MINERÍA DE DATOS

Maximiliano Ojeda

muojeda@uc.cl

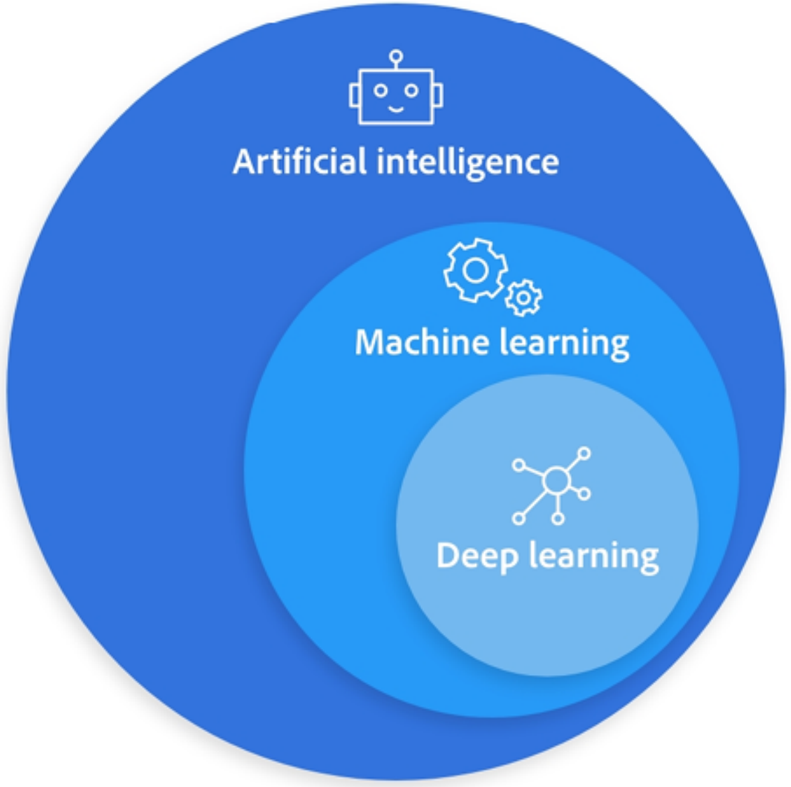
IIC-2433

Aprendizaje de Máquinas

Machine Learning

Darle a los computadores la habilidad de realizar una actividad, sin programarlos explícitamente.

La **minería de datos** y el **aprendizaje de máquina** se traslapan y no tienen límites claros



Machine Learning

Programación Explícita

- El programador escribe **reglas explícitas** para que la máquina resuelva un problema.
- Es decir, se le dice exactamente qué hacer paso a paso
- No “aprende” más allá de lo que se le pide

Kasparov vs Deep Blue (1997)

Machine Learning

Aprendizaje de Máquinas

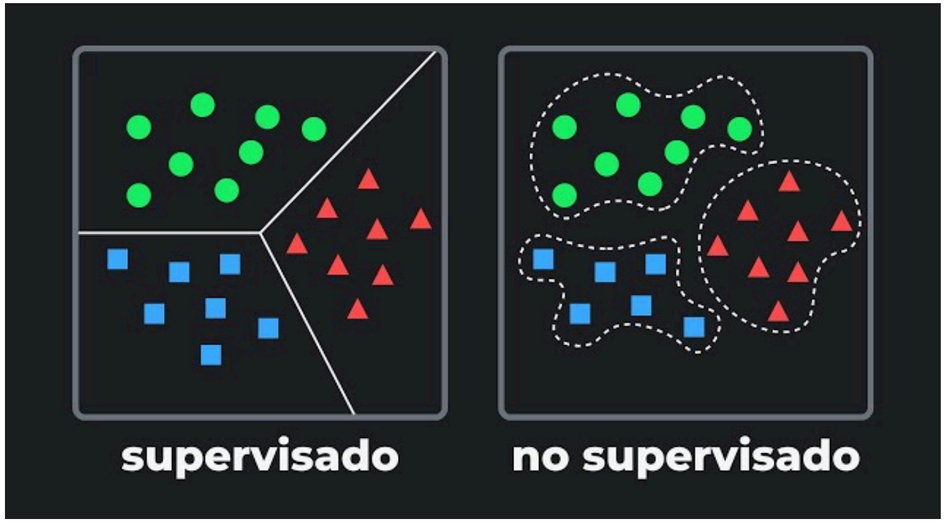
- En lugar de darle reglas, se le entrega ejemplos (datos) y un algoritmo que aprende patrones a partir de ellos.
- La máquina crea su propio modelo de reglas internas
- El modelo aprende automáticamente qué palabras, estructuras o patrones indican spam, incluso cosas que un humano no pensó.

