Tabla de contenido

[1. Vectores. 3](#_Toc145694033)

[1.1 Producto punto y sus operaciones asociadas. 3](#_Toc145694034)

[1.2 Producto interno. 5](#_Toc145694035)

[1.3 Ángulos con el producto interno. 5](#_Toc145694036)

[1.4 Proyección de un vector. 6](#_Toc145694037)

[2. Matrices. 6](#_Toc145694038)

[2.1 ¿Qué es un matriz? 6](#_Toc145694039)

[2.2 Operaciones con matrices 6](#_Toc145694040)

[2.3 El determinante de una matriz. 6](#_Toc145694041)

[2.4 Tipos de matrices. 6](#_Toc145694042)

[2.5 Transposición de una matriz. 6](#_Toc145694043)

[2.6 La convención de suma de Eistein aplicada al producto matricial. 7](#_Toc145694044)

[2.7 Matriz ortogonal: Características. 8](#_Toc145694045)

[2.8 Cambio de base de matrices. 8](#_Toc145694046)

[2.9 El proceso de Gram-Schmidt. 10](#_Toc145694047)

[2.10 Reflejando en un plano. 11](#_Toc145694048)

[3. Vectores característicos (característicos). 11](#_Toc145694049)

[3.1 ¿Qué son los vectores eigen? 11](#_Toc145694050)

[3.2 Casos especiales. 11](#_Toc145694051)

[3.3 Descripción matemática de los eigen vectors y eigen values. 11](#_Toc145694052)

[3.4 Cambio de base de vectores propios o característicos. 12](#_Toc145694053)

[3.5 Un ejemplo práctico: El page rank. 13](#_Toc145694054)

[4. Funciones. 15](#_Toc145694055)

[4.1 ¿Qué son las funciones? 15](#_Toc145694056)

[4.2 Funciones con parámetros. 16](#_Toc145694057)

[5. Derviada o gradiente 17](#_Toc145694058)

[5.1 ¿Qué es una derivada? 17](#_Toc145694059)

[5.2 Reglas de las derivadas. 17](#_Toc145694060)

[5.3 Derivadas inmediatas: 18](#_Toc145694061)

[5.1 ¿Qué es un gradiente? 19](#_Toc145694062)

[5.2 Derivadas parciales. 19](#_Toc145694063)

[5.3 Derivada total 19](#_Toc145694064)

[5.4 Regla de la cadena en funciones de varias variables. 19](#_Toc145694065)

[5.5 Vector Jacobiano. 20](#_Toc145694066)

[5.6 Matriz jacobiana. 20](#_Toc145694067)

[5.7 La matriz hessiana. 20](#_Toc145694068)

[6. Fundamentos matemáticos de redes neuronales. 21](#_Toc145694069)

[6.1 Modelos. 21](#_Toc145694070)

[6.2 Definición de una neurona a nivel matemático. 22](#_Toc145694071)

[6.3 Definición matemática formal de una red neuronal. 23](#_Toc145694072)

[6.4 Definición de una red neuronal a nivel matemático. 24](#_Toc145694073)

[6.5 Coste de la red neuronal (margen de error) 24](#_Toc145694074)

[6.6 Retropropagación en redes neuronales. 25](#_Toc145694075)

[7. Taylor series para aproximaciones. 26](#_Toc145694076)

[7.1 Introducción. 26](#_Toc145694077)

[7.2 Serie 26](#_Toc145694078)

[7.3 Serie de potencias 26](#_Toc145694079)

[7.4 Derivación en series de potencia (Formula de Maclaurin). 28](#_Toc145694080)

[7.5 Series de Taylor. 29](#_Toc145694081)

[7.6 Linealización o aproximación lineal. 29](#_Toc145694082)

[7.7 Series de Taylor con varias variables. 29](#_Toc145694083)

[8. Introducción a la optimización. 31](#_Toc145694084)

[8.1 Método de Newton-Raphson 31](#_Toc145694085)

[8.2 Descenso de gradiente. 32](#_Toc145694086)

[8.3 Optimización con Restricciones 32](#_Toc145694087)

[8.4 Multiplicadores de Lagrange. 33](#_Toc145694088)

[9. Regresiones lineales. 33](#_Toc145694089)

[9.1 Modelo de regresión lineal simple. 33](#_Toc145694090)

[9.2 Modelo de regresión no lineal. 34](#_Toc145694091)

[10. Estadística de los conjuntos de datos. 36](#_Toc145694092)

[10.1 ¿Qué es un conjunto de datos? 36](#_Toc145694093)

[10.2 Propiedades de los conjuntos de datos. 37](#_Toc145694094)

[10.3 Producto punto en el contexto de los conjuntos de datos. 40](#_Toc145694095)

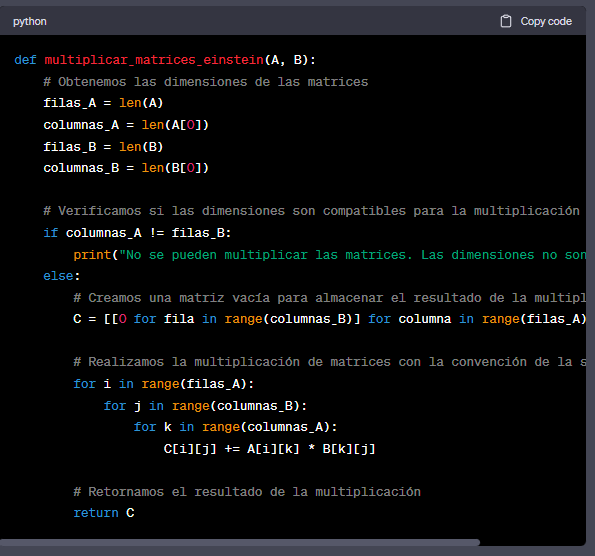
[11. Alta dimensionalidad de la información. 41](#_Toc145694096)

[11.1 Reducción de la dimensionalidad. 41](#_Toc145694097)

## La convención de suma de Eistein aplicada al producto matricial.

En el contexto de la multiplicación de matrices, lo que viene a decir está convención es que dada dos matrices A y B, las cuales se multiplican, en el calculo de una determinada posición “ij” en la matriz resultante viene dada por:

Ayudándonos de está formula, podemos construir de manera sistemática una función que calcule la multiplicación de dos matrices:



Claves:

* Nos permite lidiar con matrices no cuadradas.
* Busca simplificar la formula compleja de: . Por una más sencilla.
* Ya con la librería numpy podemos calcular fácilmente el producto punto de dos matrices.

## Matriz ortogonal: Características.

Las columnas (y también las filas) de una matriz ortogonal forman un conjunto ortonormal. Esto significa que todos los vectores columna (o fila) son perpendiculares entre sí (ortogonales) y tienen magnitud 1 (norma unitaria).

|  |  |
| --- | --- |
| **Propiedad** | **Descripción** |
| **Inversa y transposición** | La inversa de una matriz ortogonal es igual a su matriz transpuesta, es decir, |
| **Determinante** | El determinante de una matriz ortogonal es siempre 1 o -1. |
| **Columnas (y filas) ortonormales** | Las columnas (y filas) de una matriz ortogonal forman un conjunto ortonormal. |
| **Preservación de longitud y ángulos** | Las matrices ortogonales preservan la longitud y el ángulo entre los vectores al aplicar la transformación lineal correspondiente. |
| **Representación de rotaciones y reflexiones** | Las matrices ortogonales pueden representar rotaciones y reflexiones en el espacio euclidiano. |
| **Producto con su transpuesta** | Una matriz ortogonal cumple con la propiedad A^T \* A = A \* A^T = I (matriz identidad). |

Además, debemos considerar que toda matriz diagonal es una matriz ortogonal, es decir que los vectores que la conforman forman una base ortonormal.

**Importante:** En ciencia de datos siempre se están buscando matrices de este tipo.

## Cambio de base de matrices.

### ¿Qué es el cambio de base?

En matemáticas y ciencias aplicadas, a menudo es útil trabajar con diferentes bases para describir un mismo objeto en un espacio vectorial. El cambio de base de matrices es una técnica matemática que nos permite transformar una matriz de coordenadas de un sistema de referencia a otro, es decir, nos permite expresar las mismas coordenadas de un objeto en una base diferente.

Existe una matriz base que define la base del sistema de coordenadas, recordemos que una base es un conjunto de vectores que se utilizan como punto de referencia para describir la posición de un objeto en ese sistema de coordenadas (es como un “marco de referencia”). Cada uno de los vectores que componen la base deben ser linealmente independientes (esto no es sinónimo de que sean perpendiculares entre sí, ya que hay vectores que son linealmente independientes y no son perpendiculares entre sí).

Ejemplo práctico para entenderlo mejor:

Imaginemos que estamos en un avión volando en el aire y de repente recibimos una llamada de la torre de control del aeropuerto más cercano informándonos sobre la posición de una bandada de pájaros que debemos evitar. La torre de control nos proporciona las coordenadas de la posición de los pájaros utilizando un sistema de coordenadas global basado en la Tierra.

En este escenario, podemos pensar en el cambio de base como una herramienta que nos permite convertir las coordenadas de la posición de los pájaros desde el sistema de coordenadas global a un sistema de coordenadas local que esté alineado con nuestro avión. Para realizar esta conversión, necesitamos conocer la matriz de transformación que lleva de un sistema de coordenadas al otro.

Una vez que hemos convertido las coordenadas de los pájaros a nuestro sistema de coordenadas local, podemos determinar su posición relativa con respecto a nuestro avión. Si los pájaros están muy cerca de nuestro avión, podemos utilizar esta información para maniobrar y evitar la colisión.

Aquí vemos la aplicación del campo base para el mundo real.

### El marco de referencia canónico

El marco de referencia canónico es un sistema de coordenadas en el que se utilizan los vectores canónicos como base. Un ejemplo de base cartesiana está definido por los siguientes vectores:

Si, además de esto, se cumple que los ejes del marco los denominamos ejes “x”, “y” y “z” entonces hablamos del marco de referencia “cartesiano”.

### Otros marcos de referencia

Cuando tenemos una base diferente a la canónica, uno de los requisitos que se deben cumplir para que sea una base correcta, es que todos los vectores que la compongan sean linealmente independientes.

Ahora bien, supongamos lo siguiente:

* Tenemos una matriz base A que define un marco de referencia Marco(A) cuyos vectores están definido con respecto del eje de coordenadas canónico.
* Si hacemos la inversa de ese vector A, es decir, calculamos entonces tendremos la base A con respecto del marco de referencia Marco(A)
* Además, tenemos un vector “v” definido en el marco de referencia (A)
* Finalmente, tenemos un vector “w” definido en el marco de referencia (Canónico).

Con estas premisas tenemos que:

* Para transformar el vector “v” referenciado con respecto del marco(A) a un nuevo vector “ v’ ” que tiene como referencia el marco canónico tenemos la siguiente formula:
* Para transformar el vector “w” referenciado con respecto del marco (Canónico) a un nuevo vector w’ con respecto del marco(A) hacemos lo siguiente:

### Marcos de referencia ortogonales.

Vimos en el punto anterior ciertas transformaciones de vectores de un marco de referencia A hacia el marco de referencia canónico, ahora bien, dado un vector v en un marco de referencia B con vectores base ortogonales , y un marco de referencia A con vectores base ortogonales puedes calcular las coordenadas del vector v' en el marco de referencia A mediante proyecciones.

