Machine Learning e Imágenes en Python

FCEFyN

Regresión

Regresión

- Cuando buscamos predecir una variable objetivo en función de otra.
- Para esto se estudia si esas variables están relacionadas (o correlacionadas: término estadístico que significa "alineadas").
- Si consideramos que hay o puede haber relación significativa una opción es intentar matematizar esa relación, crear una fórmula matemática que materialice, formalmente, esa relación y que permita calcular pronósticos de una variable a partir del conocimiento de valores de otra (u otras) en un individuo concreto.

Regresión lineal

La relación matemática determinística más simple entre dos variables ***X*** e ***Y*** es una relación lineal

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X$$

Si las dos variables no están determinísticamente relacionadas, entonces con un valor fijo de x, el valor de la segunda variable es aleatorio.

Regresión lineal

Existen parámetros β_0 y β_1 de tal forma que con cualquier valor fijo de la variable independiente x, la variable dependiente está relacionada con x según el modelo

$$Y=\beta_0 + \beta_1 x + \varepsilon$$

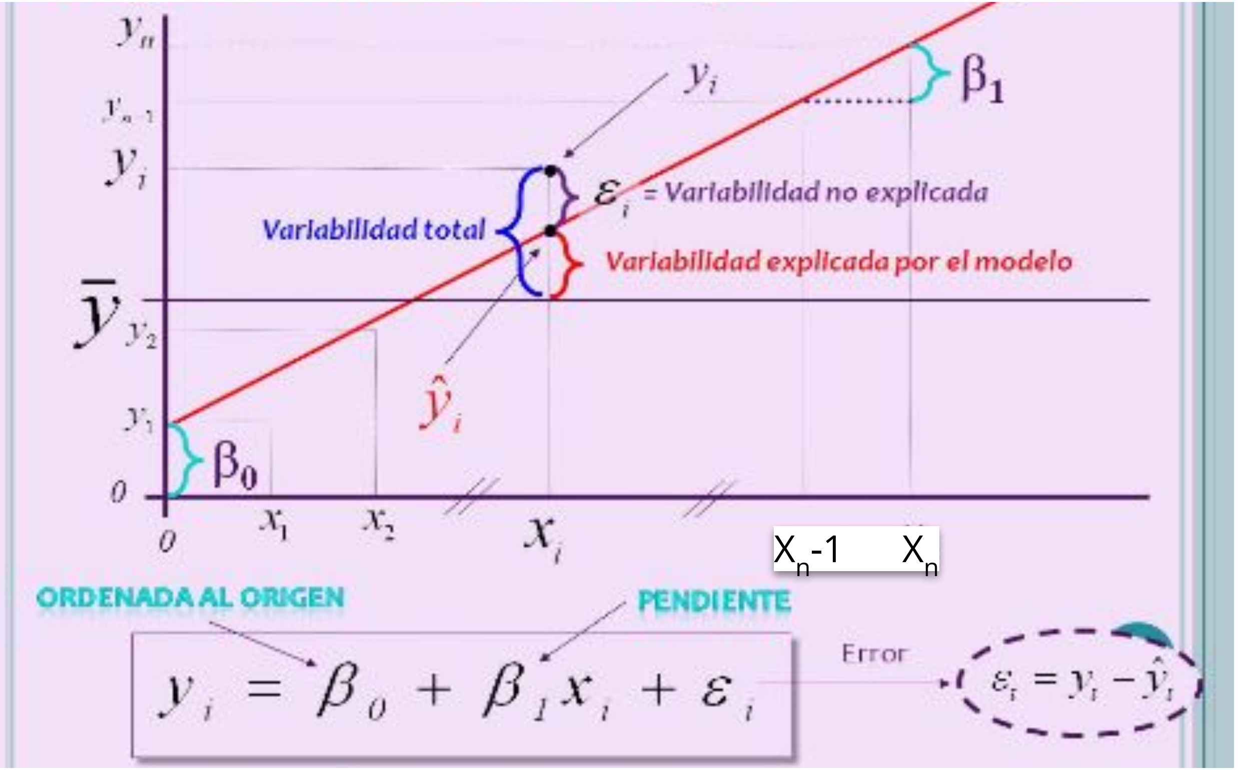
En el modelo ε es una variable aleatoria, que se supone está normalmente distribuida con $E(\varepsilon)= 0$ y $V(\varepsilon)=\sigma^2$

