Aprendizaje Automático

Regresión Logística

2023-2Q

Recordemos...

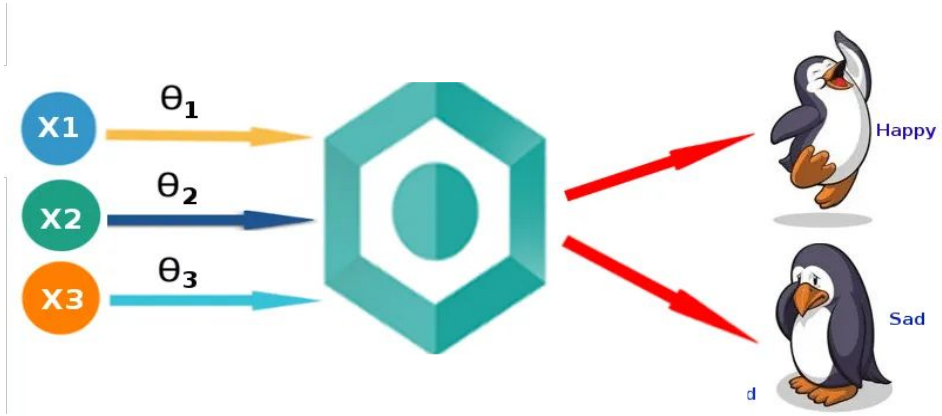
Los métodos de regresión estudian la construcción de modelos para explicar o representar la **dependencia** entre una **variable respuesta o dependiente Y** y la(s) **variable(s) explicativa(s) o independiente(s), X.**

$$Y = f(X) + e$$

Habíamos dicho que se utilizan para realizar inferencias cuando la variable respuesta no es categórica sino cuantitativa.

Regresión Logística

Es una técnica estadística que permite **predecir** el resultado de una **variable categórica** a partir de un conjunto de variables predictoras.



Regresión Logística

¿Por qué se llama "regresión" cuando se usa para clasificación?

La regresión logística se considera un método de regresión ya que su función es **predecir probabilidades**, que son resultados continuos, a partir de predictores.

Aunque en la práctica se use esta predicción para clasificar en categorías, conceptualmente sigue siendo regresión, ya que modela una variable continua como resultado.

Modelo de regresión logística

- El modelo de regresión logística se utiliza cuando la **variable respuesta es categórica** ($Y = 1$ ó $Y = 0$).
- La respuesta del modelo es una probabilidad p.
- Puede usarse para **Clasificación**: Si $p \geq 0{,}5 \Rightarrow Y = 1$ y $Y = 0$ sino.

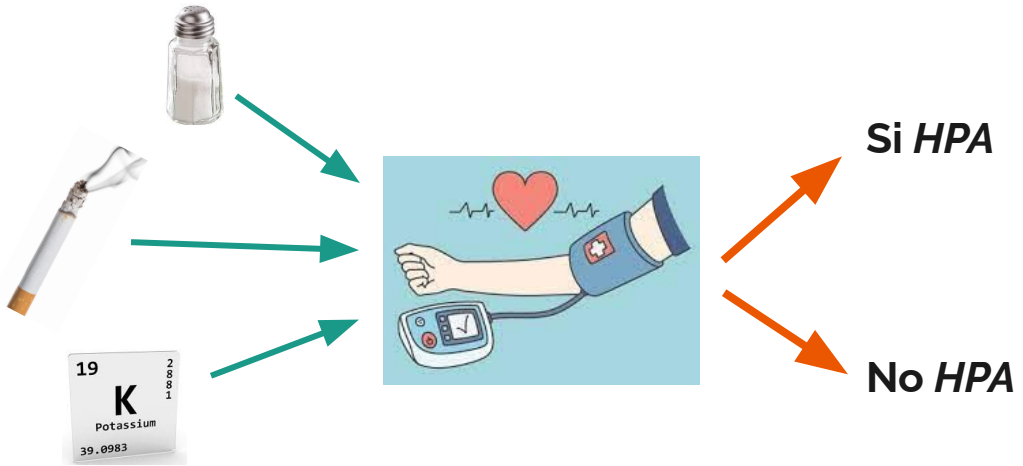
Con este modelo podemos estudiar el impacto que tiene cada una de las variables explicativas en la probabilidad de que ocurra el suceso en estudio.

