,2)$, tenemos que $x = x_0/2 \in (0, 2)$ y, además, $f(x) = x = x_0/2 < x_0 = f(x_0)$. Por lo tanto, x_0 no puede ser punto de mínimo absoluto. Dado que el razonamiento se puede repetir para cualquier $x_0 \in (0, 2)$, hemos demostrado que h no admite mínimo en su dominio.

El teorema de Weierstrass, o teorema de los valores extremos, establece que una función continua definida en un intervalo cerrado y acotado (o una unión de intervalos cerrados y acotados) admite puntos de máximo y mínimo absolutos en dicho intervalo.

2.3.2. Máximos y mínimos relativos y puntos críticos

Otros tipos de puntos de máximo y mínimo son los **punto de máximo relativo (o local) y el punto de mínimo relativo (o local)** (véase Figura 8). Un punto de máximo relativo o local de una función $f: A \subseteq R \rightarrow R$ es un punto x_0 tal que existe un entorno $B(x_0, \delta) = (x_0 - \delta, x_0 + \delta) \cap A$ y se cumple:

$$f(x_0) \geq f(x) \text{ para cada } x \in B(x_0, \delta) \cap A$$

Es decir, x_0 es punto de máximo absoluto en un entorno. De forma similar, se definen los puntos de mínimos relativos o locales. Observamos que todos los puntos extremos absolutos de una función son también extremos relativos. Si la función f es derivable, los puntos extremos coinciden con los puntos críticos de f .

Un **punto crítico (o punto estacionario)** de una función f es un punto x_0 perteneciente al dominio de f tal que la derivada f' es nula, es decir $f'(x_0) = 0$. Además, se suelen añadir a los puntos críticos también los puntos donde la función no es derivable y los extremos del dominio.

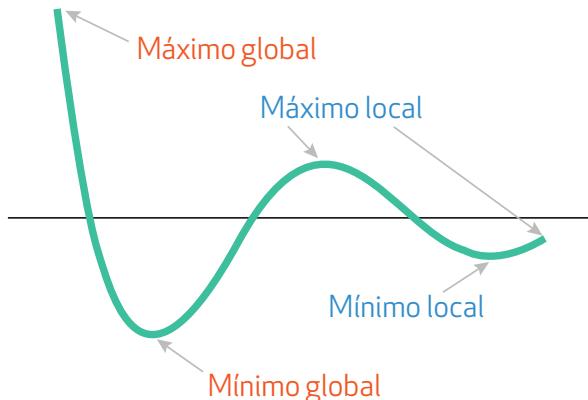


Figura 8. Máximos y mínimos de una función.

El teorema de Fermat expresa la conexión entre puntos críticos y extremos de una función. Si una función f alcanza un máximo o mínimo relativo en el punto x_0 , y si la derivada existe en dicho punto x_0 , entonces $f'(x_0) = 0$. Dicho de otra forma, el teorema de Fermat afirma que ser punto estacionario es condición necesaria para ser punto máximo o mínimo local.



Ejemplo

Sea $g(x) = x^3 - 6x^2 + 1$ una función polinómica y, por lo tanto, derivable en todos los valores. Sus puntos críticos se encuentran resolviendo la ecuación $g'(x) = 0$.

$$g'(x) = 3x^2 - 12x = 0$$

$$3x(x - 4) = 0$$

>>>

>>>

Así pues, las soluciones son $x_1 = 0$ y $x_2 = 4$, que son los dos puntos críticos de g . Para saber si los puntos críticos son puntos de máximo o mínimo locales hay que observar el signo de la derivada.

$$g'(x) = \begin{cases} g'(x) > 0 & \text{para } x < 0 \\ g'(x) < 0 & \text{para } 0 < x < 4 \\ g'(x) > 0 & \text{para } x > 4 \end{cases}$$

Así pues, está claro que $x_1 = 0$ es punto de máximo local y $x_2 = 4$ es punto de mínimo local.



Ejemplo

Consideremos ahora un ejemplo de función no derivable. Sea $f(x) = |x| + x^3$. Esta función no es derivable en 0, dado que el valor absoluto no lo es. De todos modos, podemos calcular la derivada de forma usual para valores distintos de 0:

$$f'(x) = \begin{cases} -1 + 3x^2 & \text{para } x < 0 \\ 1 + 3x^2 & \text{para } x > 0 \end{cases}$$

Vemos que la derivada es siempre positiva para $x > 0$, y es 0 para $x_0 = -\sqrt{1/3}$, cambiando de signo en este punto. En particular, $f'(x) < 0$ para $x_0 < x < 0$, y x_0 es punto de máximo local.

Dado que la función es continua en cada punto, y dado que la derivada cambia de signo en 0 (de negativa a positiva), 0 es punto de mínimo local.

Fijaos en que en 0 la derivada no está definida y este punto no se puede encontrar a partir de la ecuación $f'(x) = 0$.

2.3.3. Derivadas de órdenes superiores

Dado que la derivada de una función es una función real, podemos derivar otra vez la derivada de una función (y sucesivamente derivar una y otra vez).

De esta forma se definen las derivadas de una función de orden superior, de modo que podemos considerar la derivada de orden 2 o derivada segunda $f''(x) = (f'(x))'$, la derivada de orden 3. En general, se puede definir de forma iterativa la derivada de orden n :

$$f^n(x) = (f^{n-1}(x))' = \frac{d^n f(x)}{dx^n}$$

Donde los exponentes indican el grado de la derivada.

La derivada segunda es útil en la búsqueda de los puntos de máximo y mínimo locales.

En particular, a través del signo de la derivada segunda en el punto crítico, es posible diferenciar entre máximo y mínimo. Si $f'(x_0) = 0$, es decir, si x_0 es punto crítico, entonces:

- Si $f''(x_0) > 0$, el punto x_0 es un punto de mínimo local.
- Si $f''(x_0) < 0$, el punto x_0 es un punto de máximo local.



Ejemplo

Si $h(x) = \sin(x)\cos(x)$, la derivada prima es:

$$h'(x) = \cos(x)^2 - \sin(x)^2$$

Por lo tanto, los puntos críticos se obtienen resolviendo la siguiente ecuación:

$$\cos(x)^2 = \sin(x)^2$$

$$\cos(x) = \pm \sin(x)$$

La ecuación anterior nos proporciona el conjunto de soluciones:

$$x = k\frac{\pi}{4} \quad k = 0, \pm 1, \pm 2$$

Utilizamos ahora la derivada segunda para identificar los máximos y los mínimos:

$$h''(x) = -2\cos(x)\sin(x) - 2\sin(x)\cos(x) = -4\sin(x)\cos(x)$$

Observamos ahora que $\cos\left(\pm\frac{\pi}{4} + 2k\pi\right)$ son valores positivos y $\cos\left(\pm 3\frac{\pi}{4} + 2k\pi\right)$ son valores negativos, mientras que, para la función \sin , obtenemos que $\sin\left(\frac{\pi}{4} + 2k\pi\right) > 0$, $\sin\left(3\frac{\pi}{4} + 2k\pi\right) > 0$ y $\sin\left(-\frac{\pi}{4} + 2k\pi\right) < 0$, $\sin\left(-3\frac{\pi}{4} + 2k\pi\right) < 0$. Juntando estas desigualdades obtenemos:

- $h''\left(\frac{\pi}{4} + 2k\pi\right) = -4\sin\left(\frac{\pi}{4} + 2k\pi\right)\cos\left(\frac{\pi}{4} + 2k\pi\right) < 0$ y entonces $\frac{\pi}{4} + 2k\pi$ son puntos de máximo.
- $h''\left(-\frac{\pi}{4} + 2k\pi\right) = -4\sin\left(-\frac{\pi}{4} + 2k\pi\right)\cos\left(-\frac{\pi}{4} + 2k\pi\right) > 0$ y entonces $-\frac{\pi}{4} + 2k\pi$ son puntos de mínimo.

De forma similar podemos obtener los resultados para los otros dos conjuntos de puntos críticos.

2.4. Funciones de diversas variables

Consideremos ahora funciones con más de una variable real a valores en los números reales, es decir, una función como la siguiente:

$$f: A \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$$

Para las funciones de dos variables, es posible dibujar las gráficas asociadas en tres dimensiones. Sin embargo, si hay más de dos variables, no es posible dibujar la gráfica de la función, dado que la gráfica de una función de n variables es un subconjunto de un espacio $n+1$ -dimensional.



Ejemplo

Sea $g(x,y)$ la función de dos variables definida por $g(x,y) = x^2 + e^y$. La gráfica obtenida a través del buscador de Google se puede ver en la Figura 9.

La función g en el punto $(0,0)$ toma el valor $g(0,0) = 0^2 + e^0 = 0 + 1 = 1$ y en el punto $(2, \ln(3))$ el valor $g(2, \ln(3)) = 2^2 + e^{\ln(3)} = 4 + 3 = 7$. Es fácil comprobar que la función es definida y positiva para todos valores de $(x,y) \in \mathbb{R}^2$.

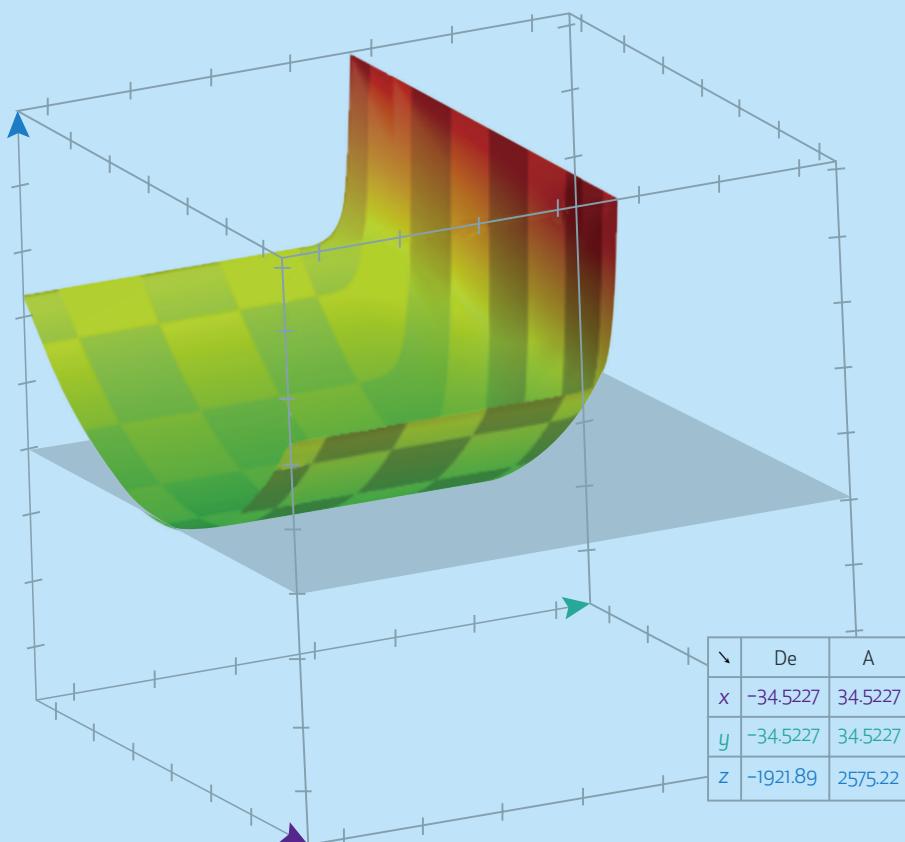


Figura 9. Gráfica de la función $g(x,y) = x^2 + e^y$.

2.4.1. Derivadas parciales

También las funciones de diversas variables se pueden derivar, pero, a diferencia de las funciones de una sola variable, no hay una sola derivada. Para cada variable, es posible definir una derivada prima de la función mientras se fijan el resto de las variables. En particular, si $f(x_1, x_2, \dots, x_n) = f(x)$ es una función de n variables, entonces la **derivada parcial** respecto a la variable x_i se denota como:

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_n)}{h}$$

A veces, el símbolo f_{x_i} se puede utilizar en vez de $\frac{\partial f}{\partial x_i}$ para la derivada parcial de f respecto a la variable x_i .

Para calcular las derivadas parciales de una función, se utilizan las mismas reglas que en el caso de una sola variable y se consideran las demás variables como constantes.



