

- Entregado el Código de Transformada de Fourier
- Energía, Zero Crossing Rate, Espectrograma, Librosa en general + audios Snore/Vicks.

(clase 5) : Problemas de Clasificación

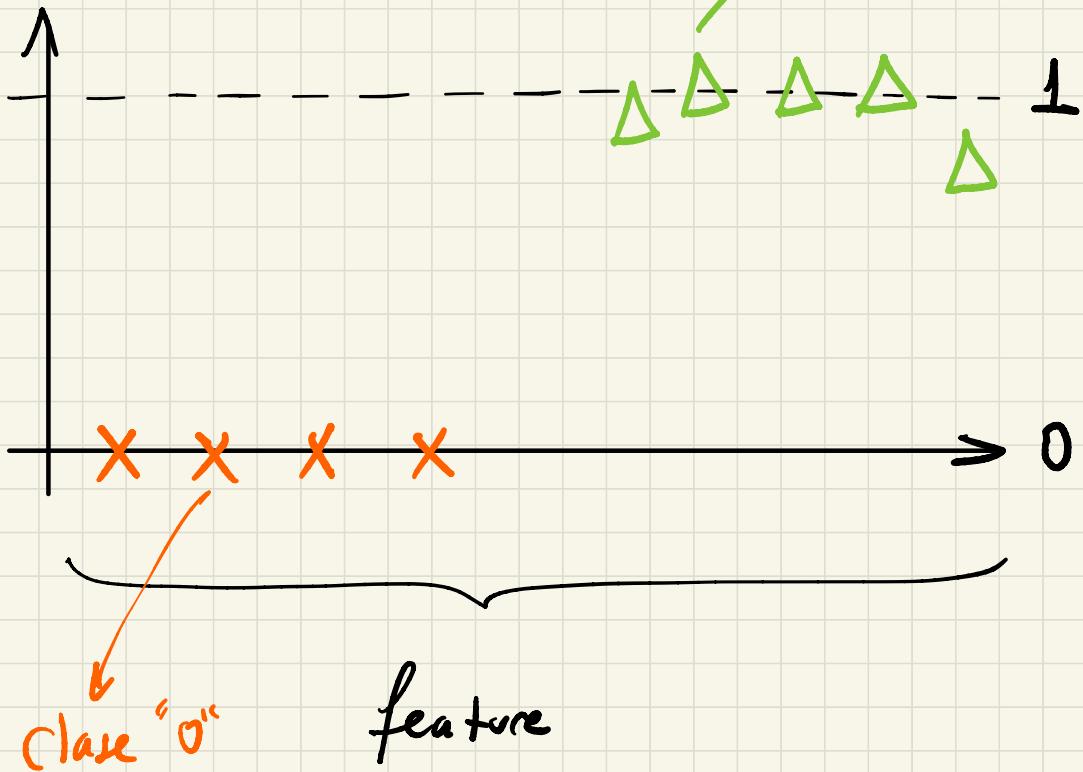
Regresión Logística

Motivación : Supongamos queremos clasificar objetos como siendo A o B
(Análogamente : $\{0, 1\}$; $\{-1, 1\}$ etc)

Esto se denomina clasificación binaria.

Ilustremos la situación :