Variable dicotómica

Por ejemplo, queremos **predecir la presencia o ausencia de hipertensión arterial** en función de factores como el consumo de sal o el hábito de fumar, etc.

La variable objetivo, HPA, toma dos posibles:

- *HPA* = 1 (Si)
- *HPA* = 0 (No)



Variable dicotómica

La **variable dependiente** solo puede tomar valores **0** ó **1** (fracaso o éxito) y no puede tomar cualquier valor real como ocurre con la regresión lineal simple o múltiple.

En realidad también puede tomar muchas clases pero por ahora lo dejamos.

Propósito de la regresión logística

- Predecir la probabilidad de que ocurra un evento basado en ciertas variables predictoras.

- Determinar qué variables influyen más para aumentar o disminuir la probabilidad de que suceda el evento.

Usos de la regresión logística

- Estimar la probabilidad de ocurrencia de un evento para un sujeto dado, en base a sus características.

- **Clasificación:** Asignar una categoría binaria (0-1) a cada sujeto según si su probabilidad estimada supera o no cierto umbral (por ej: 0.5)

Si $p(x1, \ldots, xn) \geq 0{,}5 \Rightarrow Y = 1$ y $Y = 0$ sino.

Planteo del problema

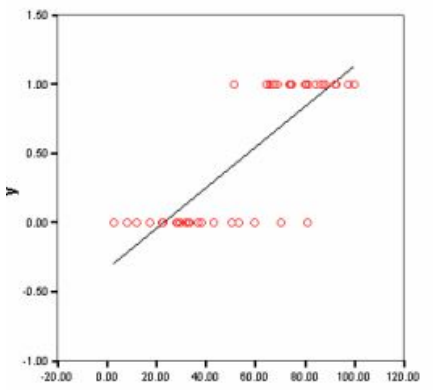
La **variable dependiente Y** (dicotómica):

- Y = 0 si NO ocurre el suceso.
- Y = 1 si ocurre el suceso.

Consideremos que existe una sola **variable explicativa X** y planteamos un modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Si estimamos este modelo y representamos gráficamente la recta de regresión. La recta de regresión no está acotada en el intervalo [0, 1].



Modelo de Regresión lineal corregido

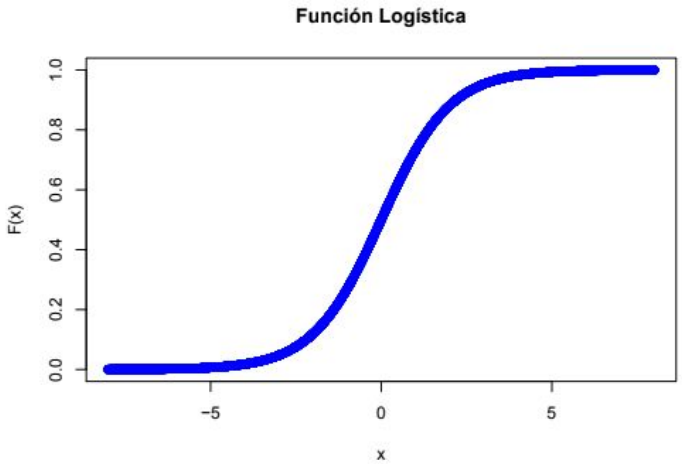
Como las funciones de distribución acumulada sí están acotadas en el intervalo [0, 1], se utiliza el modelo

$$y_i = F(\square_0 + \square_1 x_i + e_i)$$

donde F es una función de distribución.