Lee Sedol vs AlphaGo (2016)

Machine Learning

Aprendizaje Supervisado

- Clasificación: predecir **clases** o **etiquetas**
  - Ejemplo: detectar si un correo es spam o no spam.
- Regresión: predecir **valores numéricos**
  - Ejemplo: estimar el precio de una casa.



Aprendizaje No Supervisado

- Clustering: descubrir **grupos ocultos** en los datos.
  - Ejemplo: segmentar clientes según hábitos de compra.

Regresión Lineal

Regresión Lineal

- Técnica estadística donde se trata de **ajustar parámetros de una función lineal** sobre un conjunto de datos
- Se busca **predecir el valor de una variable dependiente** cuantitativa (predicha) utilizando variables independientes (predictores)
- Finalmente, queremos determinar cómo afecta nuestra variable independiente a la dependiente

$$Y = \alpha + \beta X$$

Regresión Lineal

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-114.31	34.19	15.0	5612.0	1283.0	1015.0	472.0	1.4936	66900.0
1	-114.47	34.40	19.0	7650.0	1901.0	1129.0	463.0	1.8200	80100.0
2	-114.56	33.69	17.0	720.0	174.0	333.0	117.0	1.6509	85700.0
3	-114.57	33.64	14.0	1501.0	337.0	515.0	226.0	3.1917	73400.0
4	-114.57	33.57	20.0	1454.0	326.0	624.0	262.0	1.9250	65500.0
...	...	...	...	...	...	...	...	...	...
16995	-124.26	40.58	52.0	2217.0	394.0	907.0	369.0	2.3571	111400.0
16996	-124.27	40.69	36.0	2349.0	528.0	1194.0	465.0	2.5179	79000.0
16997	-124.30	41.84	17.0	2677.0	531.0	1244.0	456.0	3.0313	103600.0
16998	-124.30	41.80	19.0	2672.0	552.0	1298.0	478.0	1.9797	85800.0
16999	-124.35	40.54	52.0	1820.0	300.0	806.0	270.0	3.0147	94600.0

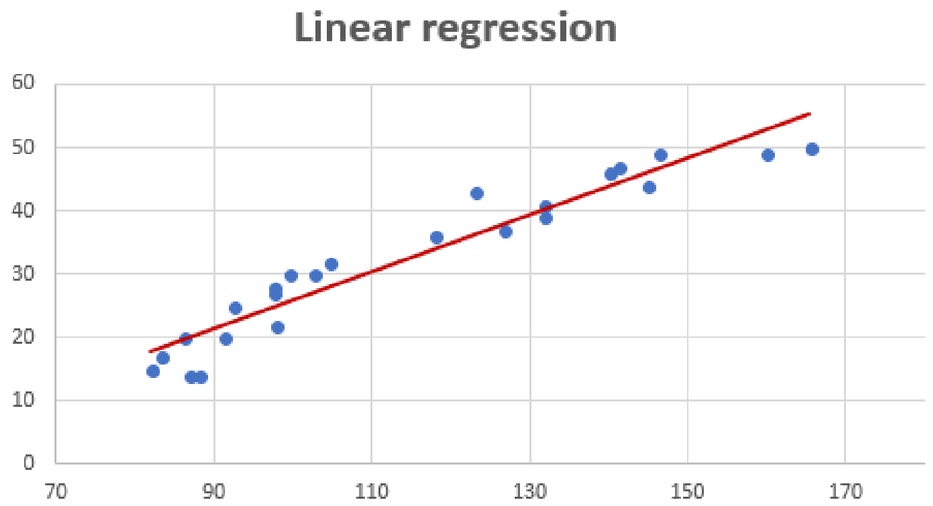
X

Y

Regresión Lineal

- Dada una tabla con un conjunto de atributos numéricos $x_1, x_2, \ldots, x_n$
- Se busca predecir un atributo numérico y
- Asumimos que esta tabla representa una función $f : \mathbb{R}^n \to \mathbb{R}$
- Y que dicha función, es una función lineal, es decir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$



Regresión Lineal

Relación lineal

- La relación entre la variable dependiente y las independientes debe ser aprox lineal.

