Intro a la Ciencia de Datos -Regresion

Falacias

Paradoja de simpson

Ejemplo extracción de cálculos renales

En un estudio se compararon los porcentajes de éxito de la extracción de cálculosrenales según el método utilizado.

Cirugía Abierta	Nefrolitotomía percutánea
78%	83%

Faltan datos

Por ejemplo:

- Como fue el post operatorio
- Rango de edad. No es lo mismo operar a gente de 80 que de 20

El estudio concluía que lo mejor era la nefrolitotomía percutánea. Sin embargo,partiendo el dataset de otra forma lo que se puede ver es distinto.

Lo que sucede es que los datos que se tienen son de tratamientos reales, en donde amás grande la piedra más chances hay de ir a cirugía. No es que la cirugía sea peor, sino que se le suelen asignar los casos más complejos.

	Cirugía Abierta	Nefrolitotomía percutánea
Piedra <2cm	93%	83%
Piedra >=2cm	73%	69%

Podemos agarrar a la mitad de los pacientes y arbitrariamente elegirlos al azar untratamiento.
En medicina esto sería bastante polémico, pero sirve para muchos otros casos en donde hacer este tipo de pruebas no mata a nadie. Esto se llama **validación cruzada**.

Ejemplo Wikipedia

Primero cartel en naranja. Despues cartel en rojo → Las donaciones aumentan 5%

¿Cómo sabemos si fue gracias al cambio del botón o por otra razón?

Aparte la diferencia es poca.

Sesgo de supervivencia

Estados Unidos de cara a la segunda guerra mundial analiza los aviones que vuelvendel combate. Concluye que debe reforzarlos en donde más se dañaron.

Resulta que estaban dañados en todos los lugares que no son críticos, porquepudieron volver.
Alguien nota esto e indica que en realidad se debe reforzar los lugares en donde novemos daños de los que volvieron.

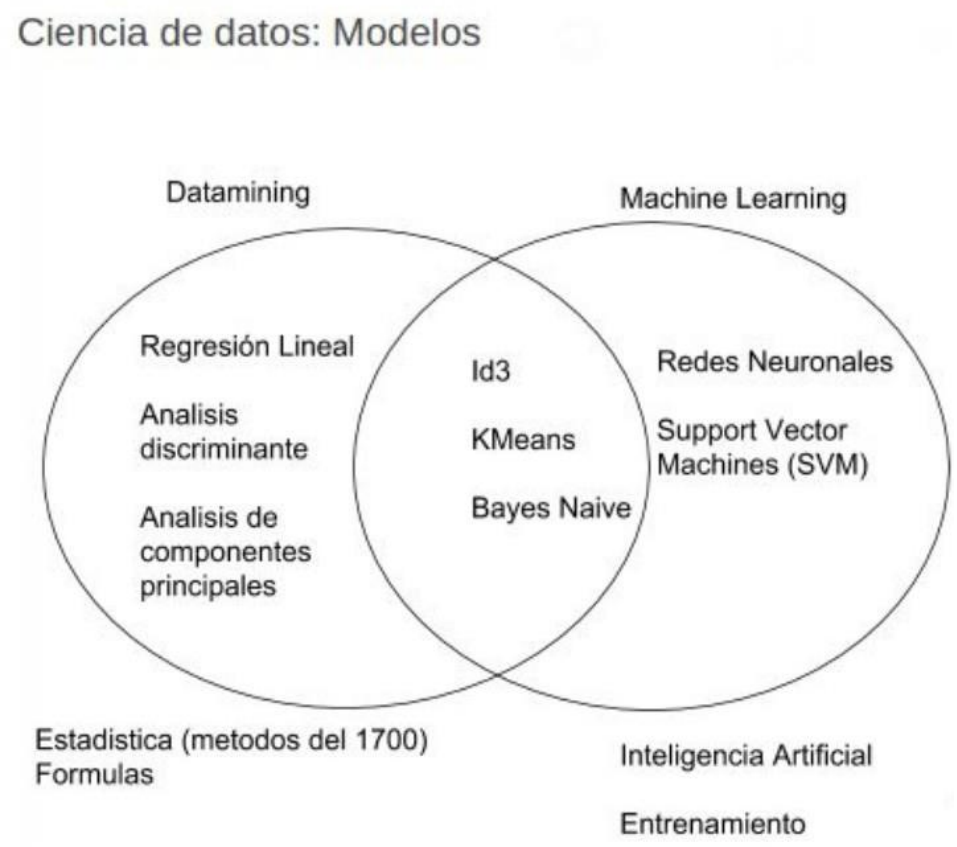
Se llama sesgo de supervivencia porque solo miramos los datos de los sobrevivientes,no del conjunto total de datos.