Recordemos que la proyección ortogonal consiste en descomponer un vector dado en términos de sus componentes paralelos y perpendiculares a cada uno de los vectores base ortogonales del marco de referencia

En este sentido, si se tiene un vector v en el marco de referencia X y se quiere transformar a un vector v' en el marco de referencia Y, se puede multiplicar el vector v por cada uno de los vectores base de Y utilizando el producto punto, cada resultado será el valor de una de las coordenadas del vector v’ en el marco de referencia Y.

## El proceso de Gram-Schmidt.

Como **lo ideal para ciencia de datos es que un conjunto de vectores sea ortonormal** formando una matriz ortogonal, dado un conjunto de vectores linealmente independientes no ortonormales podemos generar mediante el proceso de Gram-Schmidt un conjunto de vectores ortonormales.

Es decir, **el proceso de Gram-schmidt** es un método de álgebra lineal para ortogonalizar un conjunto de vectores (que deben ser linealmente independientes) en un espacio vectorial con un producto escalar, generalmente el espacio euclidiano.

De manera genérica seguimos los siguientes pasos:

1. El primer vector de la base se divide entre su modulo. Dando como resultado el primer vector de nuestra base ortonormal .
2. Tras esto para los siguientes “n” vectores hay que seguir dos procedimientos:
   1. El primero es hacer como una especie de operación que nos permite hacer el vector “n” perpendicular a los “n-1” vectores tratados anteriormente, esto se consigue con fórmulas propias de la proyección de un vector con respecto de otros. Por ejemplo, para el caso del segundo vector:

Tal que es el segundo vector de la base origen que se debe ortogonalizar con respecto de . Además, para el caso de tenemos que hacer la ecuación anterior no solo con respecto de sino también con respecto de que dando como sigue:

así, podemos concluir que para el vector tendremos que hacer la formula expresada con respecto de los n-1 vectores normalizados anteriormente.

* 1. El segundo es un proceso de normalización del vector (conseguir que su modulo sea 1), esto se consigue dividiendo el vector resultando del proceso “a” entre su modulo.

## Reflejando en un plano.

# Vectores característicos (característicos).

## ¿Qué son los vectores eigen?

Cuando multiplicamos un vector o conjunto de vectores por una matriz de transformación, el resultado es un nuevo vector o conjunto de vectores, si lo único que varía en el vector o en ciertos vectores tras realizar la operación, es el sentido y la magnitud manteniéndose la dirección igual, entonces hablamos de vectores característicos.

Es decir, podemos decir que los vectores eigen son aquellos vectores que tras multiplicarse por una matriz transformadora mantienen su dirección. Además, también hablamos de “Eigen value” o “valor característico” como el valor en el que se ha proporcionado el vector característico, es decir, ¿el vector característico es el doble de grande, o tal vez, es un medio más pequeño, o incluso, puede hacer cambiado su sentido, pero no su dirección teniendo un valor característico igual a un número negativo. Para aclarar, el valor característico λ no representa directamente la magnitud del eigenvector v, sino la escala por la cual se transforma el eigenvector cuando se multiplica por la matriz A

Cuando hablamos de un problema Eigen (problemas característicos) se está hablando de encontrar las características de algún tema.

## Casos especiales.

* Dado un conjunto de vectores que sufren una transformación lineal de manera proporcional en todas las direcciones, podemos afirmar que todos los vectores que sufren este tipo de transformación son vectores característicos.
* La rotación de un conjunto de vectores de 180 grados implica que todos los vectores siguen siendo vectores característicos.
* Cuando el eigen value es negativo significa que el vector característico ha cambiado de sentido pero que mantiene su dirección.
* En matrices de transformación diagonales (todo menos la diagonal a cero) los eigen valores son siempre los elementos de la diagonal.

## Descripción matemática de los eigen vectors y eigen values.

En ciertos casos no es fácil detectar los vectores característicos y se deben usar las matemáticas formales para poder encontrarlos. La fórmula básica queda como sigue:

Tal y como se aprecia, la multiplicación de un vector característicos por una matriz tiene que implicar que el vector mantenga su dirección y que se redimensione su magnitud en una proporción denotada por su “valor característico”. Tras realizar ciertas operaciones matriciales básicas de la anterior formula podemos llegar a la siguiente formula:

En este caso, o bien, el vector es cero, o lo que hay entre paréntesis es cero, para el estudio de los eigen vectores es mucho más interesante estudiar la formula:

De esta ultima ecuación podemos obtener los “valores característicos”, Teniendo en cuenta que a, b, c y d en los problemas serán valores conocidos. Con esto vamos a obtener una ecuación de segundo grado que nos permitirá conocer los valores característicos posibles, es decir, los posibles valores de tras resolver la ecuación, una vez obtenido este valor, volvemos a la formula:

Sustituyendo los valores de lamba en la ecuación podremos obtener los vectores característicos del sistema. Si en el polinomio característico no lograrmos obtener eigen values, entonces podemos afirmar que para esa transformación no existen eigen values.

## Cambio de base de vectores propios o característicos.

Cuando necesitamos multiplicar “n” veces una matriz de transformación sobre un vector podemos apreciar que a medida que crece “n” el problema a nivel de cálculo computacional se torna complejo y costoso, por lo tanto, para facilitar estos cálculos, se utiliza la diagonalización. Una matriz diagonal es aquella que en la que todos los elementos fuera de la diagonal principal de la matriz son cero.

Matemáticamente hablando, si tenemos una matriz de transformación y la aplicamos sobre un vector , podremos afirmar que siendo el vector el nuevo vector transformado, si bien, podemos aplicar una segunda transformación al vector nuevamente con la matriz quedándonos que:

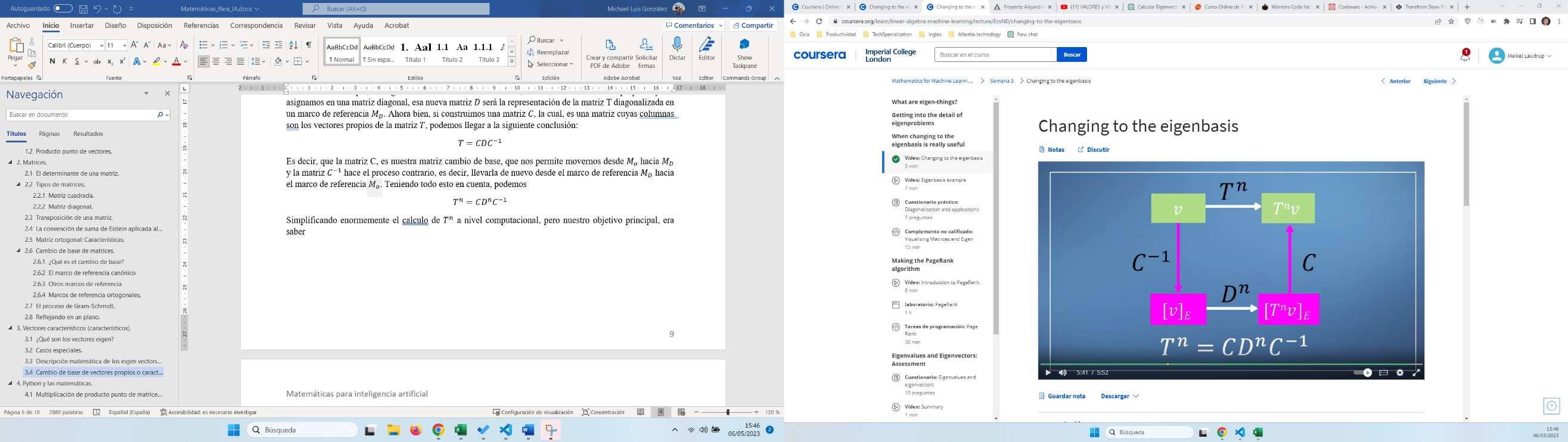
Si continuamos haciendo transformaciones de vectores nos quedaría, de manera genérica que:

Ahora bien, como habíamos comentado al principio de este epígrafe cuando “n” tiende a ser muy grande a nivel computacional el coste de realizar se puede tornar altamente costoso a medida que crece “n”, por todo esto, para optimizar el calculo se busca siempre que la matriz transformadora “T” sea una matriz diagonal.

Cuando la matriz Transformadora “” no sea diagonal, entonces buscaremos llevárnosla a otro marco de referencia cambiándola de base donde si sea una matriz diagonal, efectuar los cálculos propicios y traerla de nuevo a la base o marco de referencia origen para obtener el resultado final, simplificando enormemente el proceso de cálculo de la operación .

Pongamos dos marcos de referencia, el primero, será el **M**arco de referencia **O**rigen donde no es una matriz Diagonal, de ahora en adelante y pongamos el segundo **M**arco de referencia donde T sí que es una matriz **D**iagonal, denotado a partir de ahora como El marco de referencia tiene denotada la base y el marco de referencia tendrá la base .

Podemos afirmar que, si cogemos la matriz transformadora T, calculamos sus “valores propios” y los asignamos en una matriz diagonal, esa nueva matriz será la representación de la matriz T diagonalizada en un marco de referencia . Ahora bien, si construimos una matriz , la cual, es una matriz cuyas columnas son los vectores propios de la matriz , podemos llegar a la siguiente conclusión:

Es decir, que la matriz C, es nuestra matriz cambio de base, que nos permite movernos desde hacia y la matriz hace el proceso contrario, es decir, llevarla de nuevo desde el marco de referencia hacia el marco de referencia . Teniendo todo esto en cuenta, podemos

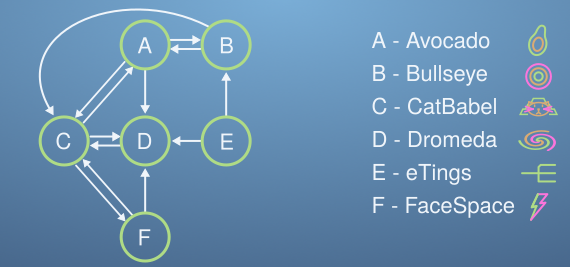
***CAMINO DIFICIL***

Simplificando enormemente el calculo de a nivel computacional.

## Un ejemplo práctico: El page rank.

El PageRank es un algoritmo desarrollado por Larry Page y Sergey Brin, los fundadores de Google, para determinar la importancia de las páginas web en función de la cantidad y la calidad de los enlaces que apuntan hacia ellas. Cuando el usuario realiza una búsqueda recibirá enlaces web ordenados en función de lo s resultados de este algoritmo La idea detrás del PageRank es que una página es más importante si otras páginas importantes enlazan a ella. En términos sencillos, el PageRank intenta determinar la probabilidad de que un usuario que navega aleatoriamente por la web termine en una página específica.

Imaginemos una micro-red de páginas webs, las cuales, son nodos de un grafo y las flechas que unen los nodos son links entre estas paginas web, por ejemplo:



Claramente se aprecia visualmente que la página web más importante será la página C, porque es las que más nodos tienen un enlace hacia a él, en este caso cuatro links externos pueden acceder hacia ella. Ahora, si imaginamos a una persona, que tienen mucho tiempo libre y se ponen a navegar en la web de manera un poco aleatoria. Mapeando todos los posibles links, se puede construir un modelo para estimar la cantidad de tiempo que gastara esta persona en cada pagina web. Podemos describir los links que hay en cada página web como un vector, en la cada posición se asignará un valor que será 0 o bien tal que “n” es el número de links externos que hay en la página web, en el caso de A, hay 3 links externos, por lo que hay una probabilidad de de que uno de los enlaces sea pulsado.