Independencia de los errores

- Los errores (residuos) deben ser independientes entre sí

Homoscedasticidad (varianza constante)

- La dispersión de los errores debe ser aproximadamente la misma a lo largo de todos los valores de *X*.

Normalidad de los errores

- Los residuos deben seguir aproximadamente una distribución normal.

No multicolinealidad (en regresión múltiple)

- Cuando hay varias variables $x_1, x_2, \ldots x_n$ estas no deben estar fuertemente correlacionadas entre sí.

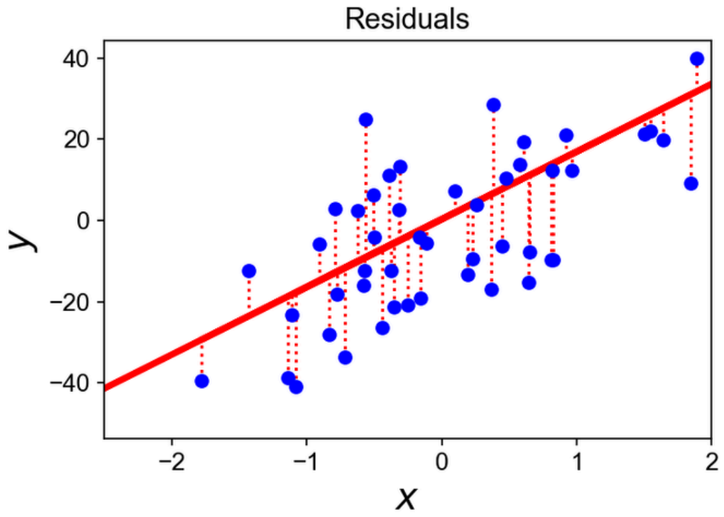
Ordinary Least Squares

La recta no pasará exactamente por todos los puntos

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$$

El método de mínimos cuadrados busca la recta que minimiza la suma de los errores al cuadrado

$$\text{SSE} = \min \sum_{i=1}^{n} e_i^2 \;=\; \min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Ordinary Least Squares

Queremos minimizar la suma de errores al cuadrado (SSE):

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Minimizando vamos a llegar a esto (en la siguiente diapositiva veremos derivarlo):

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ordinary Least Squares

Derivada respecto a β_0

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum y_i = n\beta_0 + \beta_1 \sum x_i$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Derivada respecto a β_1

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

Sustituimos $\beta_0 = \bar{y} - \beta_1 \bar{x}$

$$\sum x_i y_i = (\bar{y} - \beta_1 \bar{x}) \sum x_i + \beta_1 \sum x_i^2$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Ordinary Least Squares

