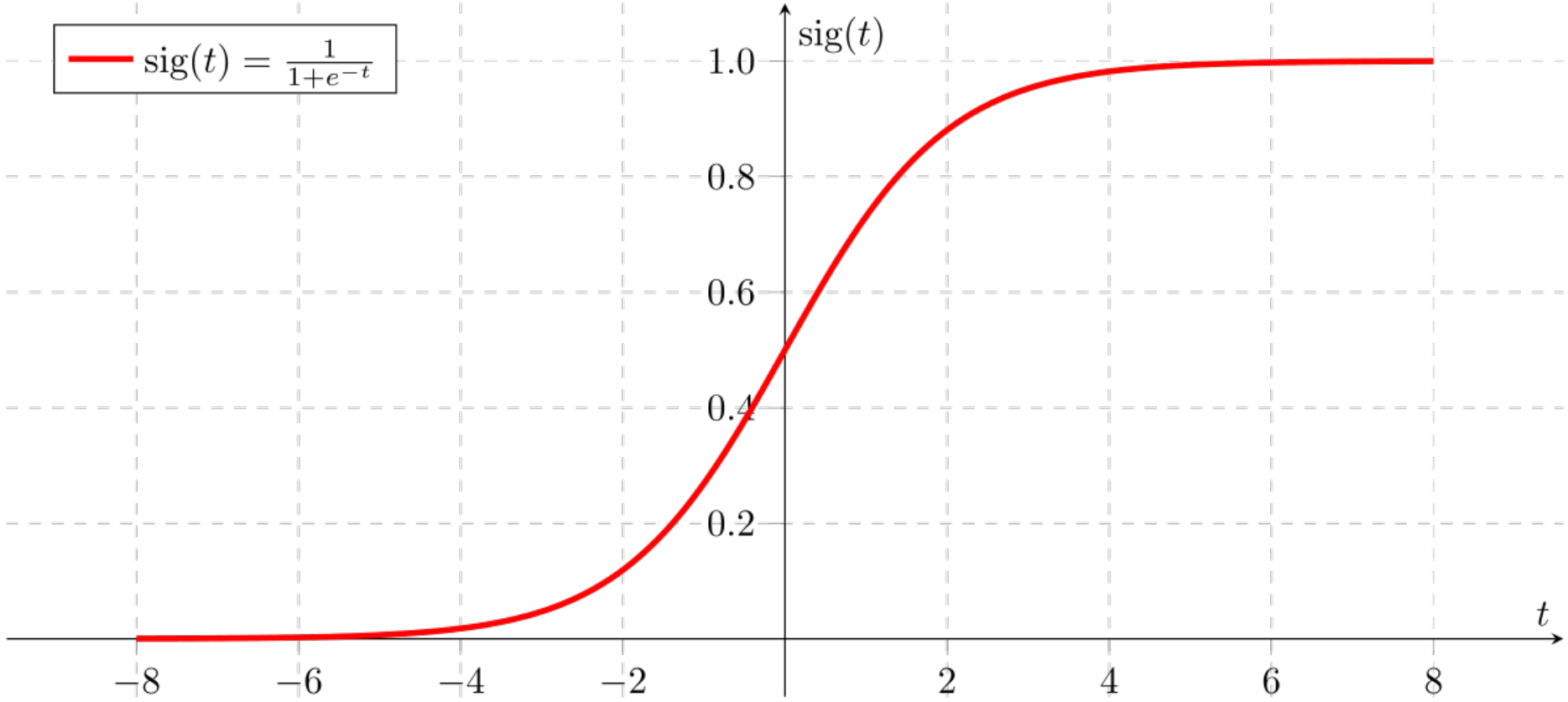
ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