Ejemplo

Si $h(x_1, x_2, x_3) = x_1^3 x_2 + x_1 \ln(x_2) + 4x_3$, podemos calcular las tres derivadas parciales respecto a las distintas variables con las mismas reglas de las derivadas en el caso de las funciones de una sola variable.

$$\frac{\partial h}{\partial x_1}(x_1, x_2, x_3) = \frac{\partial(x_1^3 x_2)}{\partial x_1} + \frac{\partial(x_1 \ln(x_2))}{\partial x_1} + \frac{\partial(4x_3)}{\partial x_1} = 3x_1^2 x_2 + \ln(x_2) + 0$$

$$\frac{\partial h}{\partial x_2}(x_1, x_2, x_3) = \frac{\partial(x_1^3 x_2)}{\partial x_2} + \frac{\partial(x_1 \ln(x_2))}{\partial x_2} + \frac{\partial(4x_3)}{\partial x_2} = x_1^3 + \frac{x_1}{x_2} + 0$$

$$\frac{\partial h}{\partial x_3}(x_1, x_2, x_3) = \frac{\partial(x_1^3 x_2)}{\partial x_3} + \frac{\partial(x_1 \ln(x_2))}{\partial x_3} + \frac{\partial(4x_3)}{\partial x_3} = 0 + 0 + 4$$

Por ejemplo, para calcular la derivada $\frac{\partial(x_1^3 x_2)}{\partial x_1}$, consideramos x_2 como una constante y derivamos respecto a x_1 . La constante x_2 se puede sacar de la derivación y nos queda $x_2 \frac{\partial(x_1^3)}{\partial x_1}$, que es ahora una derivada usual de un monomio, de modo que obtenemos $x_2(3x_1^2)$.



Ejemplo

Si $f(x, y) = \sin(xy)$ y calculamos la derivada parcial respecto a x , tenemos que considerar y como una constante y aplicar las reglas de las derivaciones.

$$\frac{\partial f}{\partial x}(x, y) = \frac{\partial \sin(xy)}{\partial x} = y \cos(xy)$$

2.4.2. Gradiente y derivadas direccionales

El **gradiente de una función** de n variables reales es el vector de dimensión n cuyos componentes son las derivadas parciales de la función respecto a las distintas variables. Formalmente, el gradiente de una función $f(x_1, \dots, x_n)$ se indica con el símbolo $\nabla f(x_1, \dots, x_n)$ y es el vector de derivadas parciales:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

La **derivada direccional** de una función $f: A \subseteq R^n \rightarrow R$ es la derivada a lo largo de una dirección fija en el espacio vectorial de las variables. En particular, se define para cada vector $v \in A \subseteq R^N$ en el dominio de la función:

$$\frac{\partial f}{\partial v}(x) = \lim_{h \rightarrow 0} \frac{f(x + hv) - f(x)}{h}$$

La derivada direccional se calcula a través del gradiente de la función con la siguiente fórmula:

$$\frac{\partial f}{\partial v}(x) = \nabla f(x) \cdot v$$

Donde el producto es un producto escalar entre vectores en R^n .

Ejemplo

Consideremos la función de dos variables $g(x,y) = 2xy + y^2$ y veamos cómo calcular la derivada direccional en la dirección $v = (1,2)$. Primero, calculamos el gradiente de la función ∇g :

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{pmatrix} = \begin{pmatrix} 2y \\ 2x + 2y \end{pmatrix}$$

Ahora podemos calcular la derivada direccional de la función a través del producto escalar entre el gradiente y la dirección:

$$\frac{\partial g}{\partial v}(x,y) = \nabla g(x,y) \cdot v = (2y, 2x + 2y) \cdot (1,2) = 2y + 2(2x + 2y) = 4x + 6y$$

El gradiente de una función f en un punto $x \in R^n$ indica también la dirección que maximiza la derivada direccional de la función. Se puede demostrar que para cada vector:

$$\|v\|_2 = \max_{w \in R^n, \|w\|_2=1} v \cdot w = v \cdot \frac{v}{\|v\|_2}$$

Así pues, para el vector gradiente $\nabla f(x)$, la dirección que maximiza el producto escalar es exactamente la dirección del gradiente.

2.4.3. Segundas derivadas parciales y matriz hessiana

Igual que en el caso de las funciones de una sola variable, es posible definir las derivadas parciales de orden mayor que 1; la diferencia es que para las funciones de diversas variables podemos mezclar las derivadas.

En particular, para las derivadas parciales de orden 2 obtenemos todas las derivadas de la forma $\frac{\partial^2 f}{\partial x_i \partial x_j}$.

Si $i = j$, escribimos simplemente $\frac{\partial^2 f}{\partial x_i^2}$. Además, el teorema de Schwarz afirma que, bajo hipótesis de regularidad (es decir las derivadas segundas existen y son continuas), el orden de las derivadas parciales no influye:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

Las derivadas segundas parciales definen, pues, una matriz simétrica de dimensiones $n \times n$, que se llama matriz hessiana:

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

2.4.4. Máximos, mínimos y puntos críticos

Los puntos de máximo y mínimo (locales y globales) de una función $f: R^n \rightarrow R$ se definen exactamente como para las funciones de una variable real.

Los puntos críticos de una función de diversas variables son los puntos tales que el vector gradiente es igual al vector de ceros.

Así pues, $x = (x_1, \dots, x_n)$ es un punto crítico si:

$$\nabla f(x_1, \dots, x_n) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Es decir, $\frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = 0$ para cada variable x_i .



Ejemplo

Consideremos la función de dos variables $g(x, y) = e^{-x^2-y^2}$ y calculemos sus puntos críticos (a partir de la gráfica de la función en Figura 10, entendemos que debería haber un solo punto crítico).

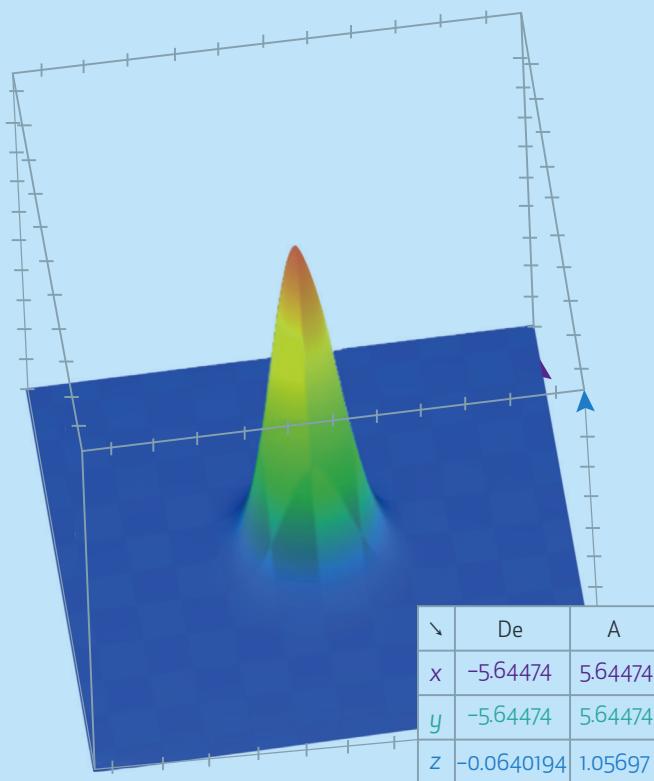


Figura 10. Gráfica de la función $g(x, y) = e^{-x^2-y^2}$.

El gradiente de la función es igual a:

$$\nabla g(x, y) = \begin{pmatrix} \frac{\partial g}{\partial x}(x, y) \\ \frac{\partial g}{\partial y}(x, y) \end{pmatrix} = \begin{pmatrix} -2xe^{-x^2-y^2} \\ -2ye^{-x^2-y^2} \end{pmatrix}$$

Está claro entonces que el único punto crítico es el punto $(0,0)$.

2.5. Integrales

Terminamos el capítulo sobre las funciones reales con uno de los conceptos más importante del análisis, la **integral de una función**. El tema de la integración de funciones reales es muy complejo y, en este apartado, solo pretendemos introducir los elementos básicos para entender la notación y la utilidad de las integrales.

2.5.1. Integrales de funciones

Existen varias definiciones de la integral de una función real, por lo que explicaremos ahora la más intuitiva: la integración de Riemann.

Sea f una función definida sobre un intervalo acotado $[a,b]$. Supongamos también que la función es positiva, es decir $f(x) \geq 0$ para cada $x \in [a, b]$.

Consideremos ahora el problema de calcular el área entre la gráfica de la función y y el eje de las x (Figura 11).

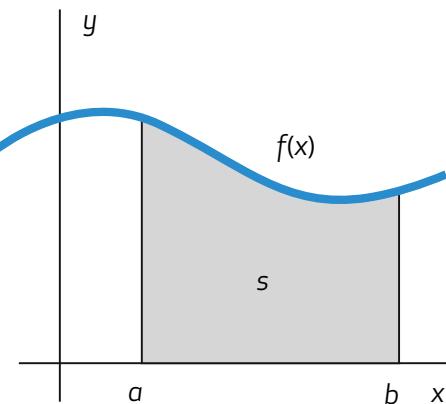


Figura 11. El área entre la gráfica de la función y y el eje de las x .

Una primera aproximación se puede construir dividiendo el intervalo $[a,b]$ mediante una partición, es decir, una secuencia de números $a = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = b$, y construir una aproximación del área mediante rectángulos (Figura 12). Escribimos la siguiente fórmula:

$$\text{Área} \approx \sum f(x_i)(x_i - x_{i-1}) = S(n)$$

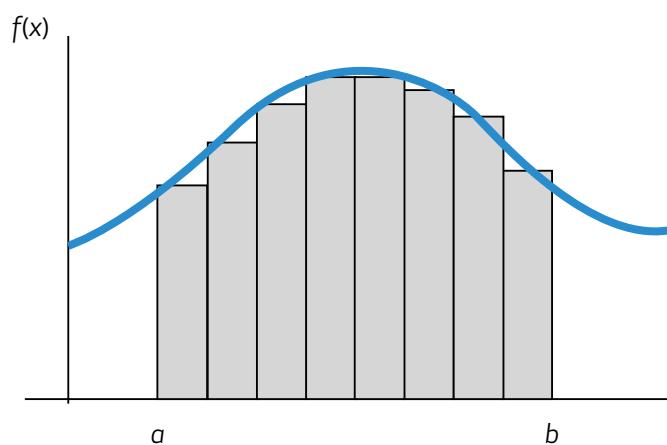


Figura 12. Aproximación del área bajo la gráfica de la función.

Intuitivamente, si aumentamos el número de puntos en la partición, la aproximación obtenida con $S(n)$ es cada vez más precisa y, en el límite para $n \rightarrow +\infty$, el valor de la suma (llamada **suma de Riemann**) es igual al área bajo la función.

Si existe el límite de las sumas de Riemann, este límite se llama integral y se denota de la siguiente forma:

$$\int_a^b f(x)dx$$

Donde se indican los **extremos de integración** a y b , y se explica con el símbolo dx cuál es la variable que nos interesa integrar.

Los límites de integración pueden también ser $a = -\infty$ y/o $b = +\infty$. Se puede, además, escribir el conjunto de integración debajo del símbolo de integral:

$$\int_{a,b} f(x)dx$$

Para las funciones no positivas, la integral se define descomponiendo las funciones en partes positivas y partes negativas:

$$\int f(x)dx = \int f^+(x)dx - \int f^-(x)dx$$

Donde $f^+(x) = \max\{f(x), 0\}$ y $f^-(x) = \max\{-f(x), 0\}$ son la parte positiva y negativa de la función.

A partir de la integral de una función, podemos definir una nueva función, llamada **integral indefinida**:

$$F(x) = \int_a^x f(t)dt$$

2.5.2. Primitivas de funciones y teorema fundamental del cálculo

La **primitiva de una función** es la inversa de la derivada, es decir, que para una función f la primitiva es una función $F(x)$ tal que $F'(x) = f(x)$.

Observamos que, partiendo del hecho de que la derivada de una función constante es igual a 0, obtenemos que la primitiva de una función no es única y, si F es una primitiva de f , entonces también $F + a$ lo es para cada $a \in \mathbb{R}$.