Siempre hay que preguntarnos: **¿cuál es el origen de nuestros datos?**

Introduccion a la Ciencia de Datos

Modelos predictivos → que nos permitan predecir casos futuros

Tenemos un conjunto de datos, que representan diferentes ejemplos u observaciones yla idea es que a partir de esos datos podamos crear un modelo tal que cuando venga una nueva observacion, un nuevo ejemplo podamos predecir el valor asociado.



Variables

Cuando estamos trabajando con los datos, entonces los vamos a considerar variablesque vamos a utilizar para entrenar a los modelos

- Variables Independientes (entradas)
- Variables dependientes (salidas, categorías)

Variables Independientes:

- Cualitativas: no tenemos rangos numericos
  - Texto
    - Nominales (categorías, ejemplo: países)
    - Ordinales (poco, mucho, muchísimo) → hay un orden, relacion
  - Númericas
    - Nominales (id, numero de telefono) → No hay orden
    - Ordinales (1,2,3)

- Cuantitativa: son las numericas, describen cantidades
  - Discreta: acotada a conjuntos de los numeros naturales (dias, minutos,…)
  - Continua: conjunto de los numeros reales. (altura de una persona)

    En informatica no existe ⇒ son variables de punto flotante, son un subconjunto acotado y discreto de los reales pero que permiten aproximar de manera satisfactoria

Variables y tipos de problemas

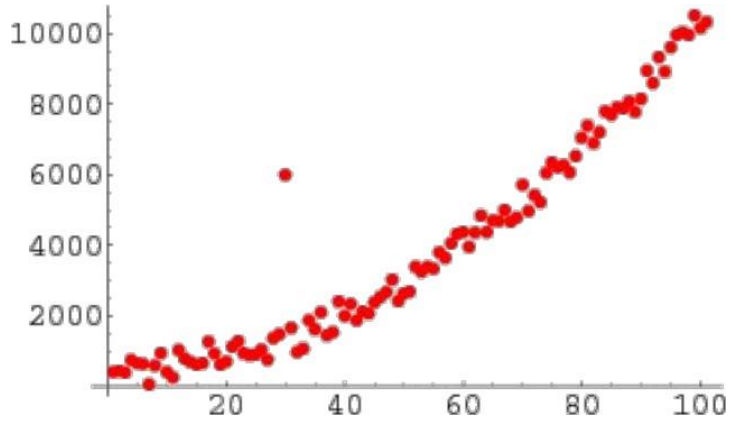
1. Si la variable dependiente es cualitativa, el tipo de problema es de clasificación ejemplo: var de salida es pais. a partir de ciertos valores de entrada tengo quedeterminar a que pais corresponde dicho ejemplo

a. Si la variable dependiente es cuantitativa, el problema es de regresión ejemplo: si tengo que determinar la temperatura, la altura de una persona,…

2. Si NO hay variable dependientes, el problema es de agrupamiento o clustering
El algoritmo automaticamente va a tratar de segmentar los valores de entrada en conjuntos. Y va a extraer caracteristicas que son propias de esos datos, innatas, ocultas, pero que los hacen similares. Las caracteristicas pueden no ser evidentes,visibles de entrada, pero el algoritmo los a encontrar.

Outliers (valor atípico)

Son valores que no se ajustan a la muestra. Tienen un comportamiento no esperado.



Pueden ser muchas cosas

- un valor mal ingresado
- dato real que por determinada circunstancia se alejan de la media, se alejan del resto de los datos

Sabiendo el problema que queremos resolver, ahi veremos si el outlier es interesantepara nosotros y lo tenemos que analizar o si lo eliminamos.