IIC2433 Minería de Datos
Regresión logística

Profesor: Mauricio Arriagada

REGRESIÓN LOGÍSTICA



https://miro.medium.com/max/1400/1*RqXFpiNGwdiKBWyLJc_E7g.png

Idea general

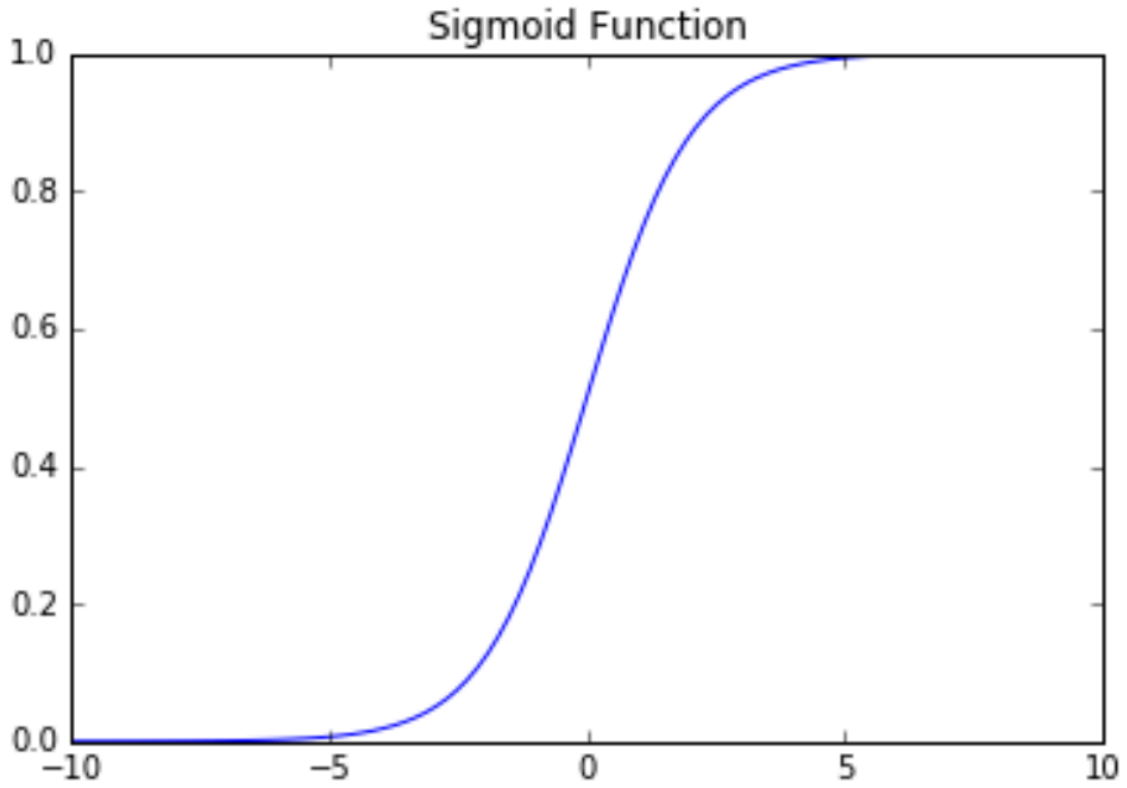
- Modelamos la salida del clasificador deseado como una distribución de probabilidad.

class(X) = 1 → P(class = 1 | X)

Class(X) = 0 → P(class = 0 | X)

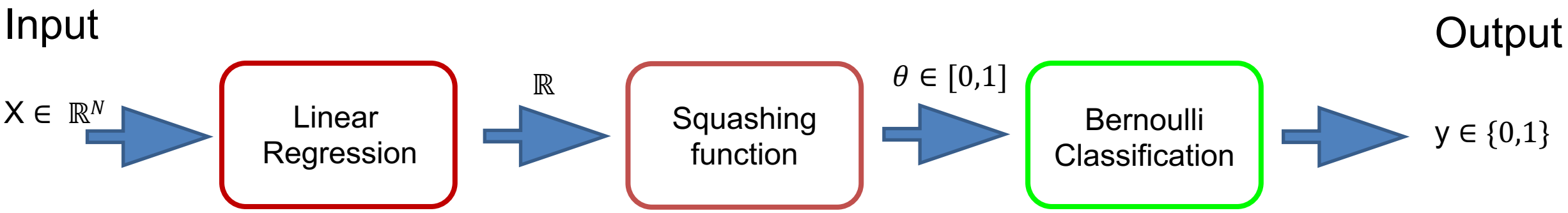
Regresión logística para clasificación binaria