Lo anterior se puede resumir algebráciamente en

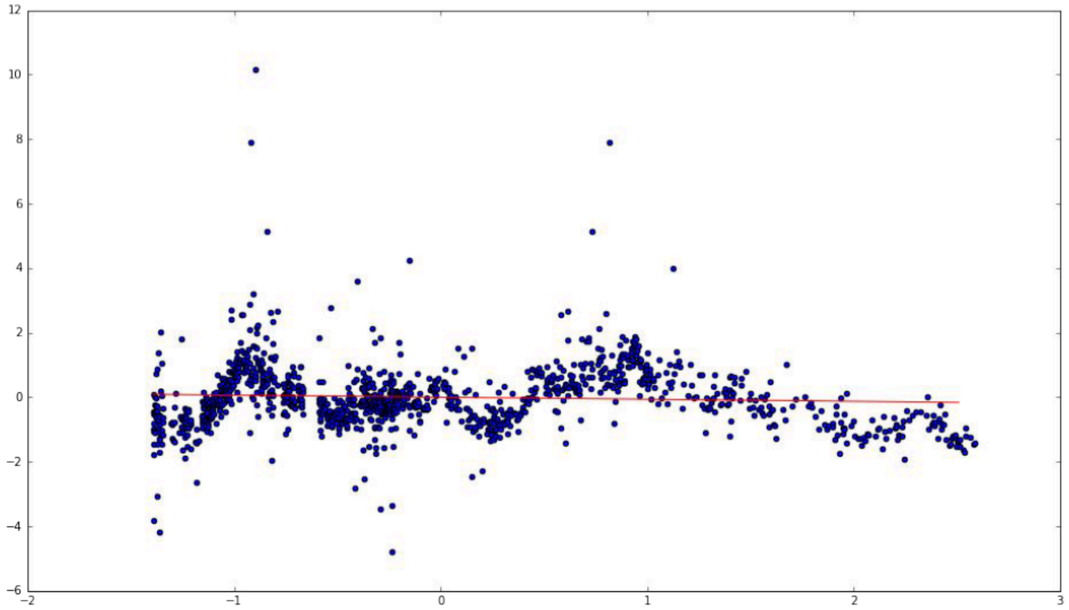
$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