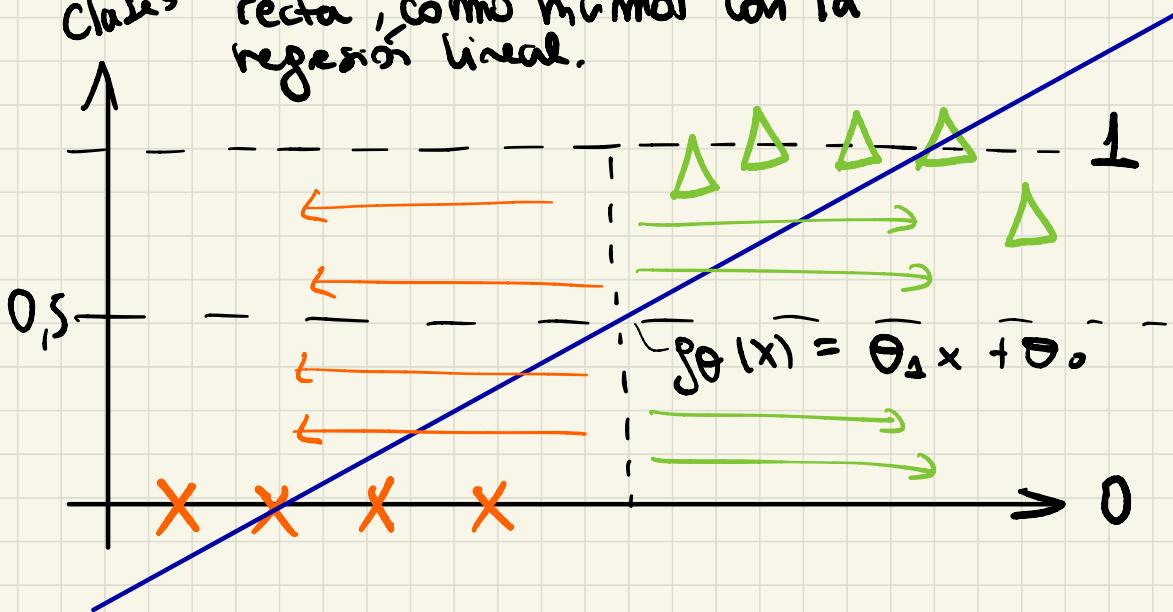
Clases



- Los cruces → triángulos representan ejemplos de entrenamiento.
- En la medida que el feature crece vemos que hay un "salto" a la otra clase.

Clases

Intentaremos utilizar una recta, como hicimos con la regresión lineal.



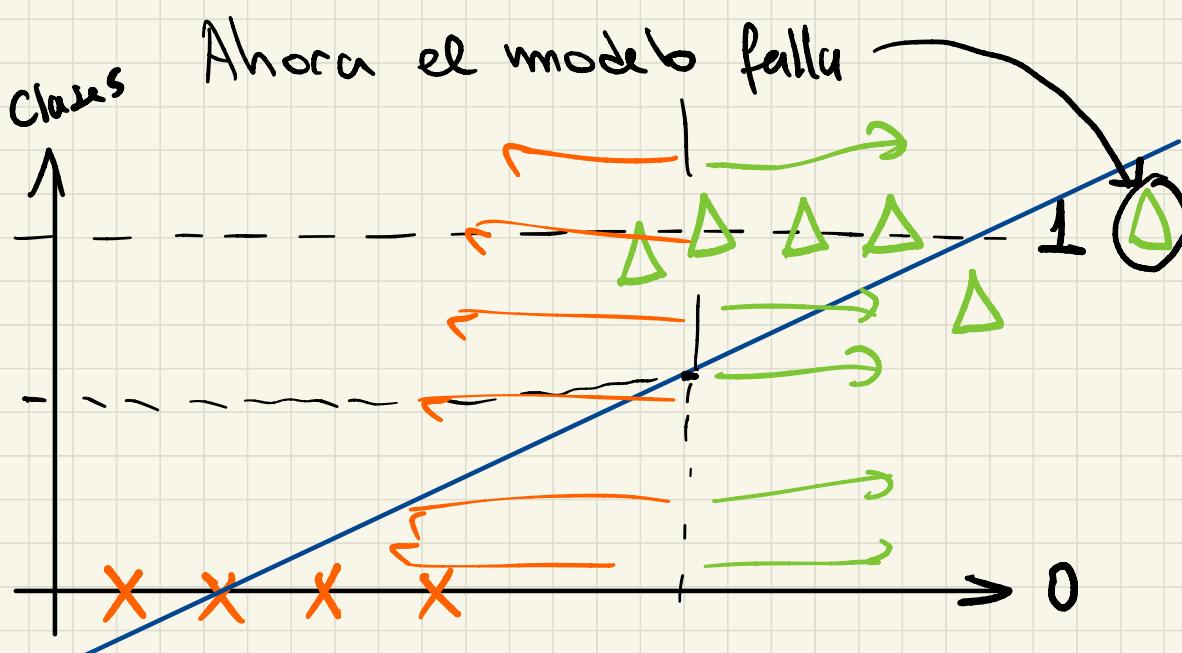
Si aplicáramos la regla :

Si $f_{\theta}(x) > 0,5 \Rightarrow \text{clase} = 1$

Si $f_{\theta}(x) < 0,5 \Rightarrow \text{clase} = 0$

Hasta aquí todo normal !!!