El teorema fundamental del cálculo dice que la integral indefinida de una función es una de sus primitivas:

$$\text{Si } F(x) = \int_a^x f(t)dt, \text{ entonces } F'(x) = f(x)$$

Además, nos permite calcular las integrales definidas con la siguiente fórmula:

$$\int_a^b f(x)dx = F(b) - F(a), \text{ donde } F \text{ es una primitiva de } f$$



Ejemplo

Una primitiva de la función $g(x) = x^2$ es $G(x) = \frac{x^3}{3}$. Es fácil de comprobar que:

$$G'(x) = \left(\frac{x^3}{3}\right)' = 3\frac{x^2}{3} = x^2 = f(x)$$

Así que ahora podemos calcular cualquier integral del tipo $\int_a^b f(x)dx$:

$$\int_0^2 x^2 dx = G(2) - G(0) = \frac{8}{3} - 0 = \frac{8}{3}$$

$$\int_{-1}^2 x^2 dx = \left(\frac{x^3}{3}\right)_{-1}^2 = \frac{8}{3} - \left(-\frac{1}{3}\right) = \frac{9}{3} = 3$$

2.5.3. Propiedades de la integral y algunas primitivas de funciones básicas

En este apartado veremos las principales propiedades de las integrales, además de algunas primitivas básicas para las funciones más usuales.

La primera propiedad es que la integral es una operación lineal:

$$\int f(x) + g(x) dx = \int f(x) dx + \int g(x) dx$$

$$\int af(x) dx = a \int f(x) dx \text{ para cada } a \in R$$

No hemos especificado los intervalos de integración porque esta propiedad es válida para cualquier conjunto de integración y también para la integral indefinida.

La integral es también lineal en el conjunto de integración:

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx \text{ para cada } a \leq c \leq b$$

Y, en particular, para cada función f :

$$\int_a^a f(x) dx = 0$$

Es posible también definir la integral para intervalos invertidos mediante la siguiente fórmula:

$$\int_a^b f(x) dx = - \int_b^a f(x) dx$$

Muchas veces es importante acotar el valor de una integral, y para eso existen varias desigualdades que se pueden emplear. Las más básicas son las siguientes:

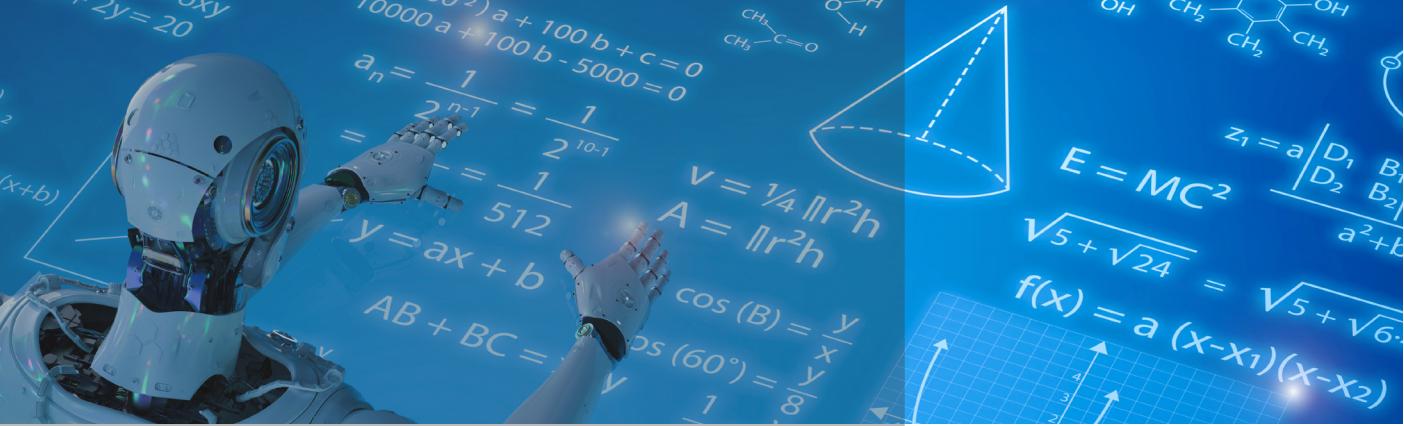
- Si $f(x) \leq g(x)$ para cada $x \in [a, b]$, entonces $\int_a^b f(x) dx \leq \int_a^b g(x) dx$

- En particular, $\int_a^b f(x)dx \leq \int_a^b |f(x)|dx$
- Si $m = \min_{x \in [a,b]} f(x)$ y $M = \max_{x \in [a,b]} f(x)$, entonces $m(b-a) \leq \int_a^b f(x)dx \leq M(b-a)$

Para calcular las primitivas de las funciones básicas, es suficiente invertir las tablas de las derivadas comunes.

- Monomios: si $f(x) = x^k$, entonces una primitiva es $F(x) = \frac{x^{k+1}}{k+1}$
- Función exponencial: si $f(x) = e^x$, entonces $F(x) = e^x$
- Si $f(x) = \frac{1}{x}$, entonces $F(x) = \log(x)$ para $x > 0$
- ...

Derivar una función es una operación simple y, si se conocen las reglas fundamentales, se reduce a un cálculo mecánico. Calcular una integral hallando la primitiva de la función es, en general, más difícil. Existen varias técnicas y trucos, pero también existen casos en los que no es posible encontrar una primitiva de la función expresa como combinación de funciones básicas, es decir, en fórmula cerrada. En muchos casos, es entonces necesario utilizar técnicas de aproximación numéricicas para calcular la integral de la función. Las técnicas más simples se basan en la suma de Riemann y algunas mejorías, y otras técnicas se basan en la probabilidad (método Montecarlo).



Capítulo 3

Probabilidad y estadística

La probabilidad y la estadística se pueden considerar el pilar matemático más importante de la inteligencia artificial y sobre todo del aprendizaje automático.

En los problemas de reconocimiento de patrones, los datos de entrenamiento se pueden ver como realizaciones de variables aleatorias bajo una distribución desconocida con el objetivo de tomar algunas decisiones o identificar la distribución que genera los datos. En todos los problemas relacionados con variables que se miden a través de sensores u observaciones, es importante considerar el ruido que estas mediciones llevan, por lo que es necesario un planteamiento probabilístico de los datos.

Además, en muchos casos es necesario considerar también la aleatoriedad en las respuestas o decisiones que los sistemas automáticos tienen que tomar. Un *bot*, o mejor dicho *chatterbot*, es un sistema capaz de conversar de forma similar a un ser humano.



Enlace de interés

Breve artículo breve de Sarah Mitroff sobre *chatbots* con algunos ejemplos y enlaces a productos en que se utilizan. Recuperado de "What is a bot? Here's everything you need to know", por Sarah Mitroff, 2016, CNET.

<https://www.cnet.com/how-to/what-is-a-bot>

Estos bots no pueden contestar de forma determinista a las conversaciones, sino que se necesita un cierto grado de creatividad y originalidad en las conversaciones. Una de las formas más sencillas y simples de introducir algo que se parezca a la originalidad humana en las máquinas es utilizar la aleatoriedad, es decir, algún tipo de respuesta probabilística.

Los conceptos probabilísticos son también necesarios para formalizar matemáticamente algunos problemas de la inteligencia artificial y para encontrar soluciones óptimas. En particular, todos los métodos estadísticos pueden ser muy útiles para implementar sistemas de inteligencia artificial.

En este capítulo veremos algunos de los conceptos básicos de la probabilidad y de la estadística para ser capaces de entender y aplicar modelos y técnicas en la inteligencia artificial.

3.1. Probabilidad básica

En este primero apartado introduciremos los principales conceptos de la **probabilidad**. El concepto matemático de probabilidad se basa (de hecho, es un caso particular) del concepto de medida y de espacio de medida. En estas notas presentaremos los conceptos de probabilidad, variables aleatorias, etc., de forma intuitiva y sin tener que utilizar los conceptos abstractos de medida. De este modo, la exposición será más sencilla, aunque a veces algo imprecisa.

3.1.1. Algunos ejemplos

Consideremos el lanzamiento de una moneda equilibrada. Es adecuado asumir que, al lanzar muchas veces la moneda, la mitad de las veces saldrá cara y la otra mitad cruz. Además, si tuviésemos que apostar entre cara y cruz, nos daría igual qué opción elegir (sabiendo que la moneda es equilibrada).

Imaginemos ahora que tenemos que asignar un valor de certidumbre antes de lanzar la moneda para comunicar nuestra confianza en el evento "Saldrá cara". Supongamos que este valor tenga que ser entre 0 y 1, siendo un valor cercano a 0 sinónimo de baja probabilidad de suceso y un valor cercano a 1 sinónimo de alta probabilidad. Es natural asignar el número 0,5 a ese valor, dado que no podemos decir si "Saldrá cara" es más o menos probable.

En matemáticas, este valor numérico se llama probabilidad y se suele indicar con la letra P (más adelante veremos que la notación puede variar).

Así pues, en el ejemplo de la moneda equilibrada, sabemos que:

$$P(\text{cruz}) = 0,5 \quad P(\text{cara}) = 0,5$$

Consideremos ahora el clásico ejemplo de la urna y las bolas de dos colores. Imaginamos que hay una urna llena de 100 bolas. Sabemos con seguridad que 30 son blancas y 70 negras, y extraemos una bola al azar (las bolas están bien mezcladas). Es razonable decir que la probabilidad de que la bola extracta sea blanca es 0,3.

3.1.2. Espacio muestral y probabilidad

Ahora intentaremos definir algo más formalmente qué es la probabilidad. Empezamos definiendo el **espacio muestral** como el conjunto de todos los posibles resultados de un experimento aleatorio. El espacio muestral se suele indicar con la letra S o la letra griega Ω .

En el caso del lanzamiento de una moneda, el espacio muestral es $\Omega = \{\text{cara, cruz}\}$; en el caso del lanzamiento de un dado con seis caras, $\Omega = \{1, 2, 3, 4, 5, 6\}$ y, en el caso de la urna del ejemplo del apartado “3.1.1. Algunos ejemplos”, $\Omega = \{\text{blanca, negra}\}$. Cada elemento del espacio muestral se llama **suceso elemental**, y cada subconjunto del espacio muestral se llama **suceso o evento**.



Ejemplo

Sea $\Omega = \{1, 2, 3, 4, 5, 6\}$ el espacio muestral asociado al lanzamiento de un dado. Consideremos el evento $\Omega \supseteq E = \{\text{Ha salido un número impar}\} = \{1, 3, 5\}$, donde E es un evento del espacio muestral compuesto por los tres sucesos elementales: 1, 3 y 5.



Ejemplo

En un ejemplo algo más complejo, consideremos el lanzamiento de dos dados de seis caras. En este caso, el espacio muestral es igual a todas las posibles parejas (i, j) , $1 \leq i, j \leq 6$:

$$\Omega = \{(1, 1), (1, 2), (1, 3), \dots, (2, 1), (2, 2), \dots, (3, 1), \dots, (6, 1), \dots, (6, 6)\}$$

Consideremos ahora el evento $E = \{\text{Salen dos números iguales}\}$. Este evento está compuesto por los seis sucesos elementales:

$$E = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$$

Los eventos de un espacio muestral son subconjuntos, por lo que podemos utilizar las operaciones entre conjuntos para generar otros eventos.

Así pues, si tenemos dos eventos $E, F \subseteq \Omega$, podemos definir el evento unión $E \cup F$, que incluye todos los sucesos elementales que pertenecen a F o a E ; el evento intersección $E \cap F$, que incluye los sucesos que pertenecen a ambos conjuntos, y el evento complementario F^c (o \bar{F} o \tilde{F}), que está formado por todos los eventos que no pertenecen al conjunto inicial F . Además, siempre existe el conjunto vacío \emptyset , que no contiene ningún suceso elemental. Si dos eventos E_1 y E_2 tienen intersección vacía ($E_1 \cap E_2 = \emptyset$), entonces se dice que son dos **eventos mutuamente excluyentes**. Los eventos de un conjunto son mutuamente excluyentes si cada pareja de eventos en el conjunto es mutuamente excluyente. Es decir E_1, E_2, \dots, E_n son mutuamente excluyentes si y solo si $E_i \cap E_j = \emptyset$ para cada $i \neq j, 1 \leq i, j \leq n$.