- Si reducimos la salida de una regresión lineal al intervalo [0,1], podríamos usar esa salida como P (Y = y)



Regresión logística para clasificación binaria

- Si reducimos la salida de una regresión lineal al intervalo [0,1], podríamos usar esa salida como P (Y = y)

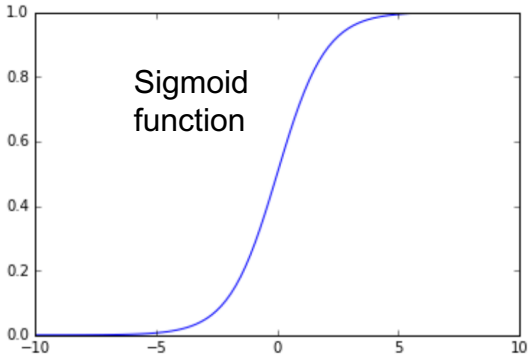


Regresión logística para clasificación binaria

- Si reducimos la salida de una regresión lineal al intervalo [0,1], podríamos usar esa salida como P (Y = y)

Linear Regression → $\mathbb{R}$ → Squashing function → $\theta \in [0,1]$ → Bernoulli Classification

$$\beta_0 + \sum_{j=1}^{d} \beta_j x_j$$



$$f_{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

$$P(Y = y|\theta) = \theta^y (1 - \theta)^{1-y}$$

Regresión logística para clasificación binaria

- Otro punto de vista: estamos haciendo una regresión sobre las probabilidades (log odds)
- Log odds: log de la proporción de obtener un "éxito" (codificado como 1) sobre obtener "fracaso" (codificado como 0)
- Entonces, lo que realmente estamos haciendo es una regresión en log odds:

$$\log \frac{\theta}{(1-\theta)} = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

$$\theta = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{d} \beta_j x_j)}} \quad \text{(sigmoid)}$$

Ejemplo

- Probalidad de pasar un examen de acuerdo a la cantidad de horas de estudio

Un grupo de 20 estudiantes pasa entre 0 y 6 horas estudiando para un examen. ¿Cómo afecta la cantidad de horas dedicadas al estudio a la probabilidad de que el alumno apruebe el examen?