Veamos que ocurre si añadimos un ejemplo más :



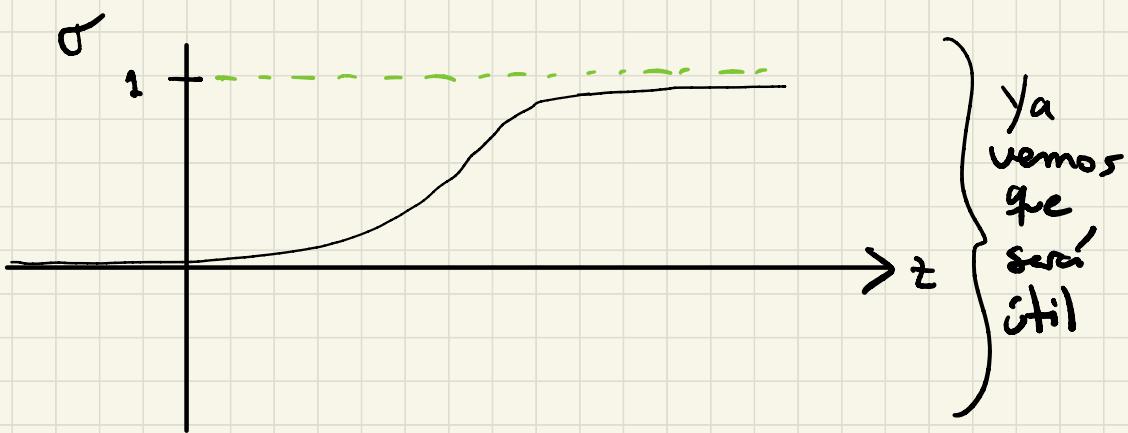
- Agregamos un dato nuevo y la regla
y e no funcional.

Función Sigmoidal

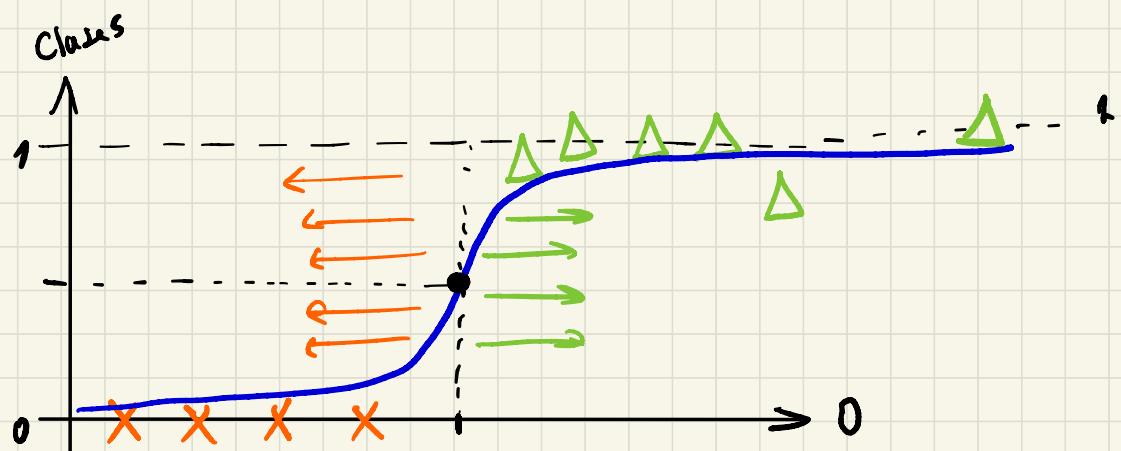
Se $\sigma : \mathbb{R} \rightarrow [0, 1]$ dada por

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

Grafiquemos usando matlab, creamos
el método!



Retomando el ejemplo anterior :



Visualmente corroboramos que resuelve
el problema.