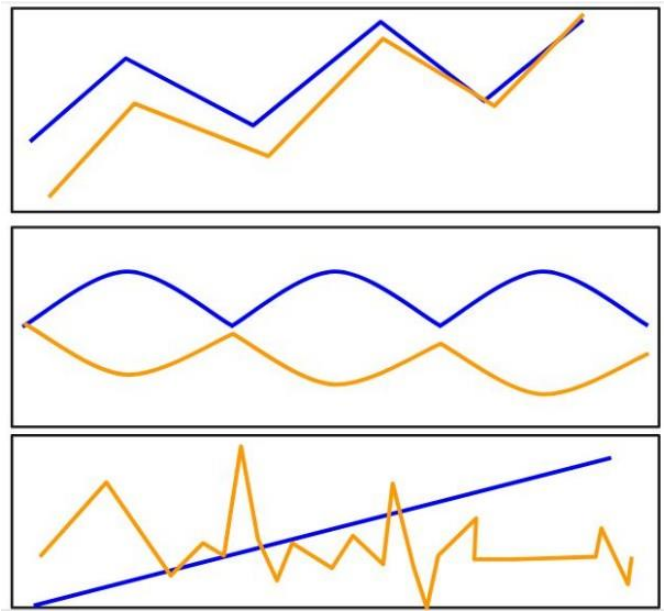
Correlación de variables

Dos variables están correlacionadas cuando varían de igual forma sistemáticamente.

Positiva
Se comportan masomenos igual

Negativa
Varian de forma opuesta

Sin correlacion



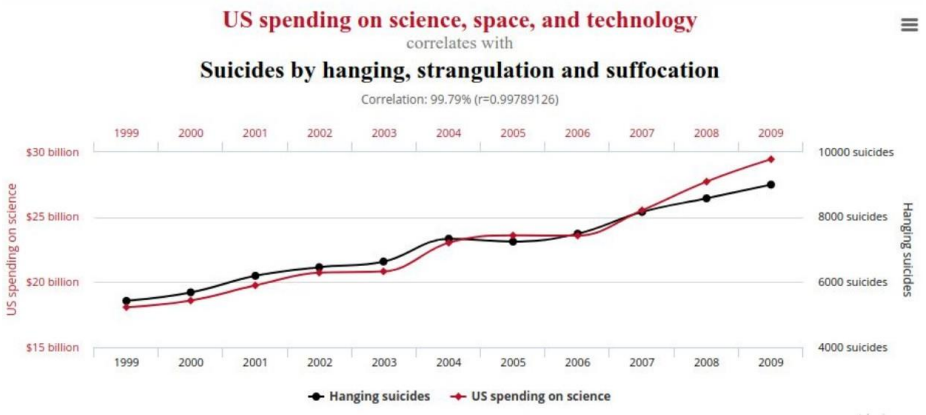
Correlación NO IMPLICA Causalidad

● Que dos variables tengan alto índice de correlación no significa que una cause la otraNo implica que una este forzando o modificando la otra

● Las relaciones de causalidad son mucho más difíciles de encontrar y demostrar

● Las correlaciones pueden suceder por otros motivos como: Una tercer variable que "empuja" a ambas o simplemente azar

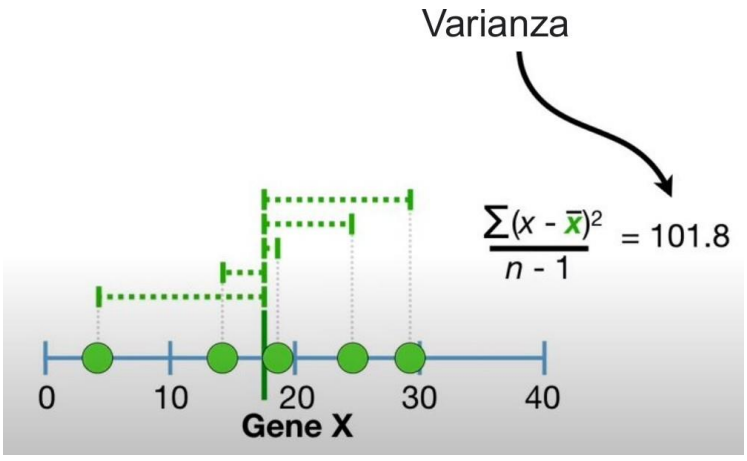
Ejemplos de correlaciones sin sentido



Varianza

Promedio de la diferencia, entre todas las observaciones, respecto de su media.Nos dice que tan dispersos estan los datos respecto de su media.

La media es: 17.6



Si la varianza es chica ⇒ estamos muy cerca de la media

Si es grande ⇒ están más alejados, dispersos

Covarianza

En probabilidad y estadística, la covarianza es un valor que indica el grado de variaciónconjunta de dos variables aleatorias respecto a sus medias.
Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal o la recta de regresión.