Una probabilidad P definida sobre el espacio muestral Ω es una función definida sobre los eventos (conjuntos de Ω) con las siguientes propiedades:

- $0 \leq P(E) \leq 1$ para cada evento $E \subseteq \Omega$
- $P(\Omega) = 1$
- Para cada pareja de eventos $E_1, E_2 \subseteq \Omega$ tales que $E_1 \cap E_2 = \emptyset$, resulta que $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

A partir de las propiedades fundamentales (o axiomas) de la probabilidad, se demuestran otras fórmulas útiles:

- $P(E^c) = 1 - P(E)$
- Si $E_1 \subseteq E_2$, entonces $P(E_1) \leq P(E_2)$
- $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$
- $P(\emptyset) = 0$

Observamos ahora que la teoría de la probabilidad es el conjunto de teoremas, resultados y fórmulas que nos permiten calcular las probabilidades de los eventos (y otras cantidades) a partir de las probabilidades de los sucesos elementales. De todos modos, la teoría de la probabilidad no nos permite asignar la probabilidad de los sucesos elementales. Estas probabilidades básicas de cada ejemplo específico son cantidades subjetivas o calculadas a partir de algunas consideraciones físicas o lógicas.

La pareja formada por un espacio muestral y una probabilidad sobre este se llama **espacio de probabilidad**.



Ejemplo

Volvamos al ejemplo del lanzamiento de un dado con seis caras y calculemos la probabilidad del evento $E = \{\text{Ha salido un número impar}\}$. El evento es la unión de los 3 sucesos elementales $\{1\}, \{3\}, \{5\}$. Si asumimos que cada suceso elemental tiene la misma probabilidad, obtenemos que:

$$P(1) = P(2) = \dots = P(6) = \frac{1}{6}$$

Dado que los sucesos elementales son eventos disjuntos entre sí, podemos calcular la probabilidad del evento E :

$$P(E) = P(\{1\} \cup \{3\} \cup \{5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

3.1.3. Eventos independientes

El concepto de independencia es uno de los conceptos más importantes en la teoría de la probabilidad, pues nos permite calcular muchas probabilidades de eventos complejos y es fundamental en el desarrollo de la teoría estadística. En este apartado empezamos a definir el concepto de **eventos independientes**.

Se dice que dos eventos E y F son independientes si:

$$P(A \cap B) = P(A)P(B)$$

En palabras, la probabilidad de la intersección de dos eventos independientes es igual al producto de las probabilidades de los dos eventos. A la práctica, la independencia de dos eventos se suele asumir a partir de algunas reflexiones de tipo lógico.



Ejemplo

Supongamos que lanzamos dos veces una moneda perfectamente equilibrada y anotamos en secuencia los resultados de los dos lanzamientos. El espacio muestral es entonces:

$$\Omega = \{(cara, cara), (cara, cruz), (cruz, cara), (cruz, cruz)\}$$

Es sensato decir que los 4 sucesos elementales del espacio Ω tienen todos las mismas probabilidades igual a $\frac{1}{4}$. Es fácil ver ahora que los eventos $E_1 = \{\text{cara en el 1.º lanzamiento}\}$ y $E_2 = \{\text{cara en el 2.º lanzamiento}\}$ son dos eventos independientes en este espacio de probabilidad:

$$P(E_1) = P(\{(cara, cara), (cara, cruz)\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(E_2) = P(\{(cara, cara), (cruz, cara)\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Y además:

$$P(E_1 \cap E_2) = P(\text{cara, cara}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(E_1)P(E_2)$$

Así pues, los dos eventos son independientes.

También podemos ver este experimento aleatorio como la realización de dos experimentos aleatorios con espacios muestrales $\Omega_1 = \Omega_2 = \{\text{cara, cruz}\}$ y probabilidades de los sucesos elementales iguales a $1/2$, es decir, el lanzamiento de una moneda equilibrada. Entonces podemos asumir que los dos lanzamientos son dos eventos físicos sin ninguna relación entre ellos (es decir, el suceso del primer lanzamiento no influye en el segundo), de modo que los eventos E_1 y E_2 son independientes entre sí y podemos calcular directamente:

$$P(\text{cara, cara}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

El concepto de independencia formaliza entonces lo que intuitivamente entendemos al decir que la realización del evento E no influye sobre la realización del evento F . Es importante recordar que la independencia de dos eventos es una asunción en el problema específico y precisamente en el espacio de probabilidad que hemos definido.

3.1.4. Probabilidad condicionada y teorema de Bayes

Otro concepto clave de la probabilidad es la **probabilidad condicionada**, que representa la probabilidad de un evento cuando tenemos la información de que otro evento dado se ha verificado. Formalmente, la probabilidad condicionada de E dado F se define de la siguiente forma:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Esta probabilidad es bien definida para cada evento F de probabilidad distinta de 0 ($P(F) \neq 0$) y define otra probabilidad sobre el mismo espacio muestral: $Q(E) = P(E|F)$.

El teorema de Bayes enlaza las dos probabilidades condicionadas $P(E|F)$ y $P(F|E)$:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

A partir del teorema de Bayes podemos demostrar también la fórmula que se llama regla de la cadena:

$$P(E \cap F) = P(E|F)P(F)$$



Ejemplo

Consideremos el lanzamiento de dos dados. Como hemos visto, en este caso, el espacio muestral coincide con el conjunto de las 36 parejas de valores:

$$\Omega = \{(1, 1), (1, 2), \dots, (1, 6), (2, 1), \dots, (2, 6), \dots, (6, 1), \dots, (6, 6)\} = \{(i, j) : 1 \leq i, j \leq 6\}$$

Consideremos los dados perfectamente equilibrados, de modo que para cada pareja de valores posibles $(i, j) \in \Omega$:

$$P(i, j) = 1/36$$

Consideremos ahora los eventos $E = \{(i, j) : i = j\}$ y $F = \{i \text{ es par y } j \text{ es par}\}$ y calculemos $P(E|F)$. Intuitivamente, si sabemos que el evento F se ha realizado, entonces sabemos que i y j son ambos números pares. Así pues, enumerando los sucesos elementales en F , obtenemos:

$$F = \{(2, 2), (2, 4), (2, 6), (4, 2), (4, 4), (4, 6), (6, 2), (6, 4), (6, 6)\}$$

Dado que 3 de los 9 eventos de F pertenecen a E , obtenemos que la probabilidad condicionada es igual a $P(E|F) = 1/3$. Utilizando directamente la definición para calcular $P(E|F)$, tenemos que calcular el evento intersección:

$$E \cap F = \{(2, 2), (4, 4), (6, 6)\}$$

Y dado que $P(E \cap F) = 1/36 + 1/36 + 1/36 = 1/12$ y $P(F) = 9/36 = 1/4$, obtenemos que:

$$P(E|F) = \frac{1/12}{1/4} = \frac{1}{12} \cdot 4 = \frac{1}{3}$$



Si dos eventos son independientes entre sí, entonces a partir de la definición de probabilidad condicionada obtenemos que:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E)$$

En palabras, la probabilidad condicionada es igual a la probabilidad original si los dos eventos son independientes.

3.1.5. Espacios muestrales infinitos y continuos

Para los espacios de probabilidades con un número infinito de sucesos elementales (o continuos), tenemos que modificar el axioma de la unión en la probabilidad de la siguiente forma:

Si E_1, E_2, \dots son sucesos mutuamente excluyentes, entonces

$$P(E_1 \cup E_2 \cup E_3 \cup \dots) = \sum P(E_i) = E_1 + E_2 + E_3 + \dots$$

Para los espacios muestrales continuos, la teoría es más compleja y se necesitan conocimientos de teoría de la medida para definir los espacios de probabilidad de forma correcta. Intuitivamente, el problema es que en los espacios muestrales continuos no se puede definir la probabilidad para todos los posibles subconjuntos, lo cual sí es posible para los espacios finitos y numerables. Por esta razón, es necesario definir una clase de subconjunto (los eventos o, mejor dicho, los conjuntos medibles) sobre los cuales es posible definir una probabilidad. Hay distintas formas de definir estos conjuntos medibles, una de las cuales (la más utilizada) es a partir de los intervalos $(a, b) \subset R$. A la práctica no se suele encontrar ningún problema, sobre todo si los eventos que se utilizan son todos construidos a partir de intervalos (abiertos o cerrados).



Ejemplo

Consideremos la función positiva $g(x) = x^2$ en el intervalo $[0,1]$. Consideremos también el cuadrado Q con vértices $(0,0)$, $(0,1)$, $(1,1)$, $(1,0)$ y área igual a 1. Observamos que la gráfica de la función g está incluida en el cuadrado Q . Imaginemos ahora que elegimos un punto al azar en el cuadrado Q . ¿Cuál es la probabilidad de que el punto esté por debajo de la gráfica de g ?

El espacio muestral es, en este caso, el conjunto de puntos en el cuadrado:

$$\Omega = \{(x,y) : 0 \leq x, y \leq 1\}$$

Dado un evento $A \subseteq \Omega$, podemos definir la probabilidad igual al área de este evento, es decir, $P(A) = \text{Área}(A)$.

Así pues, la probabilidad de que un punto esté por debajo de la gráfica de una función es igual a la integral de la función:

$$P(y \leq g(x)) = \int_0^1 g(x) dx$$

$$\text{Y en nuestro ejemplo: } P(y \leq x^2) = \int_0^1 x^2 dx = \left(\frac{x^3}{3}\right)_0^1 = \frac{1}{3}.$$

3.2. Variables aleatorias

Una **variable aleatoria** es una función real X definida sobre un espacio muestral Ω , $X : \Omega \rightarrow R$. Para cada realización de un suceso elemental $\omega \in \Omega$, la variable aleatoria asigna un valor real $X(\omega) \in R$. Si una probabilidad es definida sobre el espacio muestral, entonces es posible inducir una probabilidad sobre la imagen $X(\Omega) \subseteq R$ a través de la siguiente fórmula:

$$P_X(A) = P(X^{-1}(A)) = P(\{\omega \in \Omega \text{ tal que } X(\omega) \in A\})$$

Donde $X^{-1}(A)$ es la imagen inversa de A a través de la función X .

Las variables aleatorias son el objeto fundamental de la probabilidad y, sobre todo, de la teoría estadística. El espacio muestral Ω representa, a la práctica, el espacio de los posibles resultados de un experimento que, en general, no podemos observar directamente. Lo que podemos observar es una medición de un valor numérico para cada realización del experimento aleatorio, es decir, una variable aleatoria.

3.2.1. Variables aleatorias discretas

Si el conjunto $X(\Omega)$ es finito o numerable (es decir, infinito pero no continuo), entonces se dice que la variable aleatoria es una **variable aleatoria discreta**. En este caso, la probabilidad inducida P_X se caracteriza a partir de las probabilidades de los sucesos elementales. En particular, denotamos con $P_X(x) = f_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$ los valores de la **función de densidad o función de masa de probabilidad** f_X de la variable aleatoria X . La función de densidad satisface la condición de normalización $\sum f_X(x) = 1$, y la **función de distribución** de la variable aleatoria se define como:

$$F_X(x) = P(X \leq x) = \sum_{t=-\infty}^x f_X(t)$$



Ejemplo

Sea $\Omega = \{\text{población de España}\}$ y $X: \Omega \rightarrow R$ la variable aleatoria que asigna a cada persona el año de nacimiento. Está claro que $X(\Omega)$ es un conjunto finito de valores y X es, por lo tanto, una variable aleatoria discreta (finita).



Ejemplo

Consideremos el siguiente experimento aleatorio: lanzamos una moneda varias veces hasta que se obtiene cara y paramos. Sea X la variable aleatoria que cuenta cuántas veces hemos lanzado la moneda antes de obtener cara. $X(\omega) > 0$, pero en principio no hay ninguna cota superior al número de veces que podemos lanzar la moneda y obtener siempre cruz. El espacio muestral $X(\Omega) = \{0, 1, 2, 3, 4, \dots\}$ es, pues, infinito numerable.