Para agregar parámetros θ ajustables
al modelo propuesto por $\sigma(\cdot)$ podemos
basarnos en el modelo lineal previamente

estudiados, es decir :

$$\sigma_{\theta}(x_i) = \frac{1}{1 + e^{-\theta^T x_i}}$$

$$\sigma'_{\theta}(x_i) = \sigma_{\theta}(x_i) \cdot (1 - \sigma_{\theta}(x_i))$$

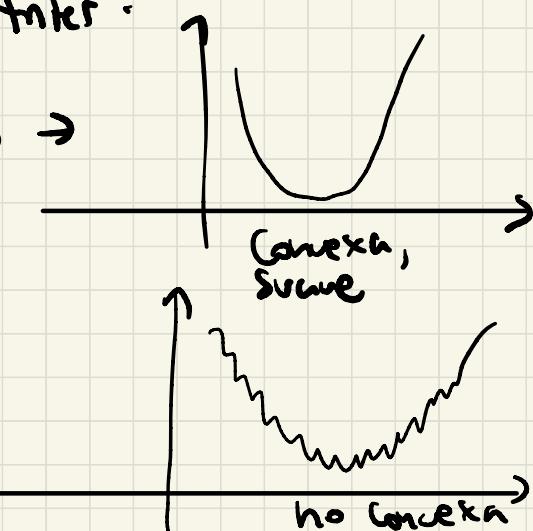
E intentar resolver :

$$J(\theta, (x, y)) = \frac{1}{2m} \sum_{i=1}^m \left(\sigma_{\theta}(x_i) - y_i \right)^2$$

Sin embargo J ahora se comportara de forma extraña. Antes :

$$\text{Con } g_{\theta}(x) = \theta_0 + \theta_1 x \rightarrow$$

Pero ahora con σ_{θ} :



∴ Necesitamos una nueva función de Costo, para evitar mínimos locales.

Intuitivamente : la función logarítmica puede suavizar este comportamiento.

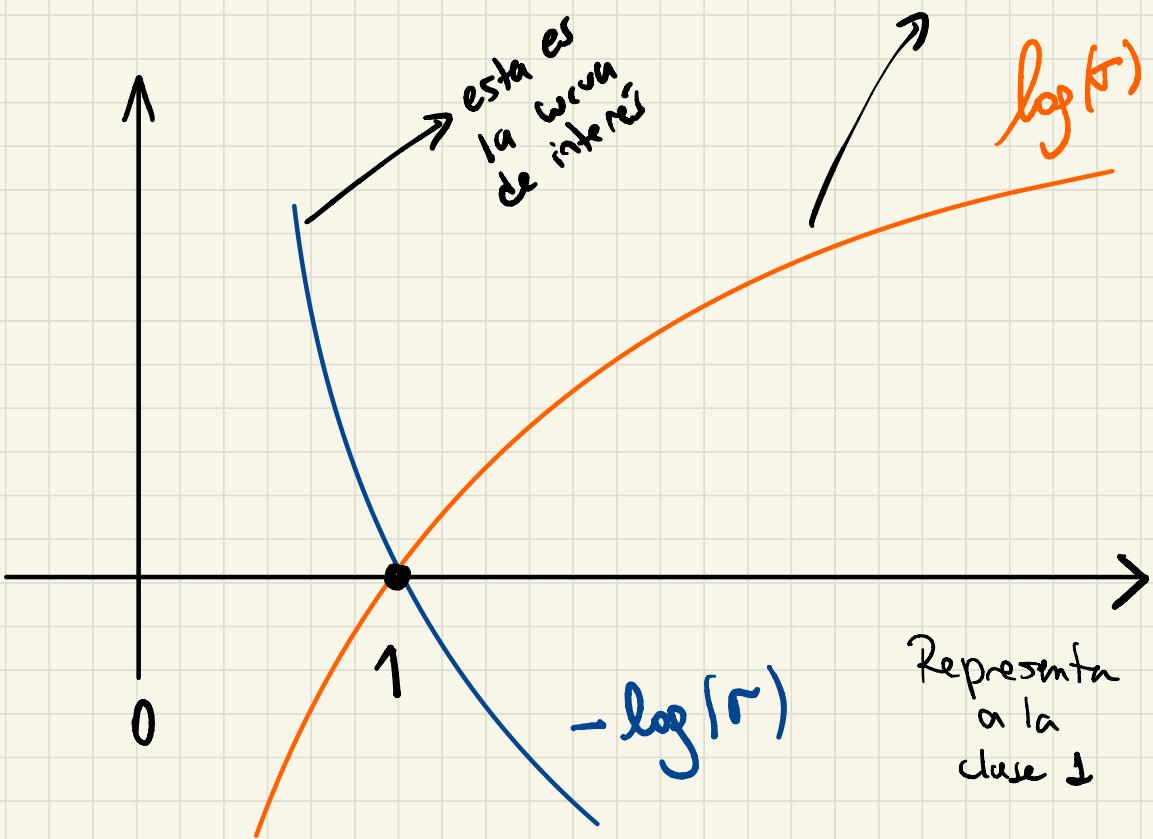
Formalmente : la función de Péndida proviene de las probabilidades (En el anexo de esta clase).