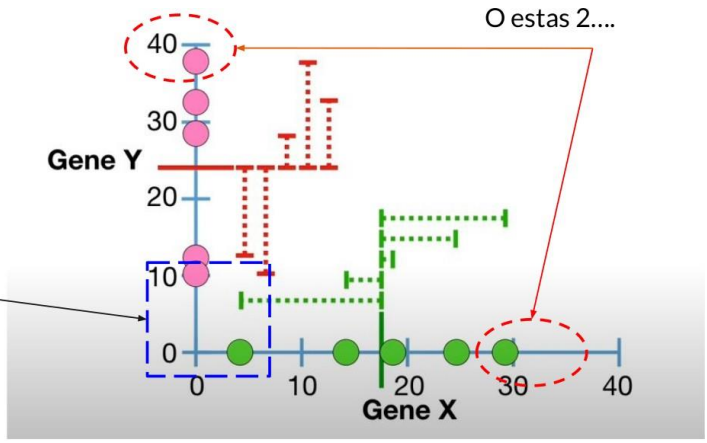
Verificamos si la variacion conjunta de cada variable respecto de su media guarda algun tipo de relacion, si las variables estan alejando de la media de la misma forma enambos ejes.

Dos variables.

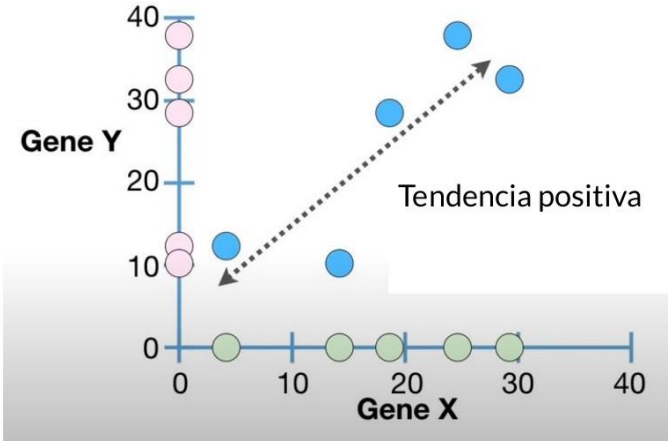
Calculamos la varianza de cada una....

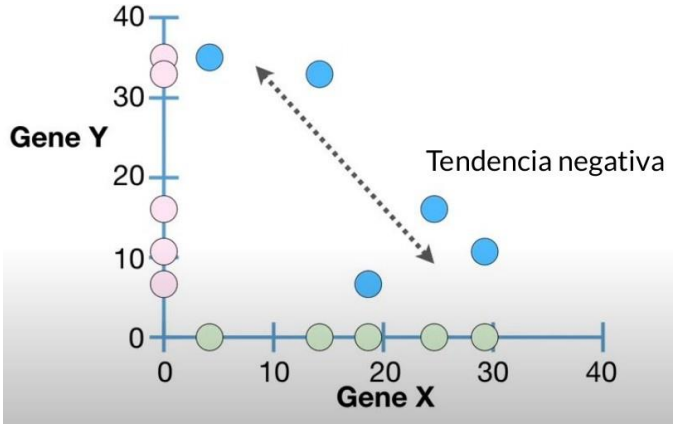
Estas dos medidas corresponden a una misma medición. A la misma observación.

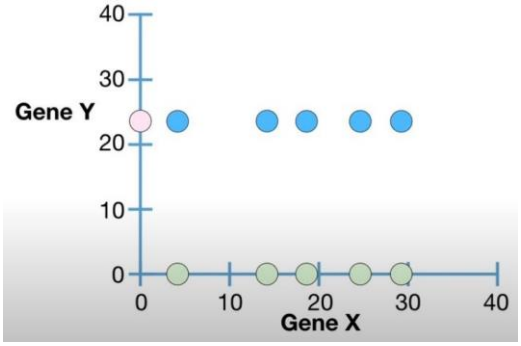
¿Son sus variaciones respecto de la media simiares?



Cómo los datos en Y y X, pertenecen a una misma medición podemos graficarlos en 2D y ver si hay alguna tendencia...







No hay tendencia

Correlación de Pearson

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

donde

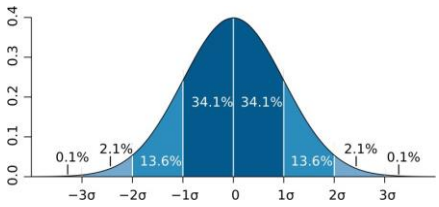
- σ_{XY} es la covarianza de (X, Y)
- σ_X es la desviación estándar de la variable X
- σ_Y es la desviación estándar de la variable Y

Para 2 variables podemos medir su correlación lineal con el coeficiente de correlación r (Pearson). Este coeficiente, es una función que mide cuán relacionada estan 2 variables de forma lineal.