3.2.2. Variables aleatorias continuas

Si el conjunto $X(\Omega)$ es continuo (por ejemplo, $X(\Omega) = (a, b)$ o $X(\Omega) = R$), entonces la variable aleatoria X es una **variable aleatoria continua**. Definimos de forma similar la función de distribución:

$$F_X(x) = P(X \leq x)$$

Decimos que la variable aleatoria admite una densidad si existe una función no negativa e integrable f_X tal que:

$$P(X \in A) = \int_A f_X(t) dt$$

Donde $A \subseteq \mathbb{R}$ es un evento (un subconjunto) de los números reales. Así pues, podemos escribir para la función de distribución:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

La función de densidad tiene que satisfacer la condición de normalización:

$$\int_R f_X(t) dt = P(X \in R) = 1$$

Una de las consecuencias de la existencia de una densidad para una variable aleatoria es que la función de distribución es una función continua:

$$P(X < x) = \lim_{t \rightarrow x^-} F_X(t) = F_X(x) = P(X \leq x)$$

Dado que $P(X < x) = P(X \leq x)$, la probabilidad de que la variable aleatoria tome un valor $x \in \mathbb{R}$ es igual a 0:

$$P(X = x) = 0 \text{ para cada } x \in \mathbb{R}$$

Esta propiedad es muy importante e indica que, para una variable aleatoria continua (con densidad), la probabilidad de tomar exactamente un valor es siempre igual a 0. Claramente, lo mismo vale para la probabilidad de tomar un conjunto finito o numerable de valores.

3.2.3. Vectores aleatorios

Un vector de variables aleatorias es un conjunto ordenado de variables aleatorias:

$$X = (X_1, X_2, X_3, \dots, X_n)$$

Donde cada $X_i : \Omega \rightarrow \mathbb{R}$ es una variable aleatoria y, para cada $\omega \in \Omega$, el vector $X(\omega) \in \mathbb{R}^n$ es un vector de dimensión n .

Para cada vector aleatorio X , podemos definir la función de distribución:

$$F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

Esto es así del mismo modo que, para las variables aleatorias simples, podemos diferenciar entre vectores aleatorios continuos o discretos, pero además podemos considerar vectores con algunos componentes discretos y algunas continuos.

En general, para un vector aleatorio con densidad, podemos escribir:

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} f_X(t_1, t_2, \dots, t_n) dt_1 dt_2, \dots, dt_n$$

Si la variable X_i es discreta, leemos la integral correspondiente como suma sobre los posibles valores e interpretamos la densidad como función de masa de probabilidad.

La densidad de todos los componentes de un vector se llama densidad conjunta.

Para cada vector aleatorio con una densidad f_X , es posible obtener la densidad de algunos de los componentes a través de la marginalización. Si $X = (X_1, \dots, X_n)$ es un vector aleatorio con densidad $f_X(x_1, \dots, x_n)$, entonces podemos integrar algunas de las variables y obtener la densidad de las restantes variables aleatorias. Por ejemplo, en el caso de un vector aleatorio de dos componentes $X = (X_1, X_2)$ con función de densidad conjunta $f_X(x_1, x_2)$, obtenemos:

$$f_{X_1}(x_1) = \int_R f(x_1, t_2) dt_2$$

$$f_{X_2}(x_2) = \int_R f(t_1, x_2) dt_1$$

Como siempre, leemos las integrales como sumas si las variables son discretas. Las funciones cuya densidad se ha obtenido de tal forma se llaman funciones de densidad marginal.

Dos variables aleatorias son independientes si la función de densidad conjunta se puede factorizar en el producto de las funciones de densidad marginal:

$$f_{(X_1, X_2)}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

En general, las variables aleatorias de un conjunto son independientes si la función de densidad conjunta se puede factorizar:

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

Está claro que, si dos variables son independientes entre sí, entonces los eventos relativos a las dos variables son eventos independientes.

La distribución condicionada de una variable aleatoria X respecto a otra variable aleatoria Y es en el caso discreto:

$$F_{X|Y}(x|y) = P(X \leq x | Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}$$

Para la densidad condicionada, obtenemos la fórmula:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Para cada valor de la variable condicionante $Y = y$, obtenemos entonces otra variable aleatoria $X|Y=y$, cuya distribución se obtiene condicionando los eventos para X al evento $Y=y$. Claramente, si las dos variables son independientes, la probabilidad condicionada es igual a la probabilidad marginal.

3.2.4. Valor esperado y varianza

El **valor medio** o **valor esperado** (llamado también **esperanza** o, simplemente, **media**) de una variable aleatoria es el valor:

$$E[X] = \begin{cases} \sum xP(X = x) & \text{si } X \text{ es una variable aleatoria discreta} \\ \int xf_X(x)dx & \text{si } X \text{ es una variable aleatoria continua} \end{cases}$$

El valor esperado de una variable aleatoria no siempre está bien definido y no siempre existe. Las condiciones bajo las cuales es posible definir el valor absoluto necesitarán una definición más rigurosa de variable aleatoria y de integración que la que utilizamos en estas notas.



El valor esperado de una variable es una operación lineal:

$$E[X + Y] = E[X] + E[Y] \text{ para cada } X \text{ e } Y \text{ variable aleatoria}$$

$$E[aX] = aE[X] \text{ para cada } a \in R \text{ y } X \text{ variable aleatoria}$$

Además, si dos variables aleatorias son independientes, entonces el valor esperado del producto es igual al producto de los valores esperados:

$$E[XY] = E[X]E[Y] \text{ para cada pareja de variables aleatorias } X, Y \text{ independientes}$$

De forma similar, podemos definir el valor esperado de las funciones de variables aleatorias:

$$E[g(X)] = \begin{cases} \sum g(x)P(X=x) & \text{si } X \text{ es una variable aleatoria discreta} \\ \int g(x)f_X(x)dx & \text{si } X \text{ es una variable aleatoria continua} \end{cases}$$

En particular, nos interesan las funciones del tipo $g(x) = x^n$. Para $n \geq 1$ números naturales, llamamos el valor $E[X^n]$ el **momento de orden** n .

Otra cantidad de interés es la **varianza** de una variable aleatoria, que se define de la siguiente forma:

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

La varianza existe y se puede calcular solo si existe el segundo momento de la variable, en cuyo caso es siempre positiva, y es 0 solo si la variable aleatoria es constante. Además, valen las siguientes propiedades:

- $\text{Var}(aX) = a^2\text{Var}(X)$ para cada $a \in R$.
- $\text{Var}(X + a) = \text{Var}(X)$ para cada $a \in R$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ para cada X, Y independientes

La varianza de una variable aleatoria es un valor que indica la dispersión alrededor del valor medio. Los valores pequeños de la varianza indican una variable aleatoria muy concentrada, es decir, que la probabilidad de que la variable tome valores cercanos a su media es muy alta. Los valores grandes de la varianza indican una alta dispersión de la variable aleatoria. Este concepto se formaliza en la desigualdad de Chebyshev:

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

La raíz cuadrada de la varianza se llama **desviación estándar** y se suele denotar con la letra griega σ :

$$\sigma_X = \sqrt{\text{Var}(X)}$$



Ejemplo

Sea B una variable aleatoria binaria que toma dos valores $\{-a, a\}$ con la misma probabilidad (es decir, $P(X = -a) = P(X = a) = 0,5$). Calculamos la varianza de B en función del valor de $a \geq 0$.

$$\text{Var}(B) = E[B^2] - E[B]^2 = \sum_{b \in \{-a, a\}} \frac{1}{2} b^2 - \left(\sum_{b \in \{-a, a\}} \frac{1}{2} b \right)^2 = \frac{1}{2} (a^2 + a^2) - \left(\frac{1}{2} (-a + a) \right)^2 = a^2$$

Cuando el valor de a aumenta y claramente la variable aleatoria está más alejada de su media ($E[B] = 0$) es cuando más aumenta la varianza.

Si una variable aleatoria X es tal que $E[X^2] < \infty$ y es posible definir la varianza, entonces es posible estandarizar (o normalizar o tipificar) la variable aleatoria, es decir, definir otra variable Z , que es la **versión estandarizada** (o normalizada o tipificada) de la variable X :

$$Z = \frac{X - E[X]}{\sigma_X}$$



La variable Z se obtiene trasladando X de tal forma que la media sea 0 ($E[Z] = E[X] - E[X] = 0$) y dividiendo por la desviación estándar, de tal forma que:

$$\text{Var}(Z) = 1$$



Ejemplo

Sea U una variable aleatoria continua uniforme entre 0 y 3, es decir, una variable aleatoria con densidad:

$$f_U(t) = \begin{cases} 0 & \text{para } t < 0 \\ \frac{1}{3} & \text{para } 0 \leq t \leq 3 \\ 0 & \text{para } t > 3 \end{cases}$$

Su media es:

$$E[U] = \int_R t f_U(t) dt = \int_0^3 \frac{t}{3} dt = \frac{1}{3} \left(\frac{t^2}{2} \right) \Big|_0^3 = \frac{1}{6} (9 - 0) = \frac{3}{2}$$

Como esperábamos por intuición, el momento de orden 2 es igual a:

$$E[U^2] = \int_0^3 t^2 dt = \frac{1}{3} \left(\frac{t^3}{3} \right) \Big|_0^3 = \frac{1}{9} (27 - 0) = 3$$

La varianza de U es entonces igual a:

$$\text{Var}[U] = E[U^2] - E[U]^2 = 3 - \frac{9}{4} = \frac{3}{4}$$

La versión normalizada de la variable U es entonces igual a $Z = \frac{U - E[U]}{\sqrt{\text{Var}(U)}} = 2 \frac{\left(U - \frac{3}{2} \right)}{\sqrt{3}}$, que

es una variable aleatoria continua y uniforme entre $\frac{-3}{2}$ y $3 \frac{2}{\sqrt{3}} - \frac{3}{2} = 2\sqrt{3} - \frac{3}{2} = \frac{4\sqrt{3}-3}{2}$.

3.2.5. Covarianza y matriz de covarianza

La **covarianza** entre dos variables aleatorias X y Y se define como:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

La covarianza es un valor que representa cómo varían juntas dos variables aleatorias. Si la covarianza es positiva, entonces los valores mayores de una de las dos variables se corresponden, muy probablemente, con valores mayores de la otra variable. Por el contrario, si la covarianza es negativa, los valores mayores de una de las variables se corresponden con valores menores de la otra.



Si las variables son independientes, entonces la covarianza entre ellas es 0, debido a las propiedades del valor esperado. Es importante observar que hay variables aleatorias que no son independientes entre sí y tales que su covarianza es 0, es decir, que es condición suficiente pero no necesaria.

Las siguientes son las propiedades más importantes de la covarianza:

- $\text{Cov}(X, a) = 0$ para cada $a \in R$
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$
- $\text{Cov}(X + a, Y) = \text{Cov}(X, Y)$

La covarianza entre las versiones estándar de las variables aleatorias se llama **correlación** o coeficiente de correlación, que se define directamente como:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right)$$

Consideremos ahora un vector de variables aleatorias $X = (X_1, X_2, \dots, X_n)$ y construyamos la siguiente matriz:

$$\Sigma = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Cov}(X_n, X_n) \end{pmatrix}$$

Σ se llama matriz de covarianza y es una matriz de dimensión $n \times n$ simétrica semidefinida positiva y cuya diagonal es igual a:

$$\text{diag}(\Sigma) = (\text{Var}(X_1), \dots, \text{Var}(X_n))$$

Si, en vez de la covarianza, consideramos el coeficiente de correlación, entonces obtenemos la matriz de correlación.

3.3. Algunas distribuciones

En este apartado veremos algunos ejemplos de distribuciones de variables aleatorias. La muestra no es completa y solo se pretende presentar algunos ejemplos.

La notación usual para indicar que una variable aleatoria sigue una distribución es $X \sim \text{Distribución}$. Por ejemplo, si X es una variable aleatoria uniforme entre los puntos -1 y 1 , entonces escribiremos $X \sim \text{Unif}([-1, +1])$.

3.3.1. Distribuciones discretas

La distribución discreta más básica es seguramente la distribución de Bernulli. Se dice que una variable aleatoria sigue la distribución de Bernulli de parámetro $p \in [0,1]$, y escribiremos $X \sim \text{Bernulli}(p)$ si X toma valores en un conjunto binario (se suele asumir $X \in \{0,1\}$) y:

$$P(X = 0) = q = 1 - p \quad y \quad P(X = 1) = p$$

El valor esperado y la varianza de una variable Bernulli se obtienen con las siguientes fórmulas:

$$E[X] = p \quad \text{Var}(X) = p(1-p) = pq$$

Si ahora consideramos n distintas variables aleatorias independientes y cada una con distribución Bernulli de parámetro p , entonces la distribución de la suma $X = X_1 + X_2 + \dots + X_n$ de las variables aleatorias Bernulli se llama distribución binomial y se denota con $X \sim B(n,p)$. Su distribución de masa (o densidad) se obtiene a través de cálculos de combinatoria y es igual a:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

La distribución binomial modeliza la probabilidad de éxito en k pruebas de las n totales en una secuencia de experimentos aleatorios binarios.