Definimos nuestra nueva función de Péndida por:

$$l(\sigma_\theta(x_i), y_i) = \begin{cases} -\log(\sigma_\theta(x_i)) & \text{Si } y_i = 1 \\ & \text{Clase 1} \\ -\log(1 - \sigma_\theta(x_i)) & \text{Si } y_i = 0 \\ & \text{Clase 0} \end{cases}$$

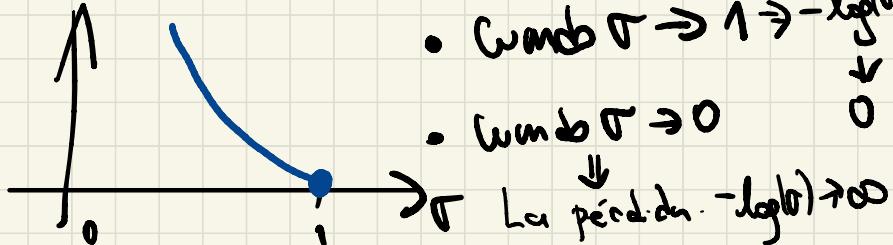
Analizaremos cada parte de esta función por partes:

esta es solo
de referencia

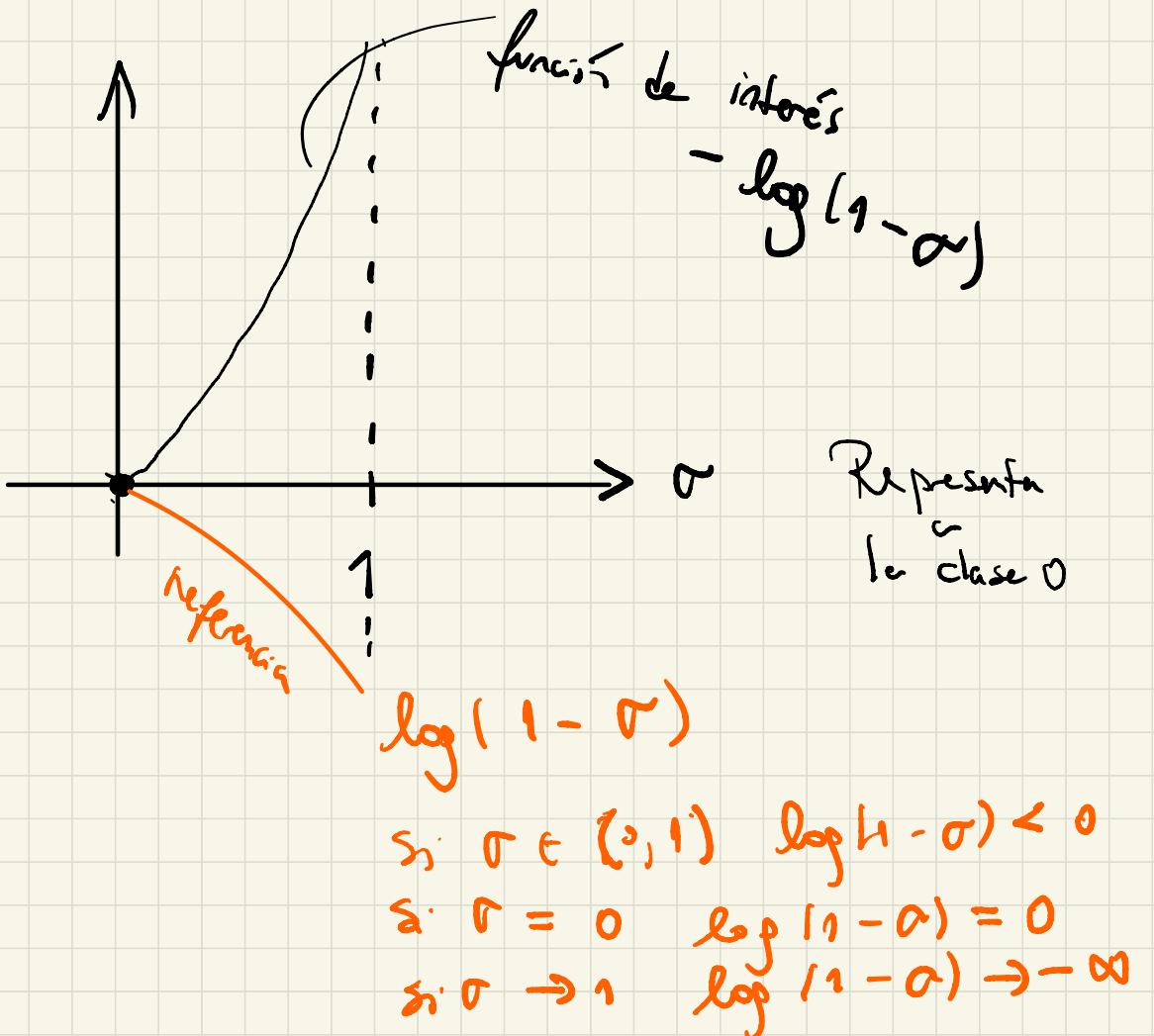


Representa
a la
clase 1

- Observemos la curva de interes solo entre $0 < r < 1$ (dominio de σ)



Vamos por la segunda parte :



Si: $\sigma \rightarrow 1 \Rightarrow -\log(1-\sigma) \rightarrow \infty$

Si: $\sigma \rightarrow 0 \Rightarrow -\log(1-\sigma) \rightarrow 0$

Simplificando la función por partes:

$$\text{Si } y_i = 0 \quad \underbrace{-y_i \log(\sigma(x_i))}_{\text{se cum}} - (1-y_i) \log(1-\sigma(x_i)) \quad \overbrace{\text{Sobrevisa este}}^{\text{se cum}}$$

Subraya este término

\downarrow

Repente
la clase 1