- Si da 0 ⇒ NO existe correlación
- Si da 1 ⇒ Están relacionadas linealmente de forma perfecta (todos los puntos estánen una línea)
- Si da -1 ⇒ Existe una correlación negativa perfecta.

Desvío estandar

Es una medida que se utiliza para cuantificar la variación o la dispersión de un conjunto de datosnuméricos.
Una desviación estándar **baja** indica que la mayor parte de los datos de una muestra tienden a estar agrupados cerca de su media (también denominadael valor esperado).
Una desviación estándar **alta** indica que los datos se extienden sobre un rango de valores más amplio.



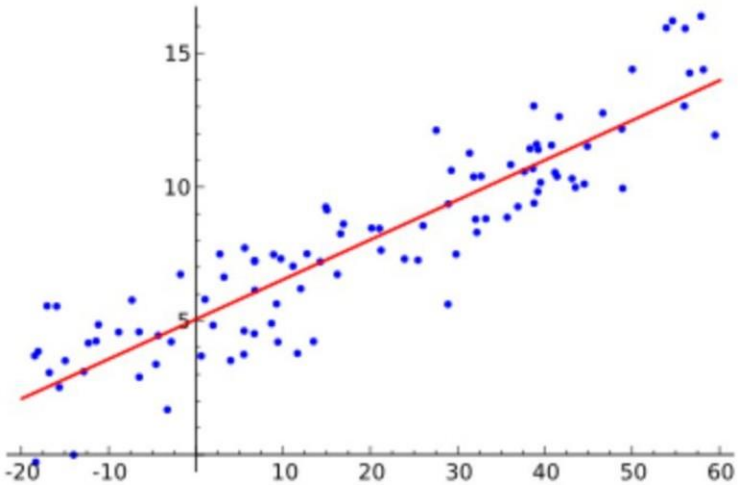
Campana de Gauss o distribución normal

Métodos de regresión

Buscamos predecir un valor en un rango continuo, para ciertos valores de entrada.

Ejemplos: Temperatura, Valor de una propiedad

Regresión lineal o ajuste lineal



Dado una serie de observaciones (puntos) encontrar una línea tal que aproxime de la mejor forma posible a todos los puntos.

Pseudo-code

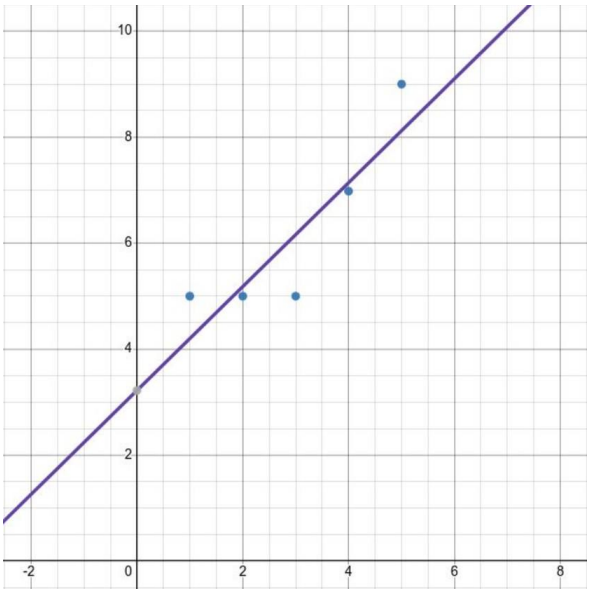
```
vars
xarray = [            1, 2, 3, 4, 5      ],
yarray = [            5, 5, 5, 6.8, 9    ],
x = y = xy = xx = a = b = resultado = 0,
cantidad = xarray.length,
```

```
for (i = 0; i < cantidad; i++) {
      x += xarray[i];
      y += yarray[i];
      xy += xarray[i]*yarray[i];
      xx += xarray[i]*xarray[i];
}
b = ((cantidad * xy) - (x * y)) /
     ((cantidad * xx) - (x * x));
a = (y - (b * x)) / cantidad;
```

$y = x*b + a$

b = 0.98

a =3.22



Generalización

Ecuación Normal:

$$\hat{\theta} = (X^T X)^{-1} \quad X^T \quad y$$

```
import numpy as np
xarray = [            1, 2, 3, 4, 5      ],
yarray = [            5, 5, 5, 6.8, 9    ]
X_b = np.c_[np.ones((5, 1)), xarray] # add x0 = 1 to each instance
theta_best = np.linalg.inv(X_b.T.dot(X_b)).dot(X_b.T).dot(yarray)
```