Ejemplo

Supongamos que lanzamos 10 veces una moneda perfectamente equilibrada. ¿Cuál es la probabilidad de obtener exactamente 4 cruces?

Para contestar, tenemos que observar que la variable aleatoria N números de sucesos sigue una distribución $B(10;0,5)$. Así pues:

$$P(N = 4) = \binom{10}{4} (0,5^4)(0,5^6) = 0,205\dots$$

El valor medio y la varianza de una variable aleatoria $X \sim B(n,p)$ son:

$$E[X] = np \quad \text{Var}(X) = np(1-p)$$

Supongamos ahora que efectuamos un número de pruebas muy alto (en el límite $n \rightarrow +\infty$) y consideramos probabilidades de sucesos $p \rightarrow 0$ tales que:

$$\lim np = \lambda$$

La distribución límite de la distribución binomial con estos parámetros se llama distribución de Poisson y se denota con $X \sim \text{Pois}(\lambda)$. La función de probabilidad es:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad k = 0, 1, 2, 3, \dots$$

La distribución de Poisson se utiliza para modelizar eventos raros, es decir, eventos cuya probabilidad de éxito es muy baja. Además, se utiliza la distribución de Poisson para describir el número de fenómenos (de probabilidad constante) que ocurren en un intervalo de tiempo fijo.

El valor medio y la varianza de una variable aleatoria de Poisson son:

$$E[X] = \lambda \quad \text{Var}(X) = \lambda$$



Ejemplo

Supongamos que en la producción de un componente electrónico específico sabemos que el 1 % de las piezas producidas son defectuosas. ¿Cuál es la probabilidad de que 10 de cada 500 piezas producidas sean defectuosas?

Modelizamos el problema a través de la distribución de Poisson. En particular, pongamos $X = \text{número de piezas defectuosas de las 500} \sim \text{Pois}(\lambda)$. El valor de λ se obtiene utilizando el hecho de que $\lambda = E[X] = \text{valor promedio de piezas defectuosas sobre } 500 = 1\% \text{ de } 500 = 5$.

Así pues, la probabilidad de que 10 piezas sean defectuosas es igual a:

$$P(X = 10) = \frac{e^{-5} 5^{10}}{10!} \approx 0,018$$

Otra distribución fundamental es la distribución uniforme sobre un número finito de valores. Para simplificar la notación, asumimos que estos valores son los primeros n números enteros positivos $\{1, 2, 3, 4, \dots, n\}$. Una variable aleatoria sigue la distribución uniforme sobre n valores si $X : \Omega \rightarrow \{1, 2, 3, \dots, n\}$ y su función de probabilidad es:

$$P(X = k) = \frac{1}{n}$$

Escribimos $X \sim \text{Unif}(n)$ o $X \sim \text{Unif}(\{1, \dots, n\})$ si queremos explicitar el conjunto de valores.

3.3.2. Distribuciones continuas

Entre las distribuciones continuas, la más importante es seguramente la distribución normal o distribución gaussiana.

Se dice que una variable aleatoria continua es gaussiana o que sigue la distribución normal o gaussiana, de parámetros la media μ y la varianza σ^2 (o desviación estándar σ) y escribimos $XN(\mu, \sigma^2)$ si toma valores en todos los números reales y su densidad es:

$$f_X(x) = f_{N(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

El valor medio y la varianza son iguales a los respectivos parámetros $E[X] = \mu$ y $\text{Var}(X) = \sigma^2$. La versión estandarizada de una variable aleatoria es una **variable aleatoria normal estándar** y su densidad es:

$$f_{N(0,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Las variables aleatorias normales son empleadas muchísimo en estadística y en aprendizaje automático. Sus propiedades hacen que los cálculos asociados sean muy sencillos y que existan fórmulas cerradas. Además, muchos eventos reales siguen una distribución que se puede modelizar con una distribución normal.

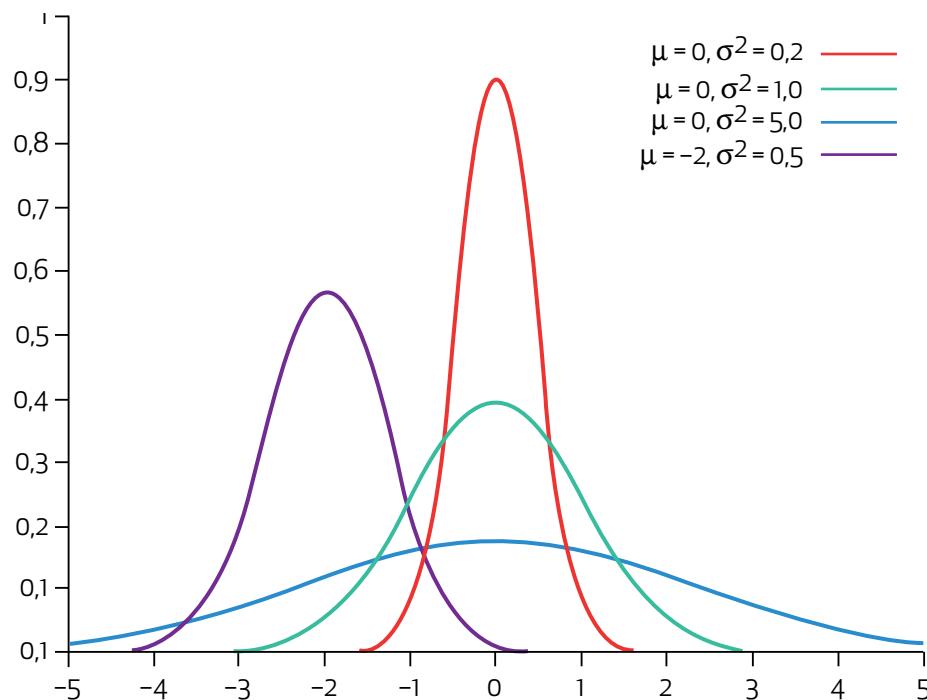


Figura 13. Gráficas de las funciones de densidad gaussianas para varios valores de los parámetros. Por D.328 bajo licencia CC BY-SA 3.0. Disponible en https://commons.wikimedia.org/wiki/File:Normal_distribution_pdf.png



Ejemplo

Supongamos que se quiere modelizar la longitud de una pieza producida en un proceso industrial. Sabemos que las piezas producidas tienen una longitud media de 5 cm y que el error medio de las piezas se mide en $\pm 0,01$ cm. Así pues, podemos modelizar la distribución de la longitud de la pieza con una variable aleatoria $L \sim N(5; 0,001)$.

Los valores de las probabilidades de los eventos de una variable aleatoria gaussiana $X \sim N(\mu, \sigma^2)$ se calculan usando una integral, por ejemplo:

$$P(X \in (a, b)) = \int_a^b f_{N(\mu, \sigma^2)}(x) dx$$

En general, estas integrales, aparte de algunos casos particulares, no se pueden resolver con fórmulas cerradas, sino que es necesario utilizar técnicas de aproximación numérica o tablas de valores.

La función de densidad gaussiana es simétrica alrededor de la media, de modo que:

$$P(X < \mu) = \int_{-\infty}^{\mu} f_{N(\mu, \sigma^2)}(x) dx = \frac{1}{2} = \int_{\mu}^{+\infty} f_{N(\mu, \sigma^2)}(x) dx = P(X > \mu)$$

La distribución gaussiana se puede extender a vectores de variables aleatorias. Un vector aleatorio $X = (X_1, \dots, X_n)$ sigue una distribución gaussiana multivariante si la densidad conjunta es igual a:

$$f_X(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^t \Sigma^{-1} (x - \mu)\right)$$

Donde el vector $\mu \in R^n$ es el vector de los valores medios y Σ es la matriz de covarianza. Las distribuciones marginales de los componentes son gaussianas, y sus parámetros son las medias marginales $E[X_i] = \mu_i$ y la varianza $\text{Var}(\Sigma) = \text{diag}(\Sigma)_i$. Es importante recordar que, si un vector aleatorio es gaussiano, entonces sus componentes siguen distribuciones gaussianas. No es verdad la implicación opuesta, es decir, que un vector aleatorio puede tener componentes normales y no seguir una distribución gaussiana multivariante.

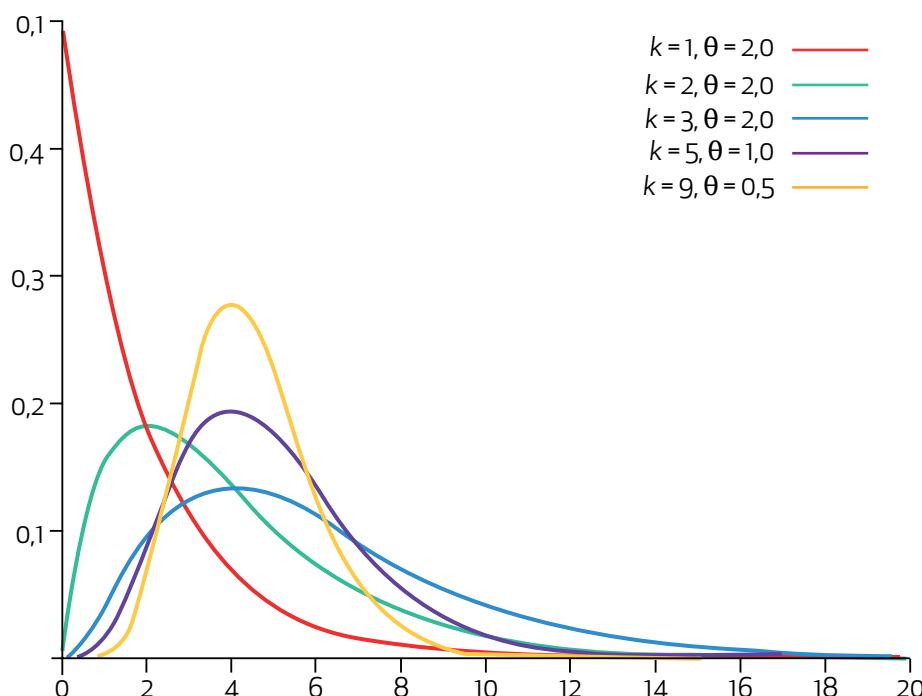


Figura 14. Gráficas de la función de densidad de la distribución gamma para varios valores de los parámetros. Por Autopilot bajo licencia CC BY-SA 3.0. Disponible en https://en.wikipedia.org/wiki/Gamma_distribution#/media/File:Gamma_distribution_pdf.svg

Para un vector gaussiano multivariante, es muy sencillo comprobar si sus componentes son independientes. Es suficiente comprobar que las covarianzas sean igual a 0. Es decir, si $(X, Y) \sim N((\mu_x, \mu_y), \Sigma)$ es un vector gaussiano multivariante, entonces X e Y son independientes si y solo si $\text{Cov}(X, Y) = \Sigma_{1,2} = \Sigma_{2,1} = 0$, es decir, si y solo si la matriz de covarianza es:

$$\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

La distribución gaussiana es simétrica y, a veces, se necesitan distribuciones para variables aleatorias no simétricas. Una de las posibilidades es la distribución gamma, cuya función de densidad es:

$$f_{\text{Gam}(k, \theta)}(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}}$$

Donde $\Gamma(t)$ es la función gamma y los parámetros son k (la forma, *shape* en inglés) y θ (escala). Si $X \sim \text{Gamma}(k, \theta)$, entonces:

$$E[X] = k\theta \quad \text{Var}(X) = k\theta^2$$

Un caso particular de la distribución gamma se da cuando $k=1$ y obtenemos la distribución exponencial.

También para las variables aleatorias continuas, podemos definir la distribución uniforme. En particular, decimos que una variable aleatoria es uniforme entre a y b , y escribimos $X \sim \text{Unif}([a, b])$ si la densidad de X es igual a:

$$f_X(x) = \begin{cases} 0 & x < a \\ \frac{1}{(b-a)} & a \leq x \leq b \\ 0 & x > b \end{cases}$$

Es importante recordar que no es posible definir una distribución uniforme sobre todos los números reales, porque, por definición, la densidad de la distribución uniforme es constante sobre el conjunto de valores y, dado que este conjunto no es acotado, la integral de la densidad no puede ser bien definida (es infinita). Esta simple observación es útil para entender que no tiene sentido decir que elegimos un número al azar con probabilidad uniforme si no decimos los límites mínimo y máximo.