$$\text{Si } y_i = 1 \quad \underbrace{(1-y_i) \log(1-\sigma(x_i))}_{\text{se cum}} \quad \overbrace{\text{Sobrevisa este}}^{\text{se cum}}$$

\downarrow

Repente
la clase 0

Luego :

$$\ell(\theta, (x_i, y_i)) = -y_i \log(\sigma(x_i)) - (1-y_i) \log(1-\sigma(x_i))$$

$$J(\theta, (x_i, y_i)) = \sum_{i=1}^m \ell(\theta, (x_i, y_i))$$

- Recibe el nombre de "Binary Cross Entropy Loss"

Un ejemplo a mano (2 iteraciones)

Programar el Gradiente Descendiente utilizando la función de pérdida logística.

$$\theta^{k+1} = \theta^k - \alpha \nabla_{\theta} \mathcal{L}(\theta)$$

↓

Hallar el gradiente para :

FuncióndeCosto:

$$J(\theta) = \sum_{i=0}^m -y_i \log(\sigma(x_i)) - (1-y_i) \log(1-\sigma(x_i))$$

$$\text{Si definimos } \sigma(x_i) = \frac{1}{1 + e^{-(\theta_0 x_i + \theta_1)}}$$

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^n (\sigma(x_i) - y_i) x_i^{(j)} \longrightarrow \text{la característica } j$$

$$\frac{\partial J}{\partial b} = \frac{1}{n} \sum_{i=1}^n (\sigma(x_i) - y_i)$$

∴ La fórmula queda :