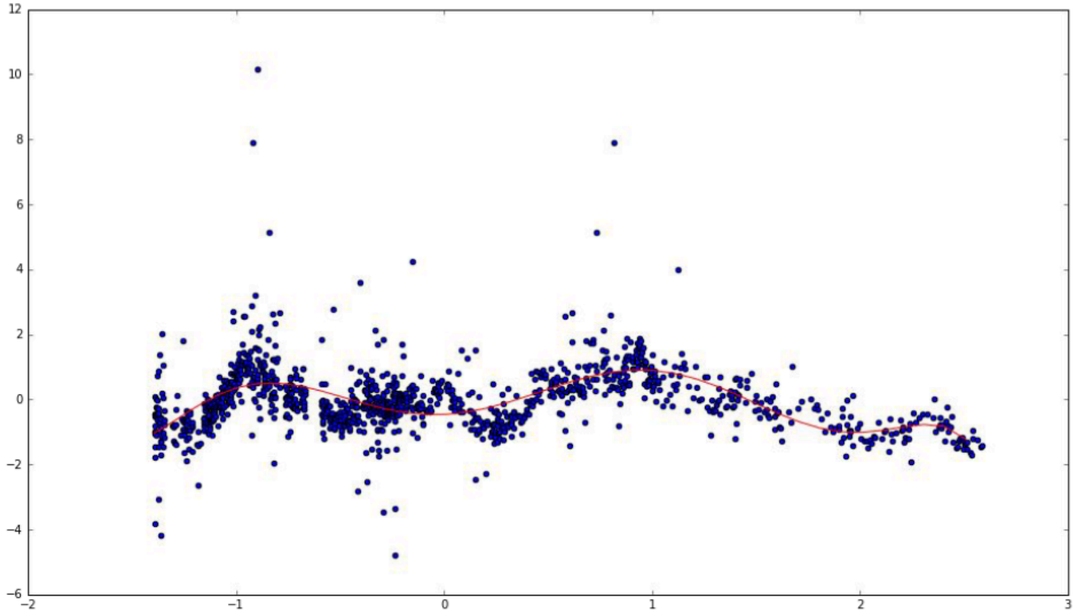
donde,

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d)$$

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)$$

Regresión Polinómica





Regresión Polinómica

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_d x^d + \varepsilon$$

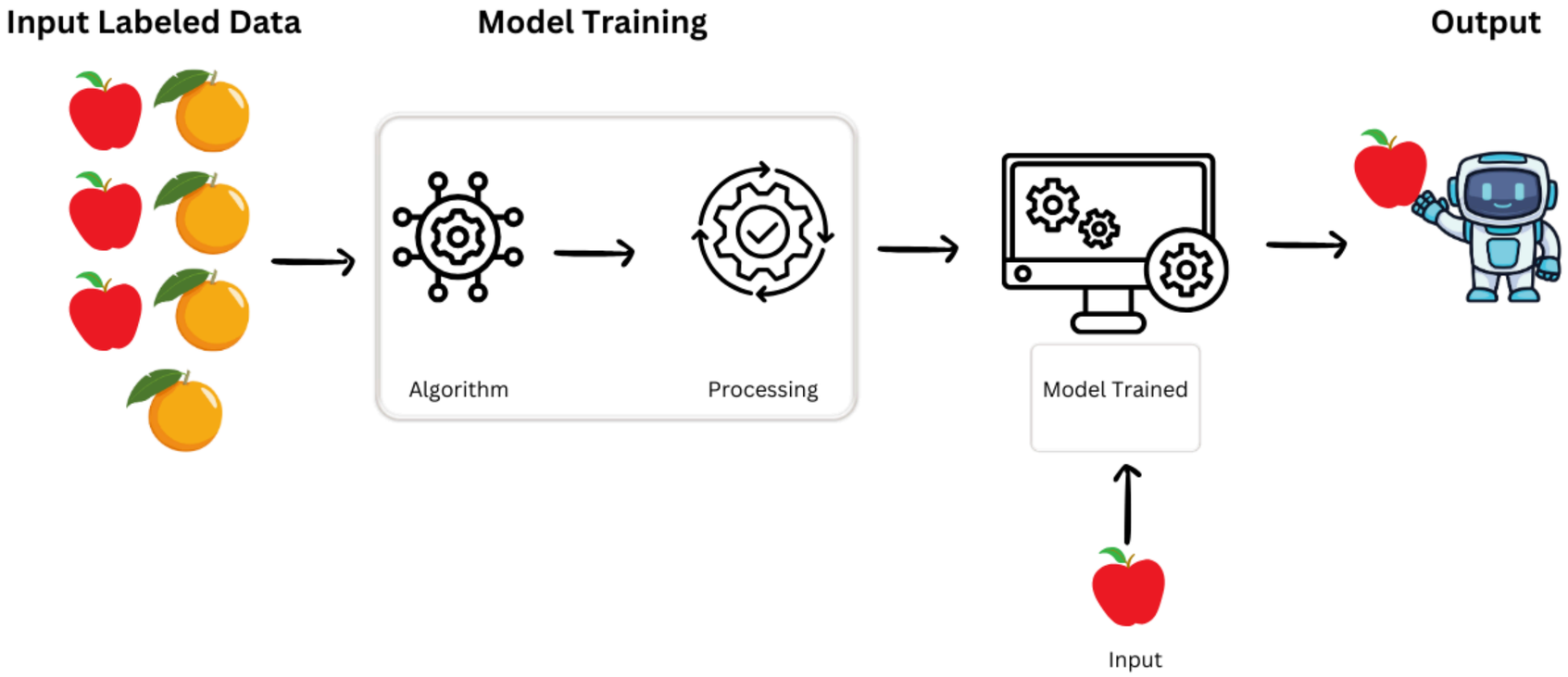
$$X_{poly} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^d \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (X_{poly}^\top X_{poly})^{-1} X_{poly}^\top y$$

- Crea una matriz con todas las potencias de X
- Luego aplica exactamente el mismo método de OLS que en la regresión lineal clásica
- El “secreto” es que el modelo sigue siendo lineal en los parámetros