Si aplicamos esto para todos los nodos del grafo, teniendo en cuenta, que cada columna será un vector de cada página web tendremos que:

Al aplicar el algoritmo de PageRank a la matriz de enlaces L, buscamos encontrar un vector de importancia (PageRank) r que sea estable bajo la transformación dada por la matriz de enlaces. Esto significa que cuando multiplicamos la matriz L por el vector r, el vector resultante sigue siendo proporcional al vector r original. En otras palabras, queremos que:

Aquí, λ es el eigenvalor y r es el eigenvector correspondiente. En el caso del algoritmo de PageRank, buscamos el eigenvector correspondiente al eigenvalor dominante (más grande), que generalmente es 1. Este eigenvector r es el vector de PageRank que buscamos.

Entonces, el problema de encontrar el vector de PageRank se reduce a encontrar el eigenvector asociado con el eigenvalor dominante de la matriz de enlaces L. Utilizamos técnicas como la iteración de potencias para encontrar este eigenvector de manera eficiente, especialmente cuando trabajamos con matrices grandes.

Imagina que el Internet es una gran red de páginas web conectadas por enlaces. Un "surfer" (navegante) en esta red comienza en una página y sigue los enlaces de una página a otra de manera aleatoria. El PageRank representa la probabilidad de que un "surfer" que navega aleatoriamente en esta red termine en una página específica después de un número suficientemente largo de pasos.

Las páginas con un PageRank más alto se consideran más importantes y, por lo tanto, tienen más probabilidades de aparecer en los primeros resultados de una búsqueda en Google. El algoritmo de PageRank originalmente fue una parte clave del sistema de clasificación de Google, aunque desde entonces se ha vuelto más complejo e incluye muchos otros factores además del PageRank.

Cuando se dice que el "PageRank es el orden clasificado de las páginas desde la más probable hasta la menos probable que el navegante esté viendo", significa que las páginas se ordenan en función de su importancia en la red (según el algoritmo de PageRank), desde la página con la probabilidad más alta de ser visitada por un "surfer" que navega aleatoriamente hasta la página con la probabilidad más baja de ser visitada. Este ranking es una medida de la importancia y relevancia de cada página en la web.

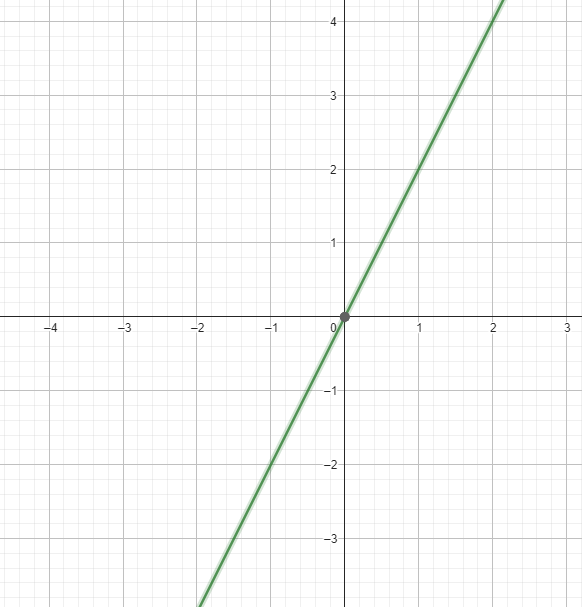
# Funciones.

## ¿Qué son las funciones?

De manera informal una función recibe es como una caja negra que recibe unos inputs y a partir de ellos genera una salida. Una manera semiformal de definirla es como:

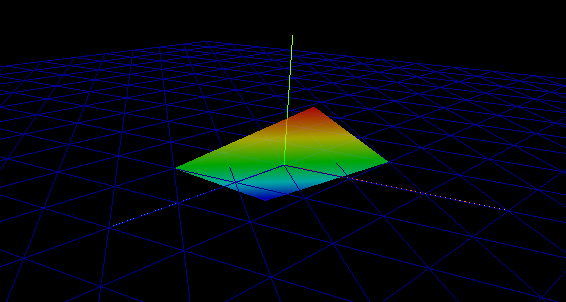
Tal que inputs puede ser un arreglo de variables y un output suele ser un calculo a partir de los inputs. Las funciones más simples son aquellas que solo tienen una variable, por ejemplo, una función que dada un valor cualquiera como entrada, te devolverá su doble:

Normalmente las funciones matemáticas tienen una representación gráfica asociada, esta se consigue dibujar asignado puntos en una gráfica cuya distancia del punto (0,0) a nivel horizontal es el valor que se le da a “x” y cuya distancia con respecto del punto (0,0) a nivel vertical es el resultado de la función, uniendo los puntos conseguimos dibujar la línea que representa la gráfica, por ejemplo, para el caso anterior tenemos que su grafica asociada es:

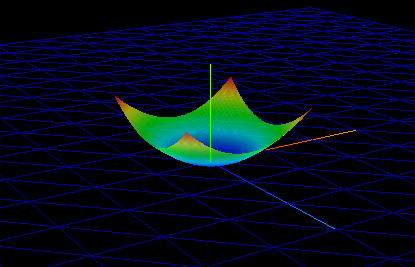


De aquí podemos concluir que las funciones de una variable son una línea (recta o curva) en un gráfico, y esta puede ser continua o discontinua, además hay otras funciones más complejas compuestas por varias variables, por ejemplo, una función que represente la suma, dado dos inputs devuelve su suma:

Las funciones de dos variables suelen representarse como superficies como planos o planos curvos, por ejemplo, para la anterior función tenemos que:



Otro ejemplo podría ser la gráfica para la función , ya finalmente, si subimos un nivel de complejidad más alta, donde tenemos funciones de tres variables, podemos hablar de figuras geométricas en el espacio.



Algunos enlaces de interés son:

* Ver funciones más conocidas: https://rstudio-pubs-static.s3.amazonaws.com/824875\_494efd5800f64deeaee802c3c4dbad5d.html
* Diagramar funciones: <https://www.geogebra.org/graphing?lang=es>
* Diagramar funciones de varias variables: <https://www.mathstools.com/section/main/3D_Functions_Plotter?lang=es#.ZFqCBXZBxD8>

## Funciones con parámetros.

La notación se utiliza para denotar una función f que depende de dos variables: x y los parámetros ​. La variable x es la variable independiente, mientras que ​ es un parámetro (o conjunto de parámetros) que afecta el comportamiento de la función f. El subíndice k en podría indicar que hay varios parámetros a en el conjunto ​ que se ajustan en algún contexto específico.

En términos más simples, la función está describiendo una relación entre la variable independiente x y el parámetro o conjunto de parámetros ​ que modulan cómo se comporta la función. Esto puede ser útil cuando tienes una función que depende de varios parámetros y quieres enfatizar uno en particular en un contexto específico.

Para diferenciar entre parámetros y funciones de varias variables, podemos pensar en un escenario práctico. Imagina que estamos modelando el crecimiento de una planta en función del tiempo y varios factores ambientales. Aquí, el tiempo y los factores ambientales serían las variables de entrada en nuestra función. Sin embargo, algunos aspectos del crecimiento, como la tasa de absorción de nutrientes y la influencia de la luz solar, pueden variar según la especie de planta. Estos aspectos específicos se considerarían parámetros, ya que son valores ajustables que afectan la relación entre las variables de entrada y la tasa de crecimiento. En resumen, mientras que las variables de entrada describen las condiciones cambiantes en un problema, los parámetros permiten ajustar la función para adaptarse a diferentes situaciones y representar cómo las variables de entrada interactúan con factores específicos.

Nota: Es importante subrayar que los parámetros “son constantes” y las variables no. Es decir, una vez elegimos los parámetros, permanecerán constantes y lo que variarias serán las variables independientes de la función.

# Derivada o gradiente

## ¿Qué es una derivada?

Las derivadas, gráficamente hablando, pueden considerarse como la pendiente de la recta tangente de una gráfica en un punto determinado de una función, al final viene a ser dada esa recta tangente calcular lo que asciende entre dos puntos dividido entre lo que recorre horizontalmente de izquierda a derecha. Formalmente, podemos afirmar que la derivada es la cantidad que crece o decrece una función entre dos puntos, dividido entre la distancia entre esos dos puntos cuando está tiende a cero. Su ecuación asociada viene definida por:

En palabras simples, **La derivada de una función en un punto específico mide cómo cambia la salida de esa función cuando la variable independiente (por ejemplo, x en f(x)) experimenta un pequeño cambio.** Si quieres ver derivadas con más detalle

* Enlaces de interés: <https://youtu.be/V7r4amUPI9k>

## Reglas de las derivadas.

Como vimos anteriormente para funciones relativamente complejas calcular la anterior función puede volverse tedioso, es por todo esto que para las funciones más comunes se tienen lo que se denominan derivadas inmediatas, algunas veces, habrá que jugar con las matemáticas para transformar derivadas no inmediatas en derivadas inmediatas.

* **Regla de la suma:**

Esta regla viene a decir que, si detectamos que nos piden calcular la derivada de la suma de “n” funciones, nos limitemos a calcular las derivadas de cada una de las funciones individualmente y luego las sumemos, simplificando los cálculos.

* **Regla de la multiplicación**:

Esta regla podría extrapolarse para una multiplicación de tres funciones, tal que:

Si aplicamos para cuatro funciones tendríamos que:

Si se estudia con un poco de detallelas 4 funciones podemos afirmar que:

Enlace recomendado: <https://www.youtube.com/watch?v=GEujiBObdRM>

* **Regla de la división**: Existe una reglar específica para la derivada de una función, la cual esta constituida por división de dos funciones:

No obstante, esta regla no es estrictamente necesaria de aprender, dado que cualquier división de funciones puede ser reescrita para ser representada como una multiplicación y así aplicar la regla de la multiplicación, un ejemplo ilustrativo sería:

Teniendo esto en cuenta podemos concluir que está regla no es estrictamente necesaria memorizarla.

* **Regla de la cadena:** La regla de la cadena es aplicable queremos calcular la derivada de una función compuesta, es decir, una función que tiene como entrada otra función, por ejemplo:

En este caso tenemos dos opciones, o extendemos de tal forma que la convertimos en una función extensa y calculamos su derivada aplicando la regla de la suma o aplicamos la regla de la cadena. Que dice así:

## Derivadas inmediatas:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## ¿Qué es un gradiente?

Se puede entender como una derivada, pero con varias variables, más formalmente, el gradiente es una generalización del concepto de derivada para funciones multivariables. Cuando trabajas con una función de una sola variable, la derivada y el gradiente son esencialmente lo mismo, ya que el gradiente sería simplemente un vector de una dimensión que contiene la derivada de la función con respecto a su única variable.

## Derivadas parciales.

Una derivada parcial es una medida de cómo cambia una función en relación con una variable específica, mientras se mantienen las demás variables constantes. En otras palabras, calcula la tasa de cambio de una función en relación con una sola variable independiente, sin considerar las otras variables. Esta se denota como donde f es la función y x es la variable con respecto a la cual se está derivando parcialmente.

Las derivadas parciales son especialmente útiles en el análisis de funciones de múltiples variables, ya que permiten estudiar cómo la función cambia en relación con cada una de las variables de forma individual.

## Derivada total

Puede considerarse como una generalización de la derivada parcial. Mientras que la derivada parcial mide la tasa de cambio de una función en relación con una variable específica, la derivada total considera cómo cambia la función en relación con todas las variables de forma simultánea, se suele definir como o simplemente df y se define como:

En esta expresión, (∂f/∂xi) representa la derivada parcial de f con respecto a la variable xi, y dxi representa el diferencial de la variable xi.