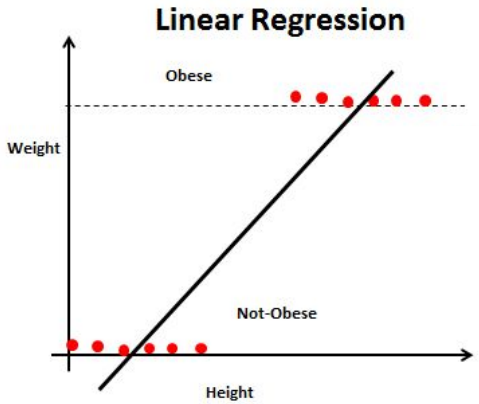
Función de Distribución Logística:

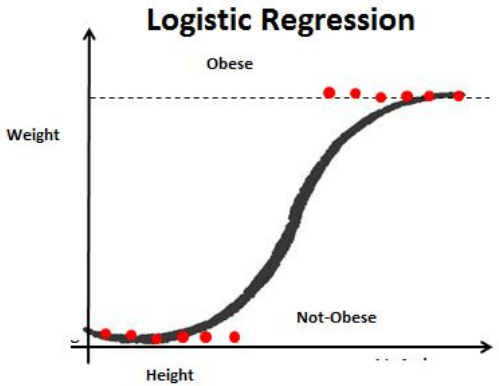
$$F(x) = \frac{e^x}{1 + e^x}$$



Modelo de Regresión lineal corregido

Cuando la función de distribución *F* es la distribución logística, el modelo se denomina *Regresión Logística*.





La idea

Suponiendo que la **variable dependiente es dicotómica** y las **variables independientes** o explicativas son todas **cuantitativas** $X_1, X_2, \ldots, X_k$ el modelo de regresión logística es:

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+\epsilon_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+\epsilon_i}}$$

donde $\boldsymbol{\beta_0, \beta_1, \beta_2 \cdots, \beta_k}$ son los **coeficientes de las covariables**.

Odds

Las **odds** en regresión logística se refieren a la razón entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra.

$$\frac{P(X)}{1 - P(X)}$$

Las usamos para comparar la influencia de las variables explicativas (o independientes) sobre la variable dependiente

Interpretación de Odds

Supongamos que lanzamos un dado y queremos calcular las odds de sacar un 6.

- P(6) = 1/6 = 0.167
- P(no 6) = 5/6 = 0.833

Entonces Odds de sacar 6 = P(6) / P(no 6) = 0.167 / 0.833 = 0.2
Significa que la probabilidad de sacar un 6 es 0.2 veces la probabilidad de NO sacar un 6.

Las odds nos permiten comparar la probabilidad de un suceso con su complementario. Mientras mayor sean las odds, mayor es la probabilidad del suceso en comparación a la probabilidad de que no ocurra.

Reemplazando la Ecuación Logística

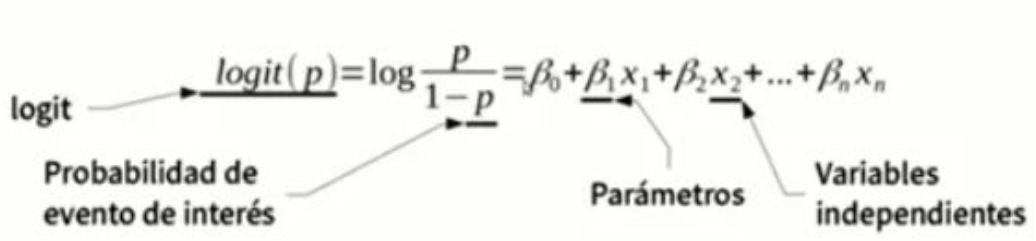
$$\frac{P(X)}{1-P(X)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+\epsilon_i}$$

y aplicando ln a ambos miembros

$$ln(\frac{P(X)}{1-P(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon_i$$

Esto es una relación lineal con la expresión llamada ***logit***.

