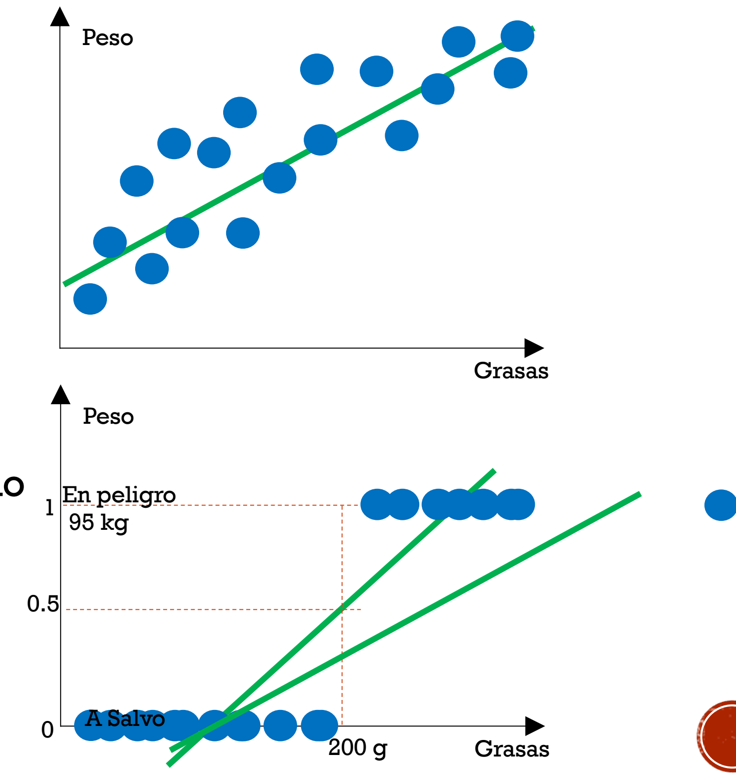
REGRESIÓN LOGÍSTICA

UNIVERSIDAD ICESI

¿REGRESIÓN LINEAL PARA CLASIFICACIÓN?

Ejemplo: tenemos los datos que relacionan la cantidad de grasas consumidas y el peso de las personas → Regresión

- Si un doctor estima que más de 95kg implica riesgo de diabetes, el problema se convierte en uno de clasificación: 0=a salvo, 1=en peligro
- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes
- No se puede interpretar estas predicciones como probabilidades (valores no están en [0;1])
- Poco robusto.



UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

- Algoritmo de **clasificación**, no de **regresión**
- Similar a la regresión lineal pero su resultado es modificado para poder obtener una salida **binaria**: sólo permite distinguir entre 2 clases.
  - Churn vs. Stay
  - Compra vs. No compra
  - Cliente valioso vs. Cliente no valioso
- Se agrega una transformación del resultado de la regresión lineal a partir de una función de distribución acumulativa logística, también conocida como función **logit** o **sigmoide**.

$$f(\boldsymbol{z}) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

- El modelo pasa de:

$$h_\Theta(\boldsymbol{X}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

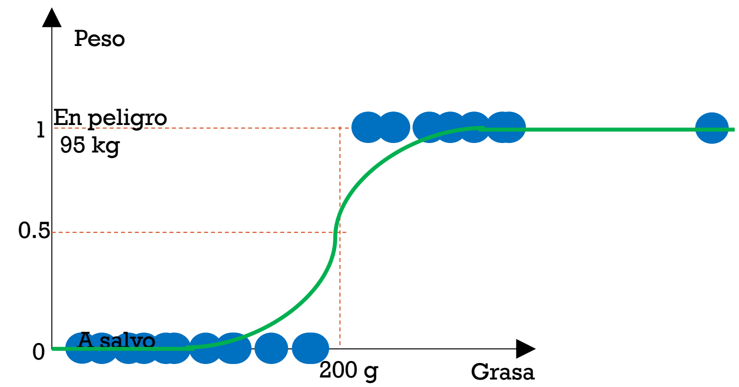
a $h_\Theta(\boldsymbol{X}) = \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{\sigma}(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)$,

con max(f(z))=1 y min(f(z))=0

- $\boldsymbol{\sigma}(z)$ es la función **sigmoide** o **logística**
- Se pueden interpretar los valores de $\boldsymbol{\sigma(z)}$ como **probabilidades** de que una instancia con atributos $\boldsymbol{X}$ pertenezca a la clase Y=1:

$$P(\boldsymbol{Y = 1} | x_1, \ldots, x_n) = p_1(\boldsymbol{X}) = \boldsymbol{\sigma}(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)$$