## Regla de la cadena en funciones de varias variables.

Antes de entrar en profundidad, para simplificar la representación formal de la regla de la cadena, podemos afirmar que notación matemática, se puede expresar:

Es decir, que la serie de variables puede ser representa como una “x” en negrita. Prosiguiendo si tenemos una función cuyas variables constituyen funciones en sí misma, podemos aplicar la regla de la cadena, tal que si cada es a su vez una función que está respecto de una variable “t”, por poner un ejemplo, entonces podemos concluir que:

Podemos resumir que es muy similar al concepto de regla de la cadena aplicado para derivadas de una sola variable, ahora bien, otra manera de calcular esto es con la siguiente formula:

## Vector Jacobiano.

Si tenemos una función de muchas variables:

Podemos definir el Jacobiano como un vector en el cada uno de sus elementos es la derivada parcial de la función con respecto de cada una de las variables, tal que:

Podemos afirmar que aquellos valores de la tupla ( en los que el vector jacobiano (es decir, el gradiente) de la función es cero, son puntos críticos de la función, los cuales podrían representar máximos, mínimos o puntos de silla. Estos puntos son donde la función puede cambiar de dirección, lo que puede indicar un máximo o mínimo local (o incluso global), o un punto de silla.

Sin embargo, el vector jacobiano por sí solo no nos proporciona suficiente información para determinar si un punto crítico es un máximo, un mínimo o un punto de silla. Para clasificar los puntos críticos, se suele recurrir a la matriz hessiana, como se verá más adelante.

## Matriz jacobiana.

La matriz jacobiana se puede construir cuando se tienen varias funciones, de tal manera de que cada fila de la matriz jacobiana se corresponde con el vector jacobiano de una función concreta, por ejemplo, si tenemos que:

## La matriz hessiana.

En este caso volvemos a enfocarnos en una única función, el hessiano de una función consiste en una matriz cuadrada de dimensión “n x n” tal que “n” es el número de variables de la función, la matriz hessiana consiste en la derivada segunda con respecto de cada variable que conforme la función para cada uno de los elementos del vector jacobiano de una función.

Es importante mencionar que para que el hessiano exista, la función debe ser al menos dos veces diferenciable. Además, el hessiano de una función es simétrico si la función es dos veces continuamente diferenciable.

Un ejemplo práctico sería el siguiente, tenemos la función:

Lo primero que hacemos es calcular el jacobiano de esta función tal que:

Luego calculamos la derivada de cada uno de los elementos del vector jacobiana con respecto de cada una de las variables de la función original, es decir “x”, “y” y “z” quedándonos que:

En la primera fila hemos derivado cada uno de los elementos del vector jacobiano con respecto de x, en la segunda fila con respecto de y, y finalmente, en la tercera fila con respecto de z, obteniendo así la matriz hessiana.

Ahora bien, la gran utilidad de la matriz hessiana es que nos permite detectar cuando un punto crítico es un máximo, un mínimo o un punto de silla, hay varias maneras, una es con el análisis de los eigenvalores de la matriz hessiana y otro, que es el que se aborda en este epígrafe, es mediante el análisis del determinante de la matriz hesiana y el elemento en la posición i = 0 y j = 0 de la matriz hessiana, tal que si denotamos el determinante de la matriz hessiana como D, es decir D = det(H), podemos concluir que, para aquellos puntos críticos que hacen el jacobiano igual a cero sí:

* Si D > 0 y > 0, entonces la función tiene un mínimo local en el punto.
* Si D > 0 y < 0, entonces la función tiene un máximo local en el punto.
* Si D < 0, entonces el punto es un punto de silla.
* Si D = 0, entonces la prueba es inconclusa, y se necesita más información para clasificar el punto.

Es importante subrayar que para funciones de tres o más variables, el criterio de la segunda derivada se vuelve más complicado y no se puede determinar simplemente con el signo de D y el primer elemento de la matriz hessiana. En general, para funciones de tres o más variables, se debe examinar los eigenvalores de la matriz hessiana para determinar el tipo de punto crítico.

# Fundamentos matemáticos de redes neuronales.

## Modelos.

Un modelo, en el contexto de la ciencia de datos y la inteligencia artificial, se refiere a una representación matemática o computacional de un sistema o proceso del mundo real, simplificándolo y pudiendo entender mejor dicha realidad. Los modelos se utilizan para entender, analizar y predecir el comportamiento de dicho sistema o proceso.

Para ilustrarlo, pensemos en un mapa como un ejemplo de modelo. Un mapa es una representación simplificada de un mundo tridimensional en un espacio bidimensional. Este modelo simplificado nos permite entender mejor nuestra ubicación y facilita la navegación en el entorno, prescindiendo de detalles innecesarios como la textura del terreno o el color de los edificios. A pesar de que el mapa no captura la totalidad del mundo real, es increíblemente útil porque destaca la información que más nos importa, por ejemplo, las distancias relativas y las ubicaciones de los lugares.

En esencia, la creación de un modelo se trata de encontrar el equilibrio correcto entre simplicidad y utilidad. Un modelo eficaz debe ser lo suficientemente simple para ser comprensible y manejable, pero también debe ser lo suficientemente complejo para capturar los aspectos más importantes del sistema que se está modelando. Los mejores modelos son los que nos permiten hacer predicciones precisas y útiles a pesar de su simplicidad.

Además, a los modelos se les puede dotar de probabilidad para atender ciertos grados de incertidumbre, dando como resultado los modelos probabilísticos.

Si tenemos datos que representan nuestra realidad y poseemos información acerca de los datos “reales” y los datos que capta “nuestro modelo” podemos afirmar que la diferencia entre los datos que capta nuestro modelo y los datos reales se denomina como “error”, la idea es que ajustando los parámetros que definen nuestro modelo acercarnos mucho más a la realidad.

Los tres elementos fundamentales son:

* Los datos: toma de contacto con la realidad, mediciones.
* Los parámetros: son aquellos que nos dan la flexibilidad para ajustar nuestro modelo y hacer que se parezca lo máximo posible a la realidad. Aunque no siempre es bueno dotar al modelo de tanta flexibilidad.
* El error: Definir una función de error que nos diga como nuestro modelo se ajusta o no a los datos de la realidad. Normalmente cuando se usan algoritmos de aprendizaje supervisado esta función de error se calcula a partir de los datos de salida suministrados y en el caso de aprendizaje no supervisado otras medidas se computan a partir de los datos de entrada. Esta función es importante porque nos sirve para hacer “optimización” y para hacer un entrenamiento o ajuste del modelo.

Enlaces:

* https://www.youtube.com/watch?v=Sb8XVheowVQ

## Definición de una neurona a nivel matemático.

### La neurona como función

Una neurona es la unidad básica de procesamiento más simple que nos vamos a encontrar en una red neuronal, en términos simples una neurona recibirá una entrada que denotaremos como “x” y generará una salida denotada como “y” dependiendo de los valores de entrada, comportándose exactamente igual que una función. Ahora bien, la función que representa la neurona viene definida por:

En esta función matemática podemos apreciar una ecuación lineal en el interior de otra función , objeto de estudio más adelante, en la que la variable de entrada tiene como producto una variable “w” que llamaremos peso (**w**eight), esta variable nos dice, cuanto de importante es para la neurona el valor de esa entrada, es decir ¿Cuál es su peso o su importancia? Ahora puede que no le veamos mucho el sentido, pero cuando la neurona reciba varias entradas, a cada una de estas entradas se le dará más o menos importancia en función de su peso asociado.

### El papel del Bias

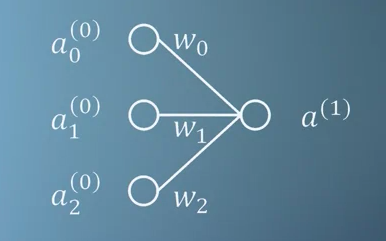
Para entender el papel del 'Bias' (b), consideremos un caso práctico. Imagina que nuestra neurona controla un sistema de calefacción, produciendo una salida de 1 para activar la calefacción y 0 para desactivarla. La entrada 'x' es la temperatura actual de la casa, y queremos que la calefacción se active cuando la temperatura caiga por debajo de los cinco grados. Ahora, si la temperatura llega a 0 grados, nuestra función sin el término de 'Bias', se simplifica a , lo que sugiere que la calefacción debería apagarse, lo cual claramente no es lo que queremos. Para tratar estos casos, introducimos el término de Bias 'b' en la ecuación lineal, resultando en , lo que nos permite ajustar la salida de nuestra neurona incluso cuando la entrada es 0, asegurando así que nuestra calefacción se active cuando la temperatura es críticamente baja.

### La función de activación.

Cuando los datos presentan patrones y relaciones que no pueden describirse de manera directa como una línea recta, se necesita la no linealidad para capturar esas complejas relaciones. Las funciones de activación en las redes neuronales introducen no linealidad al transformar las entradas ponderadas de una neurona en una forma que no se ajusta a una línea recta simple. La función de activación determina si una neurona debe activarse (emitir una señal) o no.

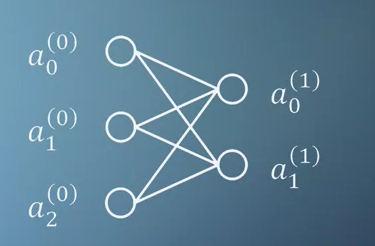
Un ejemplo de función de activación es

## Definición matemática formal de una red neuronal.

Siguiendo la nonmeclatura utilizada por la “Imperial College of London” podemos definir la función de una neurona como:

Esto podría leerse como la salida de la neurona en el nivel “1” será igual a la función de activación con parámetros de entradas igual la suma de el estado de activación de las neuronas del nivel anterior que tenga conectadas ponderadas por unos pesos y sumándole unas vías.

Es decir, que está neurona a del nivel 1 se activará según que neuronas estén activadas del nivel anterior, dándole importancia más a unas que a otras asociándoles pesos (importancia).

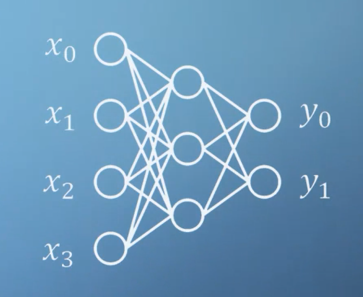
Además, nuevamente siguiendo la nomenclatura de la imperial college de londres, una red compuesta por varias neuronas podrá definirse con la siguiente expresión matemática:

De tal forma que nos queda que:

* Este es el estado de entrada a la neurona. En otras palabras, es el valor o valores (en forma de vector) que la neurona está recibiendo como entrada.
* **:** Este es el estado de salida de la neurona. Es el resultado de la función de activación aplicada a la suma ponderada de las entradas a la neurona más el sesgo.
* Esta es la **función de activación** de la neurona. Actúa como una especie de umbral, determinando si la neurona se "activa" o no en base a la entrada ponderada que recibe. Las funciones de activación son fundamentales en las redes neuronales porque permiten introducir no linealidad en el modelo, permitiendo que este aprenda y modele relaciones complejas.
* **w(weight):** Esta es la variable que representa el peso de la conexión entre dos neuronas. En términos simples, el peso determina cuánto contribuye una neurona a la siguiente. Es un parámetro ajustable que la red neuronal aprende durante su entrenamiento.
* **B (Bias):** Este término se refiere al sesgo o "bias". Funciona como una constante que se suma a la entrada ponderada antes de pasar por la función de activación. El sesgo permite ajustar la salida de la neurona junto con el peso, proporcionando un grado adicional de libertad al modelo.