Ecuación Logística



Esta ecuación relaciona las variables predictoras X con la probabilidad del evento a predecir p, a través de una transformación logística.

Los parámetros β deben estimarse a partir de datos, para esto se utiliza el **método de máxima verosimilitud.**

Estimación de los coeficientes modelo de Regresión Logística

El método de máxima verosimilitud construye una función de verosimilitud **L(β)**, que es el producto de las probabilidades individuales de cada observación.
Al maximizar L(β) encontramos los β que mejor ajustan el modelo a los datos observados. Para esto se deriva L(β) e iguala a 0 obteniendo las ecuaciones que permiten estimar los β por máxima verosimilitud.

$$\mathcal{L}(\beta_0, \ldots, \beta_n) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j))$$

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$$

log-verosimilitud

El sistema que se genera derivando con respecto a cada β_j es no lineal y no tienen solución explícita, entonces hay que resolverlo numéricamente.

Interpretación de parámetros

- El signo de β_1 indica el sentido del cambio en la probabilidad respecto a los cambios en X.
  - Si $\beta_1 > 0$: cuando X aumenta, el log-odds aumenta, por lo tanto la probabilidad de Y=1 aumenta.
  - Si $\beta_1 < 0$: cuando X aumenta, el log-odds disminuye, por lo tanto la probabilidad de Y=1 disminuye.

- Si $\beta_1 = 0$ entonces Y no depende de X y se interpreta como que la variable Y es independiente de X.
- $\beta 0$ es el término independiente, indica la probabilidad base cuando la variable predictora X es 0.

Observar

- Si $p \geq 0{,}5 \Rightarrow Y = 1$ entonces
- Pero si $p \geq 0{,}5 \Rightarrow p/(1-p) \geq 1 \Rightarrow \ln(p/(1-p)) \geq 0 \Rightarrow \text{logit}(p) \geq 0$, entonces
- Si $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \geq 0 \Rightarrow Y = 1$

$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = 0$ es un hiperplano en el espacio de predictores que separa las clases.

Por lo tanto, la regresión logística puede interpretarse como un **clasificador lineal** determinado por el signo del logit, equivalente a la combinación lineal de las X.

Ejemplo: Personas que tienen cáncer de seno

Predicción de la gravedad del cáncer de seno utilizando regresión logística:

Y = Gravedad del tumor (0 = No grave, 1 = Grave)

X = Tamaño del tumor en mm

Modelo logit:

$$\text{logit}(P(Y=1)) = \ln(P(Y=1)/P(Y=0)) = \beta_0 + \beta_1 * X$$

Donde:

- P(Y=1): Probabilidad de que el tumor sea grave
- P(Y=0): Probabilidad de que el tumor no sea grave
- X: Tamaño del tumor (variable predictora)
- β_0: Intercepto, probabilidad base cuando X=0
- β_1: Efecto del tamaño del tumor sobre el log-odds

Ejemplo: Personas que tienen cáncer de seno

Utilizando el método de máxima verosimilitud sobre los datos, se obtienen estimaciones para β0 y β1. Por ejemplo: **β0 = -2.5** y **β1 = 1.2**

Modelo estimado: $\text{logit}(P(Y=1|X)) = -2.5 + 1.2^{*}X$

Interpretación:

- β1 es positivo, luego a mayor tamaño del tumor X, mayor probabilidad de que sea grave.
- β1=1.2 indica que por cada mm de aumento en X, el log-odds aumenta en 1.2.
- β0 representa la probabilidad inicial de gravedad antes de considerar el tamaño del tumor. Un valor negativo refleja que a tamaño 0 es más probable la no gravedad.

Predicción:

- Para un tumor de tamaño X, se calcula el logit y la probabilidad de gravedad P(Y=1|X).
- Si P(Y=1|X) > 0.5 se predice que el tumor es grave.

¡Fin! ¿Alguna pregunta?