Para los vectores aleatorios, es posible generalizar la distribución uniforme sobre cualquier conjunto. Es decir $X \in \mathbb{R}^n$, es un vector aleatorio con distribución uniforme sobre $A \subseteq \mathbb{R}^n$, un conjunto acotado, si la función de densidad es igual a:

$$f_X(x) = \begin{cases} \frac{1}{\text{Área}(A)} & x \in A \\ 0 & x \notin A \end{cases}$$

3.4. Estimación de parámetros

En este apartado empezamos a ver algunos conceptos de estadística básica. La estadística es la aplicación de la teoría de la probabilidad al estudio de una población a partir de un pequeño conjunto (una muestra).

Uno de los objetivos es estimar la distribución que han generado algunos datos reales para sucesivamente poder efectuar algunas estimaciones o pronósticos. En particular, en este apartado nos centraremos en la estimación de los parámetros de una distribución.

Un concepto fundamental en estadística es el concepto de **variables independientes e idénticamente distribuidas (i.i.d.)**. Se dice que las variables X_1, \dots, X_n aleatorias de un conjunto son independientes e idénticamente distribuidas si $X_i \sim X_j$, es decir, si las variables tienen la misma distribución ($P(X_i < x) = P(X_j < x)$) y son independientes entre sí.



Ejemplo

Supongamos que lanzamos 100 veces un dado equilibrado. Así pues, podemos indicar con X_i la variable aleatoria relacionada con el i -ésimo lanzamiento. Está claro que las variables X_1, \dots, X_n son independiente e idénticamente distribuidas.

El concepto de variables i.i.d. es fundamental para definir el concepto de muestra. Supongamos que observamos un número finito de realizaciones de una variable aleatoria X : $x_1, x_2, x_3, \dots, x_n$. Podemos asumir que cada una de las realizaciones es una variable aleatoria i.i.d. con la misma distribución que X .

Se dice que un conjunto de puntos $x_1, x_2, x_3, \dots, x_n$ es una muestra de X . El objetivo de las técnicas estadísticas es extraer conocimiento, es decir, información, sobre la distribución de X (que suponemos ahora desconocida) a partir de la muestra. A partir de una muestra, es posible calcular algunos valores, los más usuales de los cuales son la media muestral y la matriz de covarianza muestral.

La **media muestral** de la muestra $x_1, x_2, x_3, \dots, x_n$ es la media aritmética:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

La fórmula es válida también para los vectores de variables aleatorias. En este caso, la media muestral es también un vector de la misma dimensión.

En general, es posible definir el momento muestral de orden k :

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

Está claro que el momento muestral de orden 1 es la media muestral.

La **matriz de covarianza muestral** se calcula a partir de una muestra $x_1, x_2, x_3, \dots, x_n$ de vectores, es decir, $x_i \in R^d$. La matriz de covarianza muestral se calcula, pues, usando la siguiente fórmula:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^t$$

El denominador es $(n-1)$ en vez de n por algunas razones teóricas que no abordaremos en estas notas. Merece solo la pena recordar que, por ejemplo, para la varianza existen las dos formas de calcular la varianza muestral:

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

3.4.1. Límites de variables aleatorias

En este apartado veremos dos resultados fundamentales de la estadística y la probabilidad. Ambos resultados estudian cómo la media muestral converge al valor medio de una variable aleatoria.

El primer resultado es la ley de los grandes números, que intuitivamente dice que la media muestral converge, cuando el tamaño de la muestra crece, al valor medio de la variable aleatoria. En particular, si X_1, X_2, \dots es una sucesión de variables aleatorias independientes con el mismo valor esperado μ y la misma varianza σ^2 (cuando las variables son i.i.d., satisfacen estas condiciones) si denotamos con $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n (X_i)$ el promedio de las primeras n variables, entonces:

$$\lim_{n \rightarrow \infty} P(|\overline{X}_n - \mu| < \epsilon) = 1 \text{ para cada } \epsilon > 0$$

Es decir, \overline{X}_n converge en probabilidad a μ .

El segundo resultado, y probablemente el resultado fundamental de la estadística, es el teorema del límite central. Consideremos la versión estandarizada de la suma de variables aleatorias i.i.d.:

$$S_n = X_1 + X_2 + \dots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Donde μ, σ son la media y la desviación estándar de las variables aleatorias X_i . Así pues:

$$\lim_{n \rightarrow \infty} P(Z_n < t) = P(N(0,1) < t)$$

Es decir, el límite de la suma estandarizada de variables aleatorias i.i.d. se distribuye como una normal estándar. Podemos también expresar el resultado en función de la media muestral \overline{X} , en particular si \overline{X}_n es la media muestral de las primeras n observaciones, de modo que \overline{X}_n tiene aproximadamente (para n suficientemente grande) una distribución gaussiana $N(\mu, \frac{\sigma^2}{n})$.

3.4.2. Método de los momentos

El método de los momentos para estimar los parámetros de una distribución se basa en resolver las ecuaciones obtenidas igualando los momentos muestrales con los momentos teóricos. En algunos de los ejemplos precedentes, ya hemos utilizado de alguna forma esta técnica para calcular los parámetros de nuestros modelos.

Si el modelo de probabilidad se especifica a partir de k parámetros, entonces el método de los momentos equivale a resolver las siguientes ecuaciones en los parámetros:

$$\begin{cases} \mu_1 = E[X] \\ \mu_2 = E[X^2] \\ \vdots \\ \mu_k = E[X^k] \end{cases}$$

Donde μ_i es el momento muestral de orden i .



Ejemplo

Supongamos que obtenemos la muestra 1; 3,2; 2,5; 3,7; 1,5; 4 y supongamos que nuestro modelo es una variable aleatoria uniforme $X \sim \text{Unif}([0, \theta])$. ¿Cuál es el estimador del método de los momentos de θ ?

Para contestar es suficiente calcular la media muestral \bar{x} y el valor medio de la variable $E[X]$, y resolver la ecuación $E[X] = \bar{x}$.

$$\bar{x} = \frac{(1 + 3,2 + 2,5 + 3,7 + 1,5 + 4)}{6} = 2,65$$

$$E[X] = \frac{(\theta - 0)}{2} = \frac{\theta}{2}$$

Así pues, tenemos que resolver $\frac{\theta}{2} = 2,65$, que nos proporciona la estimación:

$$\hat{\theta} = 5,3$$



Ejemplo

Sea $YN(\mu, \sigma^2)$ una variable gaussiana con media y varianza desconocidas, y supongamos que obtenemos una muestra i.i.d. tal que la media muestral es igual a 5 y la varianza muestral es igual a 2. Podemos, entonces, aplicar el método de los momentos para encontrar los estimadores $\hat{\mu}$ y $\hat{\sigma}$. En particular:

$$E[X] = \hat{\mu} = \mu_1 = 5$$

$$E[X^2] = \hat{\sigma}^2 + E[X]^2 = \mu_2 = s_n^2 + \mu_1^2$$

Resolviendo el sistema de dos ecuaciones en las incógnitas $\hat{\mu}, \hat{\sigma}$ obtenemos:

$$\hat{\mu} = 5\hat{\sigma}^2 = s_n^2 = 2$$

3.4.3. Método de máxima verosimilitud

El método de los momentos es, en general, bastante sencillo de aplicar, pero no tiene buenas propiedades teóricas, por lo que se suele utilizar el **método de máxima verosimilitud** para estimar los parámetros de una distribución.

La **verosimilitud** de un modelo, dada una muestra de observaciones, es:

$$L(\theta | x_1, x_2, \dots, x_n) = f_X(x_1 | \theta) f_X(x_2 | \theta) \dots f_X(x_n | \theta) = \prod_{i=1}^n f_X(x_i | \theta)$$

Donde $f_X(x|\theta)$ es la función de densidad (o de probabilidad) de la variable aleatoria en función de los parámetros θ .

El estimador de máxima verosimilitud (MLE) $\hat{\theta}_{MLE}$ de los parámetros del modelo se encuentra maximizando la verosimilitud, es decir, encontrando el punto de máximo de la función $L(\theta|x_1, \dots, x_n)$ vista como función de los parámetros θ .

Intuitivamente, la verosimilitud representa la probabilidad de la secuencia de observaciones x_1, \dots, x_n bajo la hipótesis de que las observaciones son i.i.d. con distribución $f_X(x|\theta)$. Así pues, maximizar la verosimilitud es equivalente a encontrar el modelo (en la familia paramétrica de modelos considerados) que consiga la mayor probabilidad de que se realicen las observaciones.

Encontrar el punto de máximo de la verosimilitud es equivalente a encontrar el punto de máximo de la log-verosimilitud, es decir, la función $LL(\theta|x_1, \dots, x_n)$. La ventaja de la log-verosimilitud es que, por las propiedades del logaritmo, el producto en la definición de la verosimilitud se transforma en una suma, de modo que la log-verosimilitud es, en general, mucho más sencilla de derivar (para encontrar los puntos críticos).

$$LL(\theta|x_1, \dots, x_n) = \sum_{i=1}^n \log(f_X(x_i|\theta))$$



Ejemplo

Supongamos que obtenemos n observaciones x_1, \dots, x_n de una variable gaussiana con media desconocida y varianza igual a 1, es decir, $X \sim N(\mu, 1)$. La log-verosimilitud es entonces igual a:

$$LL(\mu|x_1, \dots, x_n) = \sum_{i=1}^n \log f_{N(\mu, 1)}(x_i)$$

Donde $\log f_{N(\mu, 1)}(x) = \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2}$, de modo que:

$$LL(\mu|x_1, \dots, x_n) = \sum_{i=1}^n \log f_{N(\mu, 1)}(x_i) = n \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$$

Para encontrar el punto de máximo (si existe), podemos derivar y encontrar los puntos críticos:

$$\frac{dLL(\mu|x_1, \dots, x_n)}{d\mu} = 0 - \sum_{i=1}^n (x_i - \mu)$$

Para encontrar los puntos críticos, buscamos ahora donde la derivada es 0:

$$\sum_{i=1}^n (x_i - \mu) = 0$$

El único punto crítico es $\mu = \bar{x}$, que fácilmente vemos que es un punto de máximo (la derivada segunda es siempre negativa). Observamos que el estimador de máxima verosimilitud es igual al estimador por el método de los momentos (aunque no siempre es así).



Ejemplo

Sea $X \sim Bernulli(p)$ una variable aleatoria Bernulli con parámetro desconocido y supongamos que obtenemos una muestra i.i.d. a partir de X .

Para calcular el estimador de máxima verosimilitud \hat{p}_{MLE} , consideraremos la log-verosimilitud:

$$LL(p | x_1, \dots, x_n) = \sum_{i=1}^n \log P(X = x_i) = \sum_{i=1}^n \log(1-p) + \sum_{i=1}^n \log(p)$$

Por tanto, si calculamos la derivada:

$$\frac{dLL(p | x_1, \dots, x_n)}{dp} = \frac{n_0}{1-p} - \frac{n_1}{p}$$

Donde n_0, n_1 son el número de puntos en la muestra con valores 0 y 1, respectivamente, así que el estimador MLE se obtiene resolviendo la ecuación asociada para encontrar los puntos críticos:

$$\hat{p}_{MLE} = \frac{n_1}{n}$$

Glosario



Autovalor

Valor numérico λ asociado a una matriz A por la fórmula $\lambda v = Av$, donde $v \in R^n$ es el autovector correspondiente.

Autovector

Vector $v \in R^n$ correspondiente a un autovalor $\lambda \in R$ tal que $\lambda v = Av$.

Codominio

Conjunto de valores que una función puede asumir. En otras palabras, las salidas de la función correspondiente pueden solo ser elementos del codominio. Es uno de los tres elementos que definen una función.

Correlación

Coeficiente numérico que indica la relación entre dos variables aleatorias. Los valores positivos (o negativos) indican que, al aumentar una de las variables, la otra muy probablemente crece (o decrece). Es un valor entre -1 y 1 , y se computa como la covarianza entre las versiones normalizadas de las dos variables aleatorias.

Covarianza

Valor que, de forma similar a la correlación, indica la relación entre dos variables. Se define como $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

Definida negativa

Una matriz es definida negativa si $x^t Ax < 0$ para todos los vectores $x \in R^n$ distintos del vector de ceros.

Definida positiva

Una matriz es definida positiva si $x^t Ax > 0$ para todos los vectores $x \in R^n$ distintos del vector de ceros.