## Definición de una red neuronal a nivel matemático.

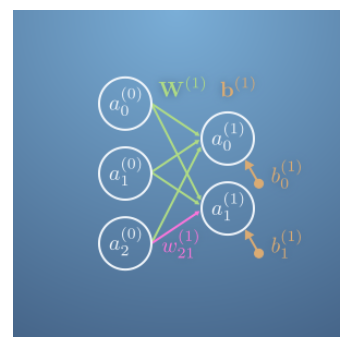
Una red neuronal puede entender como una función, que dada unas entradas genera una salida, tanto la entrada como la salida pueden ser vectores de entrada-salida.

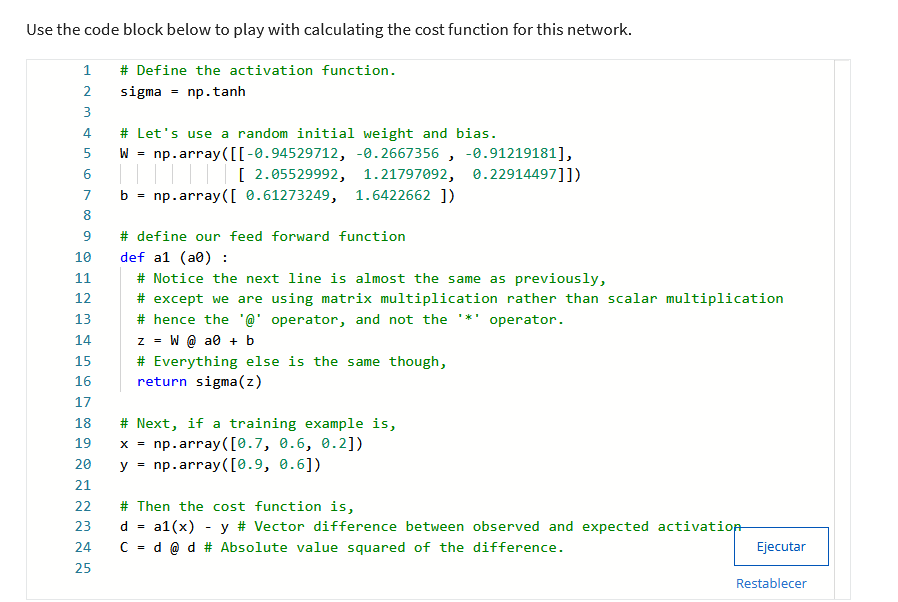


Es decir, la anterior imagen podría representarse como:

**Nota:** recordar que la “x” e “y” en formato negrita representan una tupla de variables, además, el número de variables de entrada “m” puede ser diferente del número de variables de salida “n”

## Coste de la red neuronal (margen de error)





## Retro propagación en redes neuronales.

La retropropagación (backpropagation en inglés) es un **algoritmo fundamental utilizado en el entrenamiento de redes neuronales artificiales**. Su objetivo es ajustar los pesos y sesgos de las neuronas de una red neuronal para que esta pueda aprender a realizar una tarea específica, como el reconocimiento de patrones en datos de entrada.

El proceso de retropropagación implica dos fases principales: la fase de propagación hacia adelante (forward propagation) y la fase de retropropagación (backward propagation).

### Propagación hacia adelante (Forward Propagation):

Durante esta fase, los datos de entrada se propagan a través de la red neuronal, capa por capa, desde la entrada hasta la capa de salida. Cada neurona en una capa recibe las salidas de las neuronas de la capa anterior, multiplica esas salidas por sus pesos respectivos, suma el sesgo y aplica la función de activación para generar una salida. Esta salida se convierte en la entrada para las neuronas de la siguiente capa. Este proceso se repite hasta que se alcanza la capa de salida, y la red produce una predicción.

### Retropropagación (Backward Propagation):

Una vez que se ha generado una predicción en la fase de propagación hacia adelante, **se calcula el error entre la predicción y el valor real deseado**. Este error se utiliza para ajustar los pesos y sesgos de las neuronas en la red neuronal, de modo que la red pueda mejorar su rendimiento en la tarea. **El algoritmo de retropropagación calcula las derivadas parciales del error con respecto a los pesos y sesgos en cada capa de la red**, utilizando la regla de la cadena del cálculo diferencial. **Estas derivadas indican cómo los cambios en los pesos y sesgos afectan el error final**.

Las derivadas permiten medir cómo cambia una variable (como el error) en respuesta a pequeños cambios en otra variable (como los pesos y sesgos de la red). En el contexto de las redes neuronales, las derivadas se utilizan para ajustar los pesos y sesgos con el objetivo de minimizar el error en las predicciones.

Cálculo del Error:

En el entrenamiento de una red neuronal, se necesita medir qué tan bien están haciendo las predicciones en comparación con los valores reales. Esto se logra calculando una medida de error, como el error cuadrático medio (MSE), que representa la diferencia entre las predicciones y los valores reales. **El error es la función que queremos minimizar.** Un ejemplo de esta función puede ser:

# Taylor series para aproximaciones.

## Introducción.

Si tenemos una función relativamente compleja con la que es difícil trabajar, puede llegar a ser útil, que dado un punto de esa función compleja obtengamos una función aproximada mucho más sencilla de trabajar que de valores similares a la función compleja alrededor de ese punto elegido. Para ese propósito se debe entender las series de Taylor y posteriormente el polinomio de Taylor.

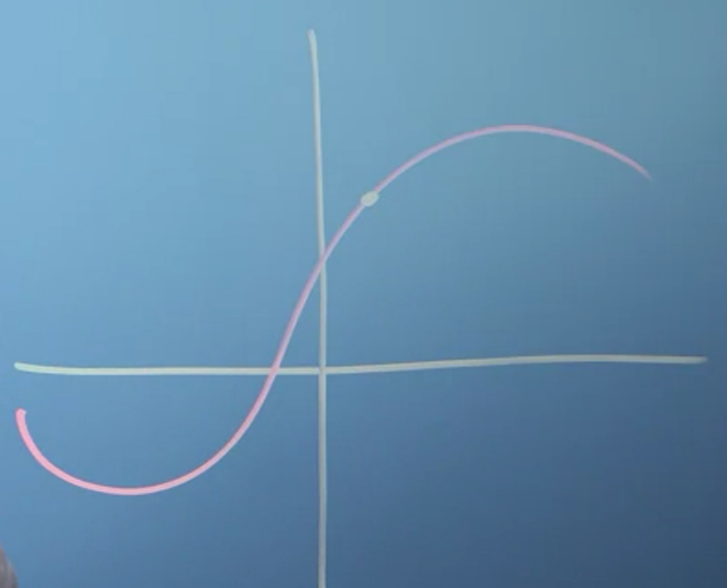
## Serie

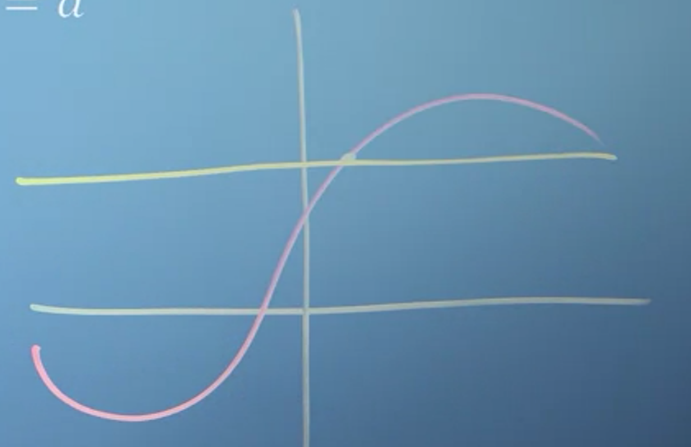
En matemáticas, una serie es una suma infinita de términos. En su forma más básica, una serie se representa como la suma de los términos de una secuencia numérica. Cada término puede ser positivo, negativo o cero, y la serie resulta de sumar todos estos términos en orden. La forma general de una serie es la siguiente:

Los aspectos más importantes de una serie incluyen su convergencia, divergencia, suma y comportamiento en relación con los términos individuales.

## Serie de potencias

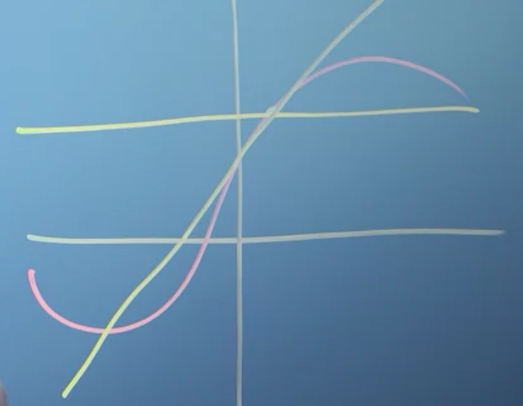
En este caso empezamos con la serie de potencia, que se visualiza de la siguiente manera:

**Ahora bien, imaginemos que tenemos la siguiente función que la consideramos como compleja y que hemos elegido el punto descrito en la imagen como nuestro punto para obtener una función similar a su alrededor.



Podemos afirmar que, si empezamos con una aproximación de orden cero, y ponemos que nuestra función alternativa tenemos una representación como la que se aprecia en la imagen de la derecha tal que:

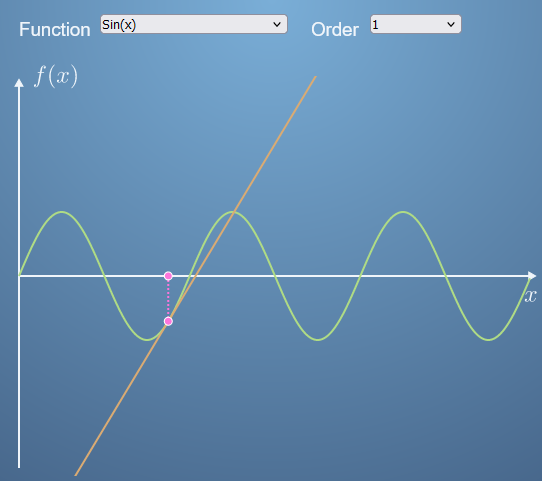
Siendo a igual al valor en el eje y del punto que hemos seleccionado. Sin embargo esta función alternativa solo se aproxima a nuestro punto únicamente en el mismo punto no en su alrededor.

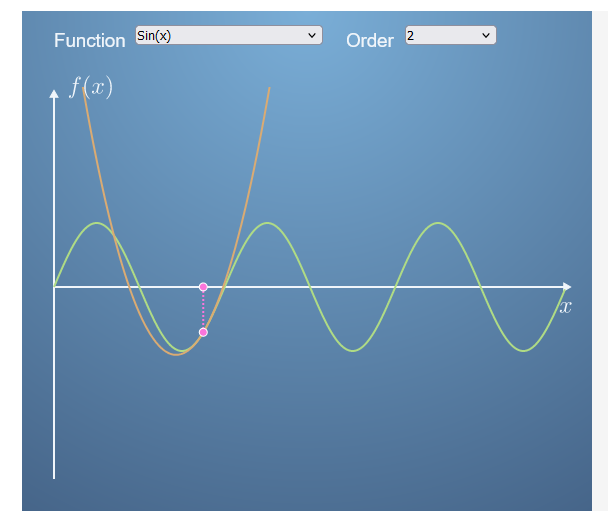
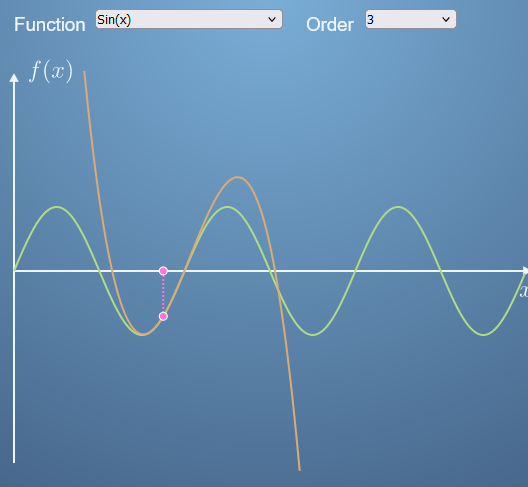
Para encontrar una aproximación más cercana podemos recurrir a:

Tal y como se aprecia en la imagen de la derecha en este caso tenemos una grafica pendiente que se aproxima más a los puntos de alrededor del punto que hemos elegido.