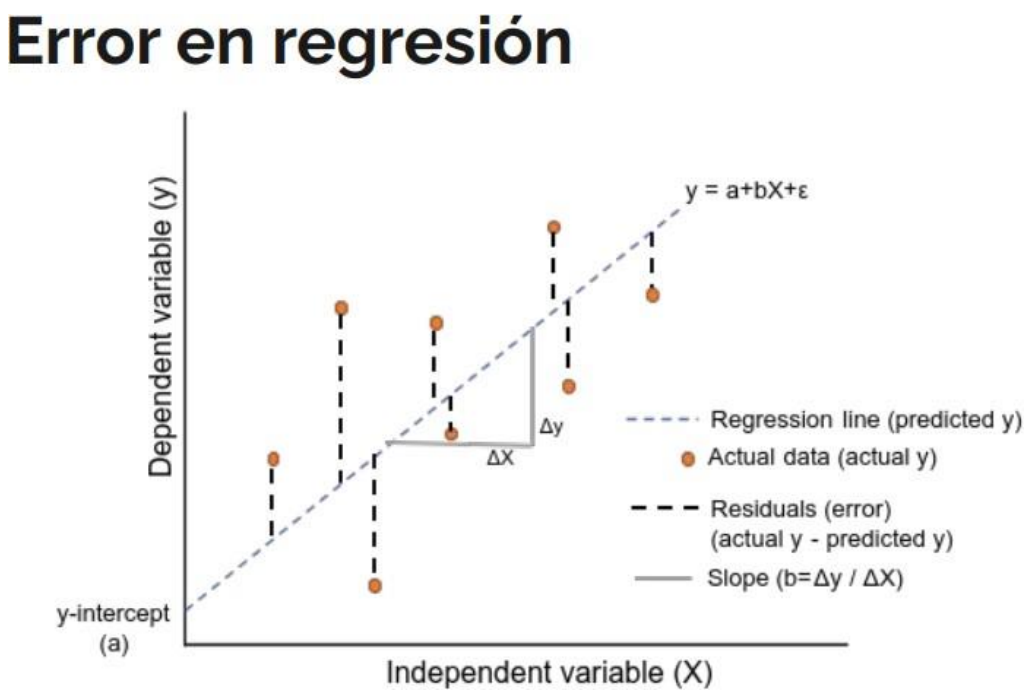
array([3.22, 0.98])

Problemas con la ecuación normal

La ecuación normal calcula el inverso de $X^T X$, que es una matriz de (n + 1) × (n + 1)(donde n es el número de características).
La complejidad computacional de invertir tal matriz es típicamente alrededor de $O(n^{2,4})$ a $O(n^3)$, dependiendo de la implementación.

En otras palabras, si se duplica el número de características, el tiempo se multiplica por aproximadamente $2^{2,4} = 5,3$ a $2^3 = 8$

Existen otros mecanismos que buscan de forma iterativa, por aproximación y son computacionalmente menos costosos como el **descenso por gradiente** → para eso necesitamos conocer el error que estamos cometiendo.



Residuo: el error que cometo en un punto. Distancia del punto a la recta

El error total es la suma de todos los residuos

Métrica para regresión

m instancias : número de instancias

h: función de hipótesis, es el modelo entrenado. En este caso regresión linealx:

todos los valores de entradas, todas las columnas

Root Mean Square Error (RMSE)

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

Raíz del error cuadrático medio

Mean absolute error (MAE)

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right|$$

Error medio absoluto

Mean Square Error

$$\text{MSE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2$$

Error cuadrático medio.

Metodo de aprendizaje supervisado → conocemos las etiquetas, los valores que queremos predecir. Vamos a poder comparar si lo estamos hacien bien.

De clasificacion: predecir si x observacion pertenece a la clase 0 o 1
Ej: persona sana o enferma, persona fuma o no fuma

De regresion: predecir algo que tiene un valor continuo
queremos predecir un valor que va a ser numérico
Ej: quiero predecir el valor de una propiedad, la popularidad de una cancion, …

El modelo se puede terminar aprendiendo de memoria los datos de entrenamiento, entonces va a dar que predice perfecto, pero al ingresar un dato nuevo no funciona. Por eso, es que guardamos datos que el modelo nunca vio para evaluar como es su performance → eso se llama capacidad de generalizacion

El modelo lineal permite usar predictoras numericas o categoricas

Lo que si es numerico la variable a predecir.