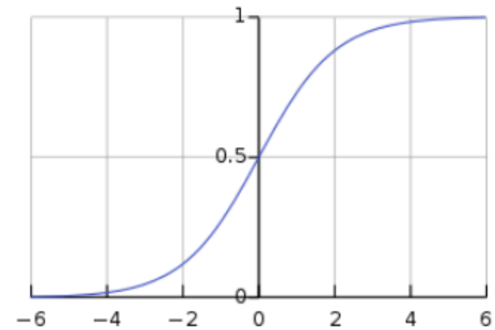
$$p_1(\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{\Theta}^T \boldsymbol{X}}}$$



REGRESIÓN LOGÍSTICA

- Comportamiento:
  - Si y=1, queremos que $p_1(\boldsymbol{X}) \approx \mathbf{1}$, luego $\boldsymbol{\Theta}^T\boldsymbol{X} \gg \mathbf{0}$
  - Si y=0, queremos que $p_1(\boldsymbol{X}) \approx \mathbf{0}$, luego $\boldsymbol{\Theta}^T\boldsymbol{X} \ll \mathbf{0}$
- Predicción: se establece un valor de umbral, por ejemplo 0.5
  - Predecir clase 1 si $p_1(\boldsymbol{X}) \geq 0.5$, cuando $\boldsymbol{\Theta}^T\boldsymbol{X} \geq 0$
  - Predecir clase 0 de otra manera
- Se puede establecer un umbral diferente si se quiere ser mas o menos robusto en la clasificación

$$p_1(\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{\Theta}^T\boldsymbol{X}}}$$

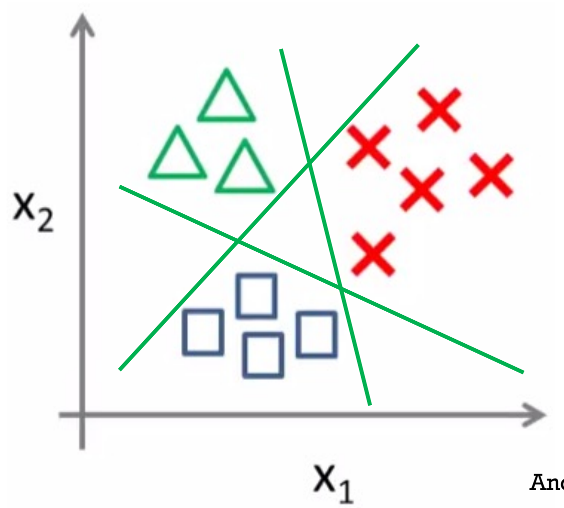


Wikipedia, 2019

REGRESIÓN LOGÍSTICA

¿Qué se puede hacer si se tienen más de 2 clases?

- Para problemas de clasificación con más de 2 clases, es necesario utilizar una aproximación de **1 vs. todos**
- Un clasificador por regresión logística es necesaria para cada clase
- Para una nueva instancia, la clase con la mayor probabilidad en su propio modelo es predicha



Andrew Ng

→ También se puede hacer regresión logística **multinomial** con la función **softmax**

UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

- Consideraciones
  - Produce estimación de probabilidades
  - No hay parámetros a afinar, solo las variables independientes a considerar.
  - Permite variables independientes numéricas y categóricas
  - Estimación de parámetros eficiente computacionalmente
  - No se ve afectado por situaciones de multicolinealidad leves. Casos importantes se pueden resolver con una regularización L2.
  - Se puede utilizar descenso de gradiente para encontrar los parámetros (mismas ecuaciones de actualización de parámetros que para regresión lineal, cambiando la función de predicción)
  - No es ideal en casos de muchas variables categóricas
  - No es muy flexible (lineal) aunque se puede extender polinómicamente.