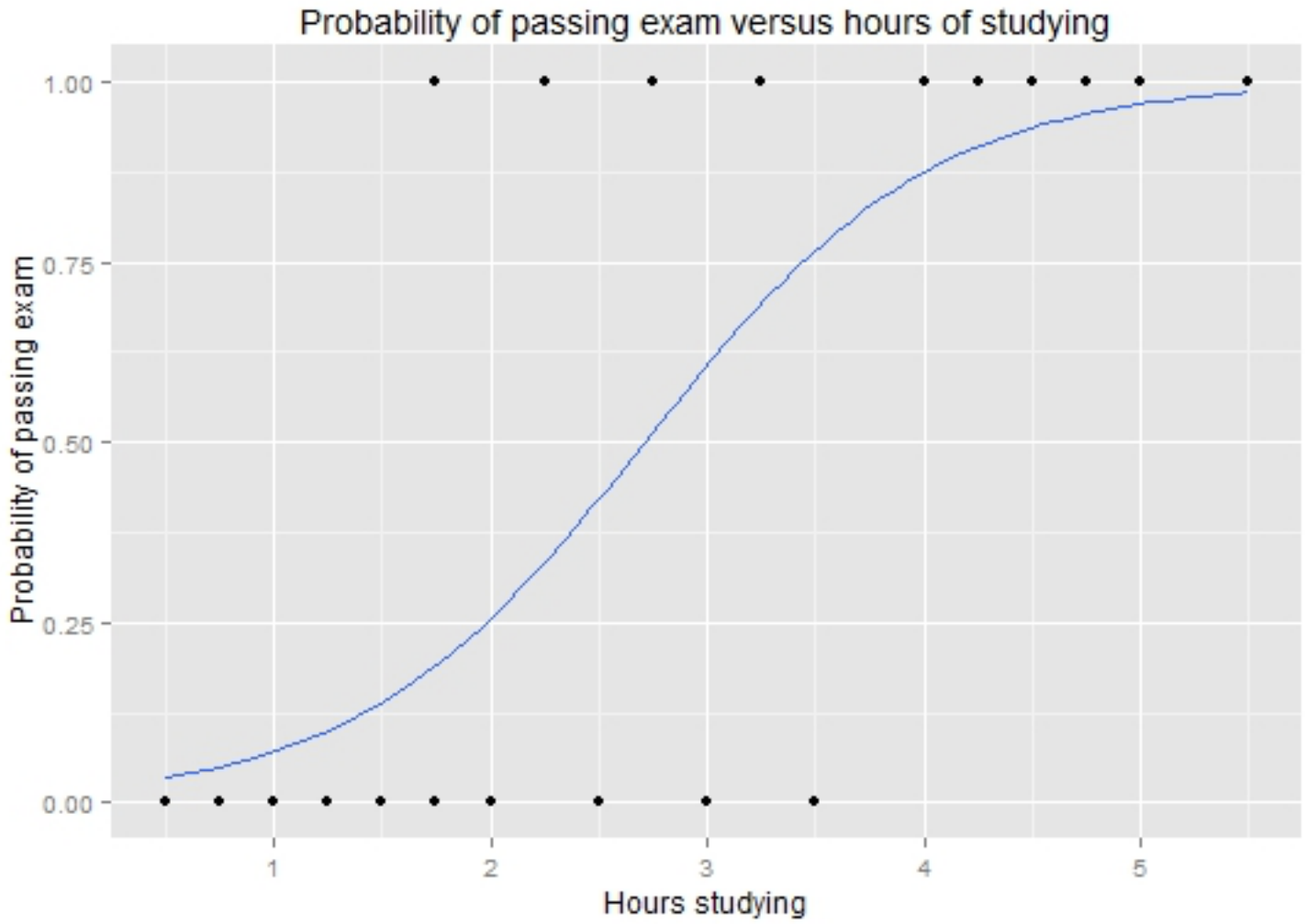
Ejemplo

- Usamos regresión logística porque el resultado de la variable dependiente tendrá 2 posibles valores: aprobar o reprobar
- Estos resultados se representan con los valores 1 y 0
- Datos:

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

- Donde la variable Pass tiene 2 valores: 1 (aprueba) y 0 (reprueba)

Ejemplo


Probability of passing exam versus hours of studying
Probability of passing exam
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Hours studying

Ejemplo

- El resultado indica que las horas de estudio se asocian significativamente con la probabilidad de aprobar el examen (p=0.0167)
- Además, los coeficientes intercept = -4.0777 y hora = 1.5046

	Coefficient	Std.Error	z-value	P-value (Wald)
Intercept	–4.0777	1.7610	–2.316	0.0206
Hours	1.5046	0.6287	2.393	0.0167

Ejemplo

$$Log - odds\ de\ pasar\ el\ examen = 1.5046 * Hours\ - 4.0777 = 1.5046\ * (Hours\ - 2.7)$$

$$odds\ de\ pasar\ el\ examen = \exp(1.5046 * Hours\ - 4.0777) = \exp(1.5046\ * (Hours\ - 2.7))$$

$$probabilidad\ de\ \ pasar\ el\ examen = \frac{1}{1 + \exp(-(1.5046 * Hours\ - 4.0777))}$$

Ejemplo

- Si un estudiante estudia 2 horas, reemplazamos este valor en la variable Hours = 2 en la ecuación:

$$probabilidad\ de\ \ pasar\ el\ examen = \frac{1}{1+\exp(-(1.5046*2\ -4.0777))} = 0.26$$

tiene una probabilidad de 0.26 de pasar el examen si estudia 2 horas

- Si un estudiante estudia 4 horas, reemplazamos este valor en la variable Hours = 4 en la ecuación:

$$probabilidad\ de\ \ pasar\ el\ examen = \frac{1}{1+\exp(-(1.5046*4\ -4.0777))} = 0.87$$

tiene una probabilidad de 0.876 de pasar el examen si estudia 4 horas

REFERENCIAS

- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- Hand, D. J. (2006). Data Mining. *Encyclopedia of Environmetrics*, *2*.
- Sainani, Kristin L. "Logistic Regression." *PM&R* 6.12 (2014): 1157-162. Web.