Regresión (lineal)

- X es denominada la variable pronosticadora, independiente o regresora.
- Y es la variable dependiente o de respuesta (target/objetivo)

- Un primer paso en el análisis de regresión que implica dos variables es construir una gráfica de puntos de los datos observados. En una gráfica como esa, cada (x_i, y_i) está representado como un punto colocado en un sistema de coordenadas bidimensional (scatter plot)

Regresión lineal



Regresión lineal

$$Y=\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x + \varepsilon$$

parámetros $\boldsymbol{\beta}_0$ y $\boldsymbol{\beta}_1$
ε variable aleatoria, que se supone normalmente distribuida con $E(\varepsilon)= 0$ y $V(\varepsilon)=\boldsymbol{\sigma}^2$

Sin considerar ε, cada observación (x, y) quedaría sobre la línea $y=\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x$, llamada línea de regresión (o de población) verdadera. La inclusión del término de error aleatorio permite que (x, y) pueda quedar o por encima de la línea de regresión verdadera (cuando $\varepsilon > 0$) o por debajo (cuando $\varepsilon < 0$).

Regresión lineal

Un investigador casi nunca conocerá los valores de $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ o $\boldsymbol{\sigma}^2$.
En cambio, estará disponible una muestra de datos compuesta de n pares observados $(x_1, y_1), \ldots, (x_n, y_n)$, con la cual los parámetros de modelo y la línea de regresión verdadera pueden ser estimados. Se supone que estas observaciones se obtuvieron independientemente una de otra.

Es decir, y_i es el valor observado de una variable aleatoria Y_i, donde

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_i + \varepsilon_i$$

y las n desviaciones, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son una m.a. de $N(0, \boldsymbol{\sigma}^2)$ (homocedasticidad)
La independencia de $Y_1, Y_2, \ldots, Y_n$ se desprende de la independencia de las ε_i

Minimizar la función de costo: Mínimos Cuadrados

Se minimiza **la suma de los cuadrados de las diferencias entre cada observado y_i y su estimado $\beta_0 + \beta_1 x_i$** .

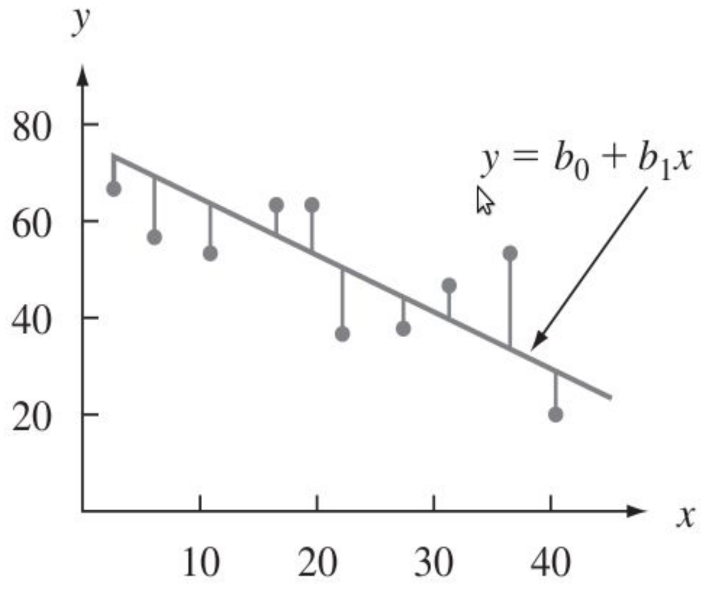
función de costo,

La **suma de cuadrados del error** (o de forma equivalente, suma de cuadrados residuales) denotada por SCE, es

$$\text{SCE} = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Esto es equivalente a minimizar el ECM. En general se busca minimizar una función de costo, cualquiera sea el tipo de regresión.