Si proseguimos aumentando el orden con los siguientes casos nos daremos cuenta que cada vez se parecerá más a la función origina alrededor del punto seleccionado:

Colectivamente, a estas pequeñas partes de la serie se le denomina “Series truncadas”. Veamos un ultimo ejemplo donde podemos apreciar que a partir de una función relativamente compleja original (verde) y un punto (en rosado) podemos obtener un función de aproximación más aproximada a medida que aumentamos el orden, por ejemplo, la función Seno:





## Derivación en series de potencia (Formula de Maclaurin).

La reescritura de la representación de una función en términos de una serie de potencias crecientes de xx es un proceso fundamental en matemáticas. El método de las series de Taylor aboga por la idea de que si tenemos un conocimiento completo de una función en un punto particular aa (esto incluye su valor, sus derivadas de distintos órdenes como la primera derivada, la segunda derivada, y así sucesivamente), entonces tenemos una comprensión completa de cómo se comporta dicha función en todas partes en su dominio.

Este poderoso principio tiene una limitación: solo se aplica a aquellas funciones que exhiben un comportamiento suave y regular, a las que se les da el nombre de "funciones bien comportadas". Para que este método funcione de manera confiable, la función debe ser continua en su dominio y capaz de ser derivada infinitas veces. En otras palabras, su estructura debe ser lo suficientemente regular como para que todas sus derivadas, incluso las de orden superior, sean definidas y finitas en cualquier lugar donde estemos interesados.

Pero antes de entrar en la serie de Taylor, centrémonos en la serie de Maclaurin, que es el predecesor de la serie de Taylor, en esta serie el punto objeto de estudio es cuando x = 0 y a partir de ese punto de va obteniendo una representación de la función más compleja de manera aproximada con una serie de potencias truncadas, que dándonos que:

Notas:

* será igual a la función original con la “x” sustituida.

## Series de Taylor.

La serie de Taylor y la fórmula de Maclaurin son muy similares, pero con una diferencia clave. La fórmula de Maclaurin es una versión específica de la serie de Taylor donde se expande una función alrededor del origen (a=0). Esta fórmula establece que, una vez que conocemos los valores y las derivadas de una función en x=0, podemos usarla para aproximar la función en cualquier lugar cercano al origen.

Por otro lado, la serie de Taylor es más general. No se limita a aproximar la función solo en x=0. Permite expandir una función alrededor de cualquier punto a, no necesariamente el origen. Una vez que conocemos los valores y las derivadas de una función en x=a, podemos usar la serie de Taylor para aproximarnos a la función en cualquier punto cercano a x=a.

En resumen, la fórmula de Maclaurin es una forma específica de la serie de Taylor, pero la serie de Taylor es más flexible al permitir la expansión en cualquier punto a, no solo en el origen. La expresión queda como sigue:

Nota: Si empezamos con el orden cero y vamos subiendo del orden 0, hacia el 1, 2, 3 , etc. Una subida de orden no siempre implica una mejora de aproximación, porque hay ciertas funciones que hacen que ciertos términos de la serie se “anule” para determinados valores de “n”, siendo series pares aquellas en las que cuando “n” es un numero par se obtiene siempre un valor de termino que mejora la precisión y series impares cuando para valores de “n” impares se mejora la precisón pero para valores pares se obtienen términos nulos.

## Linealización o aproximación lineal.

La linearización es una aproximación de una función complicada por una función lineal más simple que se parece a la función original en las cercanías de un punto específico. Esto se logra mediante el uso del polinomio de Taylor de primer orden (también conocido como el polinomio de Maclaurin **de primer orden**), que es una aproximación lineal de una función alrededor de un punto.

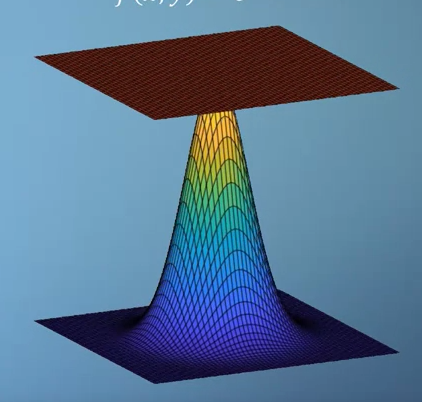
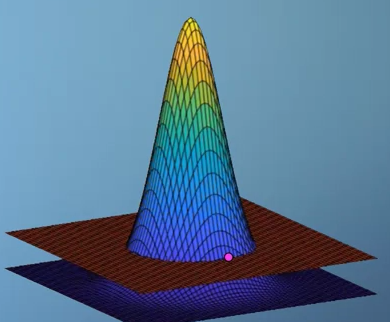
La linearización es especialmente útil cuando necesitas una estimación rápida de una función en un punto cercano. Cuanto más cerca esté x de a, más precisa será la aproximación. Sin embargo, a medida que te alejas más de a, la aproximación lineal puede volverse menos precisa debido a las características no lineales de muchas funciones. En esos casos, se pueden utilizar términos de orden superior en la expansión de Taylor para obtener aproximaciones más precisas.

Es decir, la linearización es el proceso por el cual, cuando queremos aproximarnos a una función compleja utilizando el polinomio de Taylor, ignoramos aquellos sectores que exceden el primer orden, es se debe a que a medida que vamos avanzando en los términos de la serie, se vuelve cada más pequeño, por lo que su efecto en el resultado final cada vez es más pequeño. Podemos entonces asimilar un error:

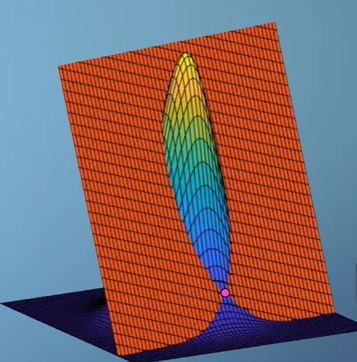
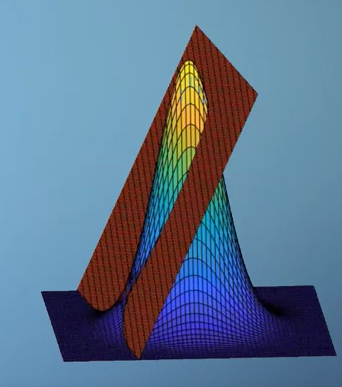
Tal que representa el término del error de orden superior, que es la diferencia entre la función real y la aproximación lineal. Este término captura cómo el error cambia a medida que Δx se hace más pequeño.

## Series de Taylor con varias variables.

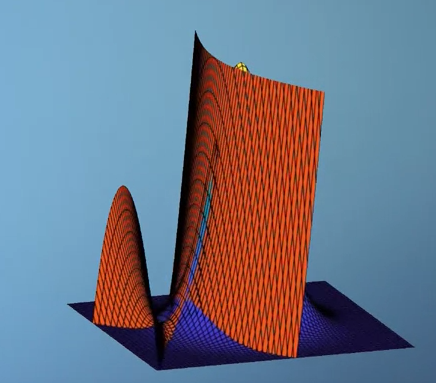
Una aproximación de orden cero se vería como un plano que se entrecruza con el punto exacto del que queremos hacer la aproximación, tal y como se aprecia en las siguientes imágenes para la función :



En el caso de una aproximación de primer orden tendremos un plano que hace como pendiente del punto escogido, tal y como se aprecia en las siguientes imagenes:

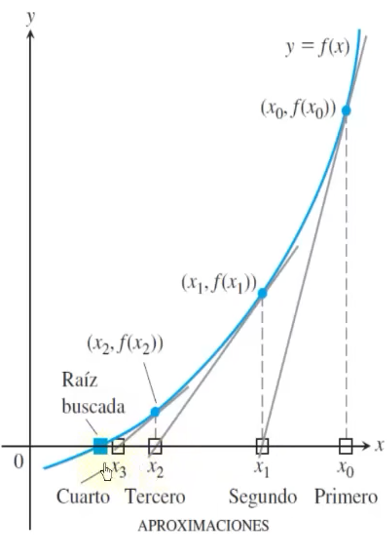


Para el caso de orden dos ya tenemos un plano curvado que se asimila a más puntos de la función original, tal y como se vislumbra en la siguiente imagen:



# Introducción a la optimización.

## Método de Newton-Raphson

Es un algoritmo utilizado para encontrar aproximaciones de las raíces (ceros) de una función real. Es decir, encontrar aquellos valores de “x” que hacen que la función sea igual a cero. Es un método iterativo que se basa en ajustar tangentes a la curva de la función para acercarse a las raíces y puede ser muy eficiente (en el contexto computacional) para encontrar raíces de funciones en comparación con otros enfoques más simples.

La idea detrás de este método es suponer una raíz que creamos que está cerca de la raíz real del problema. Una raíz, en este contexto, es un valor de x que hace que la función sea igual a cero. Una vez que elegimos este valor de x como nuestra raíz hipotética, comenzamos a realizar aproximaciones sucesivas, conocidas como iteraciones.

Cada iteración ajusta tangentes a la curva de la función y se acerca más a la raíz real. El objetivo es llegar a un valor "casi exacto" de la raíz mediante este proceso iterativo. Este método puede ser especialmente eficiente en términos computacionales en comparación con otros enfoques más simples para encontrar las raíces de funciones.

Notas:

* Gráficamente las raíces de una función son aquellos puntos donde la línea, plano, etc. que representa la función tocan el eje x.

Notas:

* Los puntos de cambio de curva de creciente/decreciente de una función suelen ser problemáticos para este algoritmo. Esto se debe a que un punto de inflexión o de silla podrían hacer que la derivada fuera igual a cero, implicando divisiones por cero y consecuentemente indeterminaciones.

## Descenso de gradiente.

El algoritmo de descenso de gradiente es un método de optimización utilizado en el campo de la optimización matemática y el aprendizaje automático para encontrar mínimos locales (o máximos locales) de una función. Este algoritmo es ampliamente utilizado en problemas de optimización en los que se busca encontrar los valores de las variables que minimizan (o maximizan) una función objetivo.

El concepto básico detrás del descenso de gradiente es bastante simple. Imagina que estás en la cima de una montaña y quieres llegar al punto más bajo en el valle. En lugar de mirar a tu alrededor para encontrar la dirección más empinada hacia abajo, el descenso de gradiente te sugiere que des un paso en la dirección de mayor aumento en la pendiente (gradiente) de la función. Luego, repites este proceso iterativamente, ajustando el tamaño del paso en función de la pendiente en cada iteración, hasta que te acerques a un mínimo local.