Derivada de una función

La derivada de una función en un punto es un valor numérico que indica cómo la función cambia en un entorno. Se define a través de un límite y se puede interpretar geométricamente como el coeficiente de la recta tangente a la gráfica de la función.

Derivada direccional

La derivada direccional es la derivada de una función de más de una variable en una dirección precisa. Se calcula con el producto escalar del gradiente por la dirección elegida.

Derivada parcial

La derivada parcial de una función de distintas variables es la derivada respecto a una de las variables fijando las demás.

Desviación estándar

La desviación estándar es una medida de concentración de las variables aleatorias. En particular, se define como la raíz cuadrada de la varianza.

Determinante

El determinante de una matriz cuadrada es un valor numérico esencial en el álgebra lineal. Una propiedad fundamental es que una matriz es invertible solo si su determinante es distinto de 0.

Diagonal de la matriz

La diagonal de una matriz cuadrada $A \in R^n$ es el vector de elementos $a_{1,1}, a_{2,2}, \dots, a_{n,n}$.

Dominio

El dominio de una función es el conjunto de valores de partida de una función.

Espacio de probabilidad

Un espacio de probabilidad es un espacio muestral, es decir, un espacio de eventos, junto con una función de probabilidad definida sobre ellos.

Espacio muestral

Un espacio muestral es un espacio sobre el que se define una probabilidad. Sus subconjuntos son los eventos o sucesos, y sus elementos se llaman sucesos elementales.

Espacio vectorial

Un espacio vectorial es un conjunto de vectores cerrados con respecto a las operaciones de suma y de producto por escalar.

Esperanza

Véase *valor medio*.

Evento

Un evento es un subconjunto de un espacio muestral.

Eventos independientes

Dos eventos en un espacio de probabilidad son independientes si la probabilidad del evento intersección es igual al producto de las probabilidades de los dos eventos.

Eventos mutuamente excluyentes

Dos eventos son mutuamente excluyentes si su intersección es el evento vacío.

Extremos de integración

Los extremos de integración de una integral son los valores $a, b \in R$ tales que la integral se escribe como $\int f(x)dx$.

Función

Una función es una regla que asocia a cada valor en un conjunto de partida (el dominio) un solo elemento en un conjunto de llegada, llamado codominio. Una función se define por tres elementos: la regla, el dominio y el codominio.

Función continua

Una función de variable real es continua si en cada punto de su dominio el límite en el punto existe y es igual al valor de la función en el mismo punto. La gráfica de una función continua es una línea ininterrumpida.

Función de densidad

La función de densidad de una variable aleatoria continua es una función f positiva y tal que $P(X < t) = \int f(x)dx$. Para las variables aleatorias discretas, la función de densidad es igual a la función de masa de probabilidad.

Función de distribución

La función de distribución de una variable aleatoria es la función $F_X(x) = P(X \leq x)$.

Función de masa de probabilidad

La función de masa de probabilidad, o simplemente función de probabilidad, de una variable aleatoria discreta es la función $f_X(x) = P(X = x)$, es decir, es la probabilidad de los sucesos elementales.

Función monótona creciente

Una función es monótona creciente si $f(x_1) < f(x_2)$ para cada $x_1 < x_2$.

Función monótona decreciente

Una función es monótona decreciente si vale $f(x_1) > f(x_2)$ para $x_1 < x_2$.

Gradiente de una función

El gradiente es el vector formado por las derivadas parciales de una función de más de una variable.

Gráfica de función

La gráfica de una función de variables reales es el conjunto de puntos del tipo $(x, f(x))$. La gráfica de una función es una imagen en dos dimensiones para las funciones de una variable y es en tres dimensiones para las funciones de dos variables.

Imagen de la función

La imagen de una función es el conjunto de puntos en el codominio, que son el resultado de computar la función en algún punto del dominio. Es siempre un subconjunto del codominio.

Imagen inversa

La imagen inversa de un subconjunto de puntos del codominio de una función es $f^{-1}(C) = \{x \in A : f(x) \in C\}$. Es decir, es igual al conjunto de puntos cuya salida por la función termina en un subconjunto C del codominio.

Integral de una función

La integral de una función es una operación que se denota con $\int f(x)dx$ e intuitivamente está relacionada con el área que hay debajo de la gráfica de la función.

Integral indefinida

La integral indefinida es la integral de una función en que el extremo de integración superior es variable. La integral indefinida es, pues, una función del extremo de integración de la integral.

Límite de una función

El límite de una función es un concepto matemático que intuitivamente representa el valor de la función considerada cuando la variable se acerca a un punto.

Mapa

Véase *función*.

Matriz

Una matriz es un objeto matemático que se representa como una tabla de valores numéricos con dos dimensiones: la primera dimensión es el número de filas y la segunda el número de columnas.

Matriz cuadrada

Una matriz cuadrada es una matriz con el mismo número de filas y de columnas, es decir, con dimensiones iguales.

Matriz de covarianza muestral

La matriz de covarianza muestral es una estimación de la matriz de covarianza que se calcula a partir de los datos de la muestra.

Matriz identidad

La matriz identidad es una matriz cuadrada compuesta de ceros en todos los elementos excepto en la diagonal, donde los elementos son unos. La matriz identidad se denota con el símbolo I_n .

Matriz ortogonal

Una matriz es ortogonal si su inversa es igual a la matriz traspuesta, es decir, $AA^t = A^tA = I_n$.

Matriz traspuesta

La matriz traspuesta es la matriz que se obtiene cambiando entre sí las filas y las columnas de una matriz.

Matriz simétrica

Una matriz es simétrica si es igual a su traspuesta.

Matriz singular

Una matriz es singular si su determinante es igual a 0. Las matrices singulares no son invertibles.

Media muestral

La media muestral es igual a la media aritmética de los valores de una muestra. Es una estimación del valor medio.

Menor complementario

El menor complementario (i, j) de una matriz A es el determinante de la matriz obtenida a partir de A eliminando la fila i y la columna j .

Menor de una matriz

El menor de una matriz es el determinante de una submatriz obtenida eliminando algunas filas y/o columnas.

Método de máxima verosimilitud

El método de máxima verosimilitud es un método de estimación de parámetros que consiste en encontrar los parámetros que maximizan la función de verosimilitud.

Momento de orden n

El momento de orden n de una variable aleatoria se define como $E[X^n] = \int x^n f_X(x) dx$. En particular, el momento de orden 1 es el valor medio.

Ortogonales

Dos vectores son ortogonales si el producto escalar entre ellos es igual a 0.

Preimagen

Véase *imagen inversa*.

Primitiva de una función

F es una primitiva de la función f si $F' = f$. La primitiva es, pues, la inversa de la derivada. Es importante observar que las primitivas de una función no son únicas.

Probabilidad

La probabilidad es un valor numérico positivo que indica cómo de probable es que un evento se realice.

Probabilidad condicionada

La probabilidad condicionada de un evento E dado otro evento F se define como $P(E|F) = P(E \cap F)/P(F)$ y representa la probabilidad de que se realice el evento E si sabemos que el evento F ya se ha realizado.

Producto escalar

El producto escalar es una operación binaria que asocia a cada pareja de vectores la suma de los productos de sus componentes.

Punto crítico

Un punto crítico de una función es un punto de su dominio donde la derivada es 0.

Punto estacionario

Véase *punto crítico*.

Punto de máximo absoluto

Un punto de máximo absoluto de una función es un punto del dominio de la función tal que el valor de la función es el mayor.

Punto de máximo global

Véase *punto de máximo absoluto*.

Punto de máximo local

Véase *punto de máximo relativo*.

Punto de máximo relativo

Un punto de máximo relativo es un punto del dominio tal que es un punto de máximo en un subconjunto, es decir, un entorno del punto.

Punto de mínimo absoluto

Un punto de mínimo absoluto de una función es un punto del dominio de la función tal que el valor de la función es el menor.

Punto de mínimo global

Véase *punto de mínimo absoluto*.

Punto de mínimo local

Véase *punto de mínimo relativo*.

Punto de mínimo relativo

Un punto de mínimo relativo es un punto del dominio tal que es un punto de mínimo en un subconjunto, es decir, un entorno del punto.

Rango de una matriz

El rango de una matriz es la dimensión del más grande menor distinto de 0.

Sistema compatible

Un sistema de ecuaciones es compatible si admite por lo menos una solución.

Sistema compatible determinado

Un sistema es compatible determinado si tiene una sola solución.

Sistema compatible indeterminado

Un sistema es compatible indeterminado si tiene infinitas soluciones.

Sistema incompatible

Un sistema es incompatible si no admite ninguna solución.

Suceso

Véase *evento*.

Suceso elemental

Un suceso elemental es un elemento de un espacio muestral, es decir, un suceso de cardinalidad 1.

Traza

La traza de una matriz cuadrada es la suma de los elementos en la diagonal.

Triangular inferior

Una matriz es triangular inferior si los elementos que hay por encima de la diagonal son iguales a 0.

Triangular superior

Una matriz es triangular superior si los elementos que hay por debajo de la diagonal son iguales a 0.

Valor esperado

Véase *valor medio*.

Valor medio

El valor medio de una variable aleatoria es $E[X] = \int xf_X(x)dx$.

Valor propio

Véase *autovalor*.

Valores singulares

Los valores singulares de una matriz (en general, no cuadrada) A son los autovalores de la matriz A^tA .

Variable aleatoria

Una variable aleatoria es una función a partir de un espacio de probabilidad en los números reales.

Variable aleatoria continua

Una variable aleatoria es continua si el conjunto imagen es un conjunto continuo.

Variable aleatoria discreta

Una variable aleatoria es discreta si el conjunto imagen es finito o numerable, es decir, si no es continuo.

Variable aleatoria normal estándar

Una variable aleatoria normal estándar es una variable aleatoria continua con distribución gaussiana (o normal) con media 0 y varianza 1.

Variables independientes e idénticamente distribuidas (i.i.d.)

Las variables aleatorias de un conjunto son independientes e idénticamente distribuidas si son todas independientes entre sí y además tienen todas la misma distribución.

Varianza

La varianza de una variable aleatoria se define como $\text{Var}(X) = E[X^2] - E[X]^2$ y es una medida de la dispersión de la variable alrededor de su valor medio.

Vector

Un vector es un objeto matemático con una dimensión n que se representa como una lista ordenada de valores $v = (v_1, v_2, \dots, v_n)$.

Vector columna

Un vector columna es un vector interpretado como una matriz de dimensión $n \times 1$.

Vector fila

Un vector fila es un vector interpretado como una matriz de dimensión $1 \times n$.

Vector propio

Véase *autovector*.

Verosimilitud

La función de verosimilitud es el producto de la densidad de la variable aleatoria en los puntos de la muestra considerando los parámetros de la densidad como variables.

Versión estandarizada

La versión estandarizada de una variable aleatoria es una trasformación de la variable obtenida con una traslación y una escala de tal forma que la nueva variable aleatoria tenga valor medio 0 y varianza 1.

Enlaces de interés



The mathematician who cracked Wall Street | Jim Simons

Charla TED con Jim Simons, profesor de matemáticas y fundador de Renaissance Technologies. Fue uno de los primeros en aplicar técnicas matemáticas y estadísticas en la predicción y la optimización en el mercado bursátil, un ejemplo de la aplicación de las matemáticas. En Renaissance Technologies empezaron a utilizar el aprendizaje automático y los datos masivos (*big data*) para entrenar modelos matemáticos.

<https://www.youtube.com/watch?v=U5kldtMJGc8>

MATHSCINET, Mathematical Reviews

MathSciNet es un repositorio de artículos de matemáticas. Es la mejor elección para encontrar artículos y autores en una fuente oficial. Lo gestiona la American Mathematical Society.

<https://mathscinet.ams.org/mathscinet/>

divulgaMAT

divulgaMAT es el centro virtual de divulgación de las matemáticas de la Real Sociedad Matemática Española.

<http://www.divulgamat.net/>

GeoGebra

Es una herramienta en línea para dibujar gráficas de funciones, realizar construcciones geométricas y resolver ecuaciones. Está disponible en español.

<https://www.geogebra.org/>

Bibliografía



Cohn, P. (1994). *Elements of Linear Algebra* (p. 69). Londres: Taylor & Francis Ltd/CRC Press.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11), 559-572. doi:10.1080/14786440109462720.



Autor
Dr. Gherardo Varando