Regresión simple y múltiple (varias regresoras)



- Una regla fundamental: Cuanta mayor correlación haya entre dos variables, en la representación bidimensional, estructurada en forma de recta, los valores estarán reunidos más próximos a la recta.
- Si atendemos al números de variables independientes, distinguiremos dos tipos de Regresión: la Regresión simple y la Regresión múltiple: $Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta} \cdot \mathbf{x_i}^t + \varepsilon_i$

  Donde $\mathbf{x_i} = [x_{i1}, \ldots, x_{im}]$ y $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$

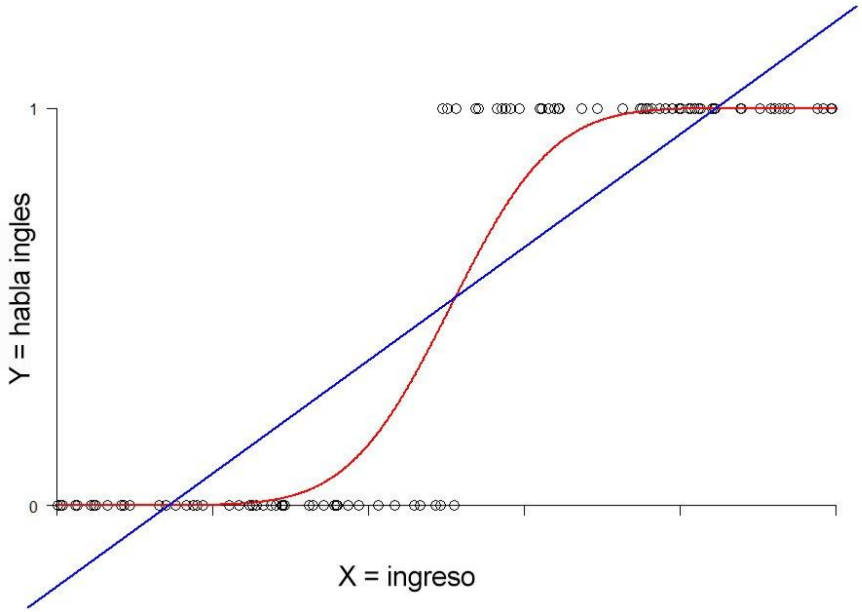
Regresión múltiple (varias regresoras)

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\,.\,\mathbf{x_i}^t + \varepsilon_i$$

Donde $\mathbf{x_i} = [x_{i1}, \ldots, x_{im}]$ y $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$

Un investigador casi nunca conocerá los valores de $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$ o $\boldsymbol{\sigma}^2$.
En cambio, estará disponible una muestra de datos compuesta de n vectores filas de datos observados $(x_{11}, \ldots, x_{1m}, y_1), \ldots, (x_{n1}, \ldots, x_{nm}, y_n)$,
n es la cantidad de observaciones.
m es la cantidad de variables regresoras. (dimensionalidad)
Los parámetros de modelo (con esto el hiperplano de regresión) pueden ser estimados.

Regresión Logística: para Y dicotómica, dos clases



Cuando la variable dependiente Y es categórica, que hacemos? Cualquiera sea la variable Y la podemos codificar en dos clases 0 y 1.

Así surge la regresión logística:

$$Y_i = logit(\boldsymbol{\beta}_0 + \boldsymbol{\beta} \, . \, \mathbf{x_i}^t) + \varepsilon_i$$

Donde $\mathbf{x_i} = [x_{i1}, \ldots, x_{im}]$ y $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$

a ***logit*** la llamamos función de enlace, hay otras...