Aquí hay una descripción general de cómo funciona el algoritmo de descenso de gradiente:

1. **Inicialización:** Se comienza con una estimación inicial para las variables de interés.
2. **Cálculo del gradiente:** Se calcula el gradiente de la función objetivo en el punto actual. El gradiente indica la dirección y la tasa de cambio máxima de la función en ese punto.
3. **Actualización de variables:** Se actualizan las variables en la dirección opuesta al gradiente multiplicado por un factor de aprendizaje (también conocido como tasa de aprendizaje). Este paso determina cuánto se mueven las variables en cada iteración.
4. **Repetición:** Se repiten los pasos 2 y 3 en iteraciones sucesivas, ajustando las variables para moverse hacia la dirección de menor incremento.
5. **Convergencia:** El algoritmo continúa iterando hasta que cierto criterio de convergencia se cumpla, como un número máximo de iteraciones, un cambio mínimo en la función objetivo o un cambio pequeño en las variables.

La fórmula viene determinada por:

## Optimización con Restricciones

La "Constrained Optimization" (Optimización con Restricciones) es un tipo de problema de optimización en el cual se busca encontrar los valores de las variables que maximizan o minimizan una función objetivo sujeta a ciertas restricciones o condiciones. Estas restricciones pueden ser limitaciones en las variables, ecuaciones o desigualdades que deben cumplirse mientras se busca el valor óptimo de la función.

En otras palabras, en la optimización con restricciones, no solo se busca el mejor valor para una función objetivo, sino que también se deben cumplir ciertas condiciones o restricciones establecidas en el problema. Estas restricciones pueden surgir debido a limitaciones físicas, recursos disponibles, reglas del problema o consideraciones prácticas.

Por ejemplo, considera el problema de maximizar los beneficios de una empresa sujeta a restricciones como el presupuesto disponible para inversiones, la cantidad máxima de recursos disponibles, o la necesidad de cumplir con ciertos requisitos legales. En este caso, estaríamos tratando con un problema de optimización con restricciones.

Resolver problemas de optimización con restricciones puede ser más complejo que la optimización sin restricciones debido a la necesidad de considerar tanto la función objetivo como las restricciones simultáneamente. Existen métodos y algoritmos específicos diseñados para abordar este tipo de problemas, como el método de los multiplicadores de Lagrange, que introduce factores de ajuste para equilibrar la función objetivo y las restricciones en el proceso de optimización.

## Multiplicadores de Lagrange.

Los multiplicadores de Lagrange son un recurso en matemáticas que te ayuda a encontrar los puntos donde una función alcanza un valor extremo (máximo o mínimo) bajo ciertas restricciones. Imagina que estás tratando de caminar en un terreno irregular y quieres encontrar el punto más alto al que puedas llegar, pero tienes que quedarte en un camino específico. Los multiplicadores de Lagrange te ayudan a encontrar el punto en el camino donde llegas a la altura máxima posible.

Aquí tienes los pasos básicos para usar multiplicadores de Lagrange:

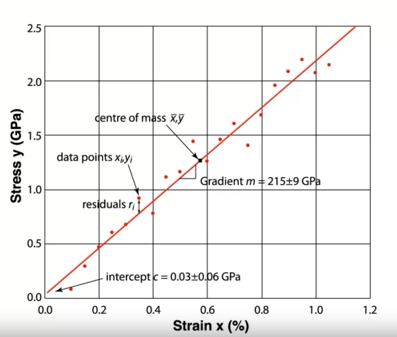
1. **Tienes una función objetivo:** Esto podría ser algo como "encuentra el valor máximo o mínimo de una función" (como la altura en el ejemplo del terreno).
2. **Tienes restricciones:** Estas son las reglas que debes seguir, como quedarte en un camino específico en el terreno.
3. **Formas una ecuación:** Para usar multiplicadores de Lagrange, formas una ecuación combinando la función objetivo y las restricciones multiplicadas por lo que se llama "multiplicador de Lagrange".
4. **Resuelves las ecuaciones**: Encuentras los puntos donde las ecuaciones se igualan y eso te da las coordenadas donde la función objetivo alcanza su valor extremo bajo las restricciones.
5. **Verificas y eliges**: Verificas si esos puntos son máximos o mínimos reales al comparar los valores de la función. El más alto o el más bajo será tu respuesta.

# Regresiones lineales.

## Modelo de regresión lineal simple.

La regresión lineal es una técnica estadística utilizada para modelar la relación entre una variable dependiente (también llamada variable de respuesta) que es el resultado de la función [f(x)] y una o más variables independientes (también llamadas variables predictoras), es decir, x, y, z, etc. El resultado final es modelar una ecuación de la forma

La idea básica es que, si tenemos en un plano de “N” dimensiones varios datos conocidos, a partir de esos datos conocidos podamos elaborar una función (modelo) que simplifique la realidad y nos permita predecir otros datos en otros puntos del plano de “n” dimensiones.

En este caso, si hablamos de regresión lineal simple, nos referimos a problemas manifestados en dos dimensiones, es decir, que la función que modeliza nuestro problema tiene una sola variable, a continuación, se muestra una imagen ilustrativa:

En la imagen de la izquierda podemos apreciar una nube de puntos, que representa como un objeto sometido a deformación (Strain) sufre una determinada tensión (stress).

Como vemos, la línea recta roja que se ha trazado busca simplificar la realidad y predecir que valor tendrá la tensión para otras deformaciones. Esto se llama método de regresión lineal y en este caso es simple porque solo hablamos del eje X e Y. La recta que buscamos es aquella que implique menos margen de error entre los puntos (datos reales) y el valor de la función estimado para ese punto x.

Recordemos que la recta tiene la forma de la función , donde tanto “m” como “b” son coeficientes, b es el valor de “f(x)” cuando x = 0, es decir, f(0) y “m” es la pendiente de la recta.

Para encontrar una línea recta que se ajuste lo mejor posible a los datos proporcionados se utiliza el método de mínimos cuadrados, que busca minimizar la suma de los cuadrados de las diferencias entre los valores observados (datos reales) y los valores predichos por la línea de regresión. La fórmula que se utiliza para calcular esta suma de cuadrados de residuos es:

Donde:

* ​ es el valor observado de la variable dependiente para el punto i.
* es el valor de la variable independiente (predictora) para el punto i.
* m es la pendiente de la línea de regresión.
* b es la ordenada al origen de la línea de regresión, es decir

Si embargo, esta fórmula de aquí tiene un coste computacional alto, por lo que se opta por hacer una redefinición usando derivadas parciales para hacer más fácil el cálculo computacional, de tal forma que:

en términos sencillos, esta expresión indica que estás buscando los valores de m y c para los cuales las derivadas parciales de la función de error con respecto a m y c son cero, lo que representa un mínimo local en la función de error y asegura que has encontrado la mejor línea de regresión que se ajusta a los datos.

## Modelo de regresión no lineal.

La regresión no lineal es una técnica estadística utilizada para modelar relaciones entre variables cuando la relación no se ajusta a una forma lineal, es decir, no puede ser representada mediante una línea recta. En contraste con la regresión lineal, que se basa en una ecuación lineal para predecir una variable dependiente y a partir de una o más variables independientes x, la regresión no lineal permite modelar relaciones más complejas y curvilíneas.

En la regresión no lineal, la ecuación utilizada para el modelado puede ser cualquier función no lineal, como polinomios de grados superiores, exponenciales, logarítmicas, trigonométricas, entre otras. Estas ecuaciones pueden capturar patrones más complejos presentes en los datos, lo que permite una mejor representación y predicción en comparación con la regresión lineal en casos donde los datos no siguen una relación lineal.

El proceso de ajuste en la regresión no lineal implica encontrar los parámetros de la función no lineal que minimizan la diferencia entre los valores observados y los valores predichos por el modelo. Esto se realiza generalmente utilizando métodos iterativos y algoritmos de optimización.

Supongamos que tienes un conjunto de datos que representan el crecimiento de una población de bacterias con el tiempo. Los datos son los siguientes:

|  |  |
| --- | --- |
| Tiempo (días) | Población de baterias |
| 1 | 10 |
| 2 | 20 |
| 3 | 30 |
| 4 | 40 |
| 5 | 50 |

Notarás que los datos parecen crecer de manera exponencial. Si intentamos ajustar estos datos utilizando una regresión lineal, no sería apropiado, ya que no seguirían una relación lineal. En cambio, podríamos usar una regresión no lineal con una función exponencial , donde y son los parámetros que ajustaremos. En este caso, el modelo sería:

Donde:

* es la población de bacterias en el tiempo t.
* son los parámetros de ajuste que queremos encontrar.
* e es la base del logaritmo natural.

El objetivo es encontrar los valores de a y b que minimizan la diferencia entre los valores observados y los valores predichos por el modelo. Utilizando métodos de optimización, podríamos encontrar y son los valores que mejor se ajustan a los datos. Luego, podríamos usar estos valores para predecir la población de bacterias en otros momentos.

Existen diversos métodos de optimización utilizados para encontrar los parámetros que minimizan una función objetivo en la regresión no lineal. Algunos de los métodos más comunes son:

* **Método de los mínimos cuadrados:** Este método busca minimizar la suma de los cuadrados de los residuos entre los valores observados y los valores predichos por el modelo. Se ajustan los parámetros del modelo para que la diferencia entre los datos observados y los valores predichos sea lo más pequeña posible.
* **Método de Descenso de Gradiente**: Este es uno de los métodos más utilizados en la optimización numérica. Consiste en iterativamente ajustar los parámetros en la dirección opuesta al gradiente de la función objetivo. Existen variantes como el Gradiente Conjugado y el Gradiente Estocástico.
* **Método de Gauss-Newton**: Es un método iterativo que se basa en aproximar la función no lineal con una función lineal en cada iteración. Requiere calcular y resolver sistemas lineales en cada paso. Es efectivo para problemas con soluciones cercanas al óptimo.
* **Método de Levenberg-Marquardt**: Una mejora del método de Gauss-Newton que incluye un término de regularización para manejar situaciones en las que la aproximación lineal no es precisa. Es ampliamente utilizado para ajustar curvas en regresión no lineal.
* **Algoritmo BFGS**: Es un método quasi-Newton que busca la inversa de la matriz hessiana de la función objetivo. Es eficiente en términos de memoria y puede ser muy efectivo en optimización no lineal.
* **Algoritmo de Powell**: Utiliza una combinación de direcciones para realizar pasos de búsqueda. Es útil en casos en los que la función objetivo puede no ser muy suave o bien comportada.
* **Método de Nelder-Mead**: También conocido como el método del simplex, es un algoritmo de optimización directa que no requiere el cálculo del gradiente. Se basa en construir un simplex en el espacio de parámetros y luego moverlo para buscar el mínimo.
* **Algoritmos Genéticos**: Se basan en procesos de selección y reproducción inspirados en la teoría de la evolución. Son especialmente útiles en problemas de optimización donde la función objetivo no es diferenciable o cuando hay múltiples mínimos locales.

### Método de los mínimos cuadrados.

El método de los mínimos cuadrados es una técnica fundamental en la regresión y el análisis de datos que busca encontrar los mejores valores para los parámetros de un modelo, de modo que la diferencia entre los valores observados y los valores predichos por el modelo sea minimizada. Esta técnica es especialmente útil cuando estamos tratando de ajustar una función no lineal a los datos y queremos encontrar los valores óptimos de los parámetros del modelo. Teniendo la siguiente formula:

Tal que:

* χ2 (chi cuadrado) es una medida cuantitativa de cuán bien se ajusta el modelo a los datos observados. Se trata de la suma de los cuadrados de las diferencias entre los valores observados yiyi​ y los valores predichos y(xi,ak)y(xi​,ak​) obtenidos a través de la función de ajuste.
* n es el número de puntos de datos en tu conjunto de datos.
* y(xi​,ak​) es el valor predicho por la función de ajuste yy en el punto xixi​, donde akak​ son los parámetros del modelo.
* σ2 es la varianza de los errores. Puede ser una constante si asumes que los errores tienen la misma magnitud en todos los puntos, o puede variar según los datos.