Regresión Logística

Aprendizaje Automático: Clasificación



Regresión Lineal

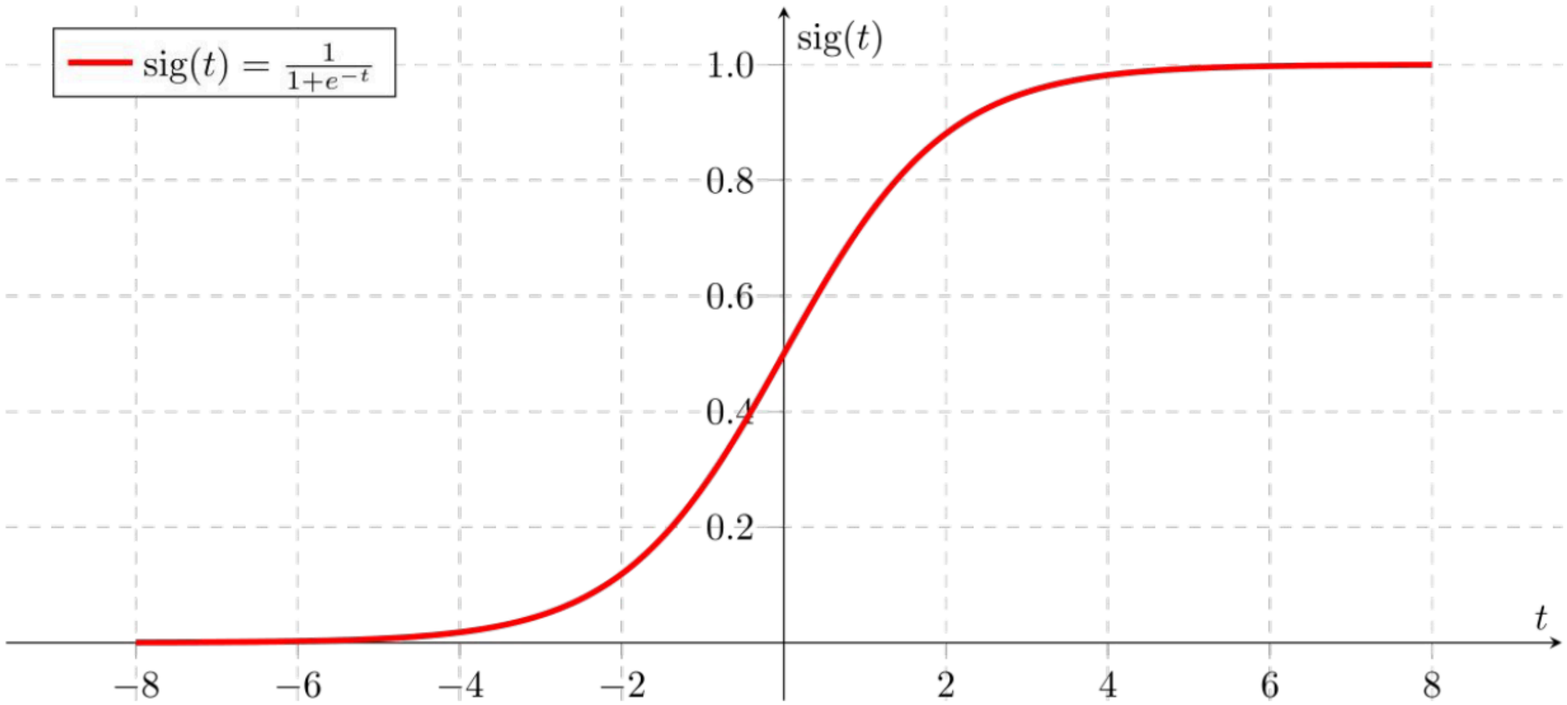
- La regresión lineal predice valores **numéricos continuos,** pero muchas veces queremos **predecir categorías** (spam/no spam, 0/1)
- No podemos usar una recta cualquiera porque podría dar valores fuera de [0,1]

$$Y = \alpha + \beta X$$

La regresión logística está pensada para variables categóricas binarias (0 o 1).

Regresión Logística

No podemos usar una recta cualquiera porque podría dar valores fuera de [0,1]. entonces usamos una función que “apriete” los valores a ese rango



Regresión Logística

La probabilidad de que *y=1* dado *x* se modela como:

$$P(y = 1 \mid x) \;=\; \sigma(\beta_0 + \beta_1 x) \;=\; \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

En general, con varias variables:

$$P(y = 1 \mid x) \;=\; \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

Regresión Logística

En regresión lineal usamos **mínimos cuadrados**, en regresión logística usamos **máxima verosimilitud**

$$L(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i)$$

o en su forma logarítmica

$$\ell(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} \left[\, y_i \log(\hat{p}_i) \;+\; (1 - y_i) \log(1 - \hat{p}_i) \,\right]$$

Regresión Logística

Una vez calculada la probabilidad, se decide:

$$\hat{y} = \begin{cases} 1, & \text{si } P(y = 1 \mid \mathbf{x}) \geq 0.5, \\ 0, & \text{si } P(y = 1 \mid \mathbf{x}) < 0.5. \end{cases}$$