La anterior formula es una representación de la suma de los cuadrados de los residuos normalizados. Aquí, cada término en la suma representa la diferencia entre el valor observado ​ y el valor predicho elevado al cuadrado. Luego, esta diferencia al cuadrado se divide por σ2, que es la varianza de los errores.

Esta división por la varianza permite dar más peso a los errores más pequeños y menos peso a los errores más grandes, lo que puede ser útil para tratar con datos con diferentes niveles de precisión. El objetivo del método de los mínimos cuadrados es encontrar los valores de los parámetros que minimizan χ2, es decir, que hacen que la suma de los cuadrados de los residuos sea lo más pequeña posible. Cuando encontramos estos valores óptimos de los parámetros, estamos encontrando el mejor ajuste posible del modelo a los datos, en términos de minimizar la discrepancia entre los valores observados y los valores predichos.

# Estadística de los conjuntos de datos.

## ¿Qué es un conjunto de datos?

Un conjunto de datos es una colección organizada de información que se utiliza para realizar análisis, investigación, estudio o entrenamiento de algoritmos en el campo del aprendizaje automático y la inteligencia artificial. Un conjunto de datos consta de observaciones o ejemplos individuales, donde cada observación representa una entidad o evento particular. Estas observaciones pueden estar en forma de filas en una tabla (como en una hoja de cálculo) o como elementos en una estructura de datos en programación.

Un conjunto de datos puede contener una o más características (variables) que describen las propiedades o atributos de las observaciones. Cada observación se registra con valores específicos para estas características. Por ejemplo, si estás recopilando datos sobre estudiantes, las características podrían incluir la edad, el género, las calificaciones, etc.

Los conjuntos de datos pueden ser pequeños o enormes, dependiendo de la cantidad de observaciones y características que contienen. Además, pueden ser utilizados para diferentes propósitos, como:

* **Análisis Estadístico:** Para comprender patrones, tendencias y relaciones en los datos.
* **Aprendizaje Automático:** Para entrenar y probar modelos y algoritmos de machine learning.
* **Validación y Prueba:** Para evaluar el rendimiento y la capacidad de generalización de los modelos.
* **Investigación:** Para respaldar investigaciones científicas y análisis en diversas disciplinas.

Los conjuntos de datos varían en términos de su formato (tablas, archivos de texto, imágenes, videos, etc.), tamaño y naturaleza de los datos que contienen. El preprocesamiento y la limpieza de los conjuntos de datos son pasos críticos para asegurarse de que los datos sean coherentes y adecuados para el análisis o el entrenamiento de modelos.

En resumen, un conjunto de datos es una colección estructurada de información que se utiliza como base para el análisis, la investigación y el entrenamiento de modelos en el ámbito de la ciencia de datos y el aprendizaje automático.

## Propiedades de los conjuntos de datos.

### Media

La media, también conocida como promedio, es una medida estadística que representa el valor central de un conjunto de datos. Se calcula sumando todos los valores en el conjunto y luego dividiendo la suma por el número total de valores. La media proporciona una idea general de la tendencia central de los datos y es ampliamente utilizada en estadísticas y análisis de datos, en el contexto de conjunto de datos si tenemos un conjunto de datos D, entonces:

Tal que “n” es el número de conjunto de datos, son los valores individuales del conjunto de datos y E[D] es el valor esperado o la media del conjunto de datos, mas comúnmente conocida como . Cabe subrayar que los conjuntos de datos no siempre son numéricos y fácilmente operables matemáticamente hablando.

### Moda

### Varianza y desviación estándar.

La varianza es una medida estadística que describe la dispersión o variabilidad de un conjunto de datos con respecto a su media (promedio). En otras palabras, la varianza indica cuán alejados están los valores individuales de los datos de su valor promedio. En una población donde todas las personas tienen exactamente la misma altura la varianza será cero, porque no varía la altura de unos a otros, si varía ligeramente, es decir, que entre los diferentes individuos de la población los cambios de altura en cm no son muy altos unos respectos de otros, entonces tendremos una varianza relativamente baja.

Finalmente, si por el contrario, tenemos individuos que miden 2 metros, otros que miden 1,5 metros, y la variabilidad de las alturas es mayor y con mayor constraste, podemos afirmar que la varianza será alta.

En el contexto de los conjuntos de datos, la varianza es una medida clave para comprender cuánto se extienden los valores alrededor del valor central. Si los valores están muy cercanos a la media, la varianza será baja; si están más dispersos, la varianza será alta.

[Varianza poblacional]

Donde:

* n es el número de observaciones en el conjunto de datos.
* xi​ son los valores individuales del conjunto de datos.
* μ es la media del conjunto de datos.

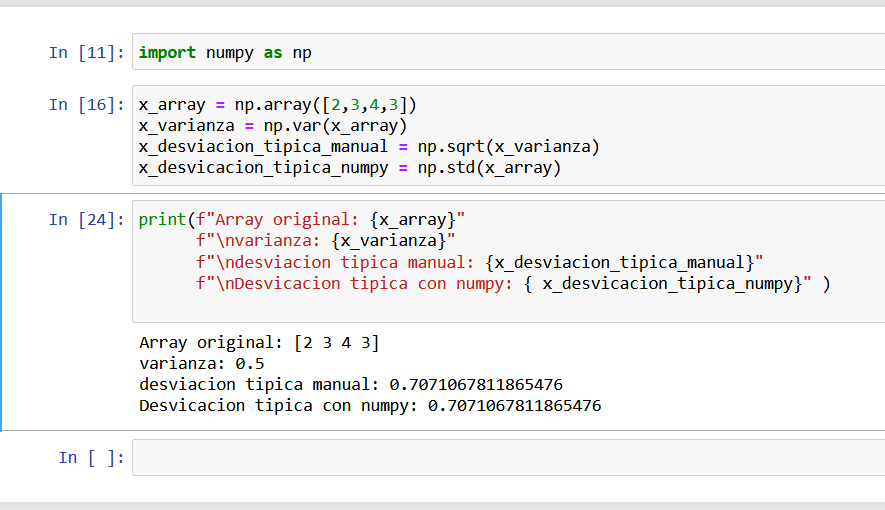
Además, tenemos la varianza muestral, que se diferencia de la anterior en que se le resta una unidad a la “N”, es decir:

[Varianza muestral]

Como podemos apreciar en ambas formulas la diferencia entre cada elemento individual de la muestra y la media es elevada al cuadrado para compensar las diferencias que den como resultado un numero negativo, es decir, intentamos céntranos más en las distancias que hay (modulo) que en los signos. Para compenzar este efecto de elevación al cuadrado tenemos la desviación estándar que consiste en la raíz cuadrada de la varianza poblacional:

Nota: no significa que se eleve al cuadrado el valor de sino que, “” es el propio nombre que se le da a la varianza.

En Python la varianza y la desviación standar se puede obtener fácilmente con la funciones “var” y “std” de numpy, tal y como se aprecia en siguiente imagen:



### Covarianza.

https://youtu.be/jo3XizfYjek?t=1078

|  |  |
| --- | --- |
| **Varianza poblacional** | **Covarianza poblacional** |
|  |  |
| **Varianza muestral** | **Covarianza muestral** |
|  |  |

Cuando tenemos conjunto de datos de “n” dimensiones o “n” catacteristicas, nos interesa estudiar la covarianza, que es

### Matriz de covarianza

La matriz de covarianza para dos dimensiones queda como sigue:

* Si miras la diagonal de la matriz de covarianza, obtienes las varianzas de cada variable, que te dicen cuánto varían estas variables respecto a sus medias.
* Si miras los valores fuera de la diagonal, obtienes las covarianzas entre pares de variables, que te indican cómo varían conjuntamente estas variables. Si la covarianza es positiva, las variables tienden a aumentar juntas; si es negativa, una tiende a disminuir cuando la otra aumenta y viceversa. Si es cero, es que no hay correlatividad.
* **La matriz de la varianza es siempre simétrica** porque .
* **Definida positiva:** Una matriz es definida positiva si cumple con ciertas condiciones. Una de las formas de determinar si una matriz es definida positiva es que todos sus eigenvalores sean positivos. Otra forma es que, para cualquier vector no nulo v, el producto es positivo, donde C es la matriz y es la transpuesta de v.

En el contexto de la matriz de covarianza, ser definida positiva garantiza que la matriz representa una estructura de variabilidad y covariabilidad válida. Esto significa que la varianza de cualquier combinación lineal de las variables será no negativa. Es una propiedad crucial, especialmente en áreas como la optimización y el análisis multivariante, porque asegura que no hay direcciones en las que la variabilidad sea negativa, lo que no tendría sentido en muchos contextos prácticos.

Esta matriz siempre es una matriz simétrica definida positiva

## Producto punto en el contexto de los conjuntos de datos.

En ocasiones, cuando tenemos un conjunto de datos representado como una nube de puntos en un espacio vectorial, resulta relevante explorar la similitud entre los distintos puntos de datos. El propósito fundamental es reducir la dimensionalidad del conjunto, buscando conservar la información esencial. En otras palabras, si dos puntos reflejan información muy similar, consideramos eliminar una de las dimensiones correspondientes, ya que esta dimensión podría no añadir información única y podría ser redundante, algo que podríamos obtener a partir de otras dimensiones.

La medida que nos indica el grado de similitud puede variar dependiendo del problema y los datos específicos. Puede ser distancia euclidiana, coseno, correlación, etc. Además, la elección de la medida de similitud puede influir en cómo se agrupan y reducen los datos. En este caso vamos a utilizar la ortogonalidad, estrechamente relacionada con el concepto de correlación.

Se utilizan las proyecciones ortogonales de los puntos de datos como una manera de comprimir los datos sin perder información, si pensamos en los puntos de datos como vectores en un espacio vectorial. La distancia entre estos puntos y el Angulo entre ellos

# Alta dimensionalidad de la información.

Datos de alta dimensionalidad se refiere a conjuntos de datos que tienen un gran número de dimensiones o características. En otras palabras, cuando se trabaja con datos de alta dimensionalidad, cada entrada (muestra) en el conjunto de datos está representada por un gran número de características.

Por ejemplo, considera una matriz de datos donde cada fila representa una persona y las columnas representan diferentes características como edad, altura, peso, ingresos, etc. Si tienes muchas columnas, digamos cientos o miles, entonces estarías tratando con datos de alta dimensionalidad.

El término "alta dimensionalidad" no solo se refiere al número total de características, sino también a la relación entre el número de características y el número de muestras en el conjunto de datos. Cuando el número de características es considerablemente mayor que el número de muestras, esto puede llevar a problemas y desafíos específicos en el análisis y procesamiento de datos, como la maldición de la dimensionalidad, que puede dificultar la interpretación de los datos, aumentar los requisitos de cómputo y afectar negativamente el rendimiento de los algoritmos de machine learning.

Además, de todas estas dimensiones, ciertas, pueden ser redudantes es decir, que contengan información que ya se puede obtener de otras dimensiones, por ejemplo, si tenemos que una persona tiene dos características como la edad y el año de nacimiento, realmente, es redudante está información, porque con saber la fecha de nacimiento, ya sabemos la edad.

## Reducción de la dimensionalidad.

La reducción de dimensionalidad explota la estructura y la correlación y nos permite trabajar con una representación de la información o los datos más compacta idealmente sin perder información.