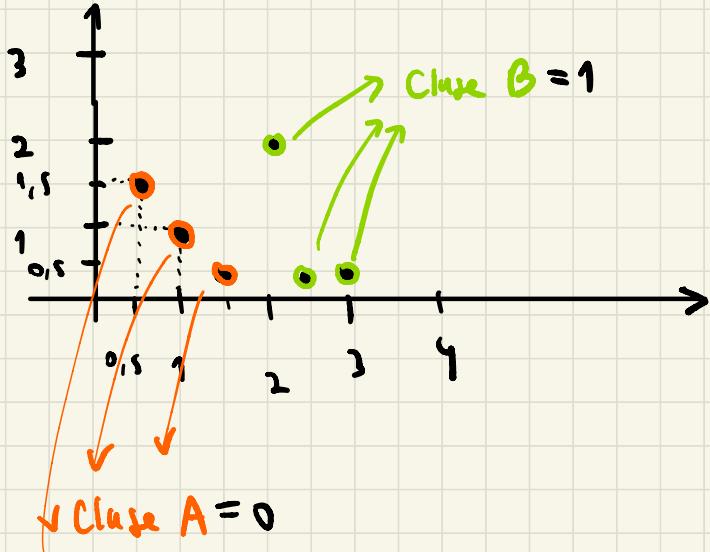
$$\theta^{k+1} = \theta^k - \alpha \left[\frac{1}{m} \sum_{i=1}^m (\Gamma_{\theta, b}(x_i) - y_i) x_i^{(k)} \right]$$

$$b^{k+1} = b^k - \alpha \left[\frac{1}{m} \sum_{i=1}^m (\Gamma_{\theta, b}(x_i) - y_i) \right]$$

Resumiendo :

- Regresión Lineal $\hat{y}_{\theta}(x_i) = \vec{\theta} x_i + b$
- Regresión logística $\Gamma_{\theta, b}(x_i) = \frac{1}{1 + e^{-(\theta x_i + b)}}$
- Los conceptos que revisamos
como cuidar de la tasa de
aprendizaje, implementación
y feature scaling aplican de la
misma forma.

Ejemplo :



$$g_{\theta}(x_i) = \theta_1 x_i + \theta_0 \quad \theta_1 = 0$$

$$\theta_0 = 0$$

Fórmulas :

$$\theta_0^{k+1} = \theta_0^k - \alpha \left[\frac{1}{6} \sum_{i=1}^6 (\sigma(x_i) - y_i) \right]$$

$$\theta_1^{k+1} = \theta_1^k - \alpha \left[\frac{1}{6} \sum_{i=1}^6 (\sigma(x_i) - y_i) \cdot x_i \right]$$

Primeras iteraciones :

$$\begin{aligned} & \boxed{R=0} \quad \boxed{\alpha=0,1} \\ & \theta_0^1 = \theta_0^0 - 0,1 \left[\frac{1}{6} \sum_{i=1}^6 \left(\frac{1}{1 + e^{-(0 \cdot x_i + 0)}} - y_i \right) \right] \approx -0,05 \\ & \theta_1^1 \approx -0,0875 \end{aligned}$$

Segunda iteración :

$$\theta_0^2 = \theta_0^1 - \alpha \left[\frac{1}{6} \sum_{i=1}^6 (f(x_i) - y_i) \right]$$

$$\theta_1^2 = \theta_1^1 - \alpha \left[\frac{1}{6} \sum_{i=1}^6 (f(x_i) - y_i) x_i \right]$$

Reemplazando θ_0^1) θ_1^1 :

$$\theta_0^2 = -0,05 - 0,1 \left[\frac{1}{6} \sum_{i=1}^6 (f(x_i) - y_i) \right] \approx -0,0875$$

$$\theta_1^2 = -0,0875 - 0,1 \left[\frac{1}{6} \sum_{i=1}^6 (f(x_i) - y_i) x_i \right] \approx -0,05$$

Aquí también reemplazamos θ_0^1 , θ_1^1

etc. El ejemplo completo está en el notebook!

+ Entrenando con los features de audio!

Anexo : Binary Cross Entropy Loss

Recordando ideas de probabilidad y estadística.
Para poder construir formalmente la función de costo vista en clases debe mas recordar los siguientes conceptos :

- 1) Probabilidad o Ley de Probabilidad
- 2) Probabilidad Condicionada, Independencia
- 3) Variables Aleatorias
- 4) Funciones de Distribución de Probabilidad
- 5) Distribución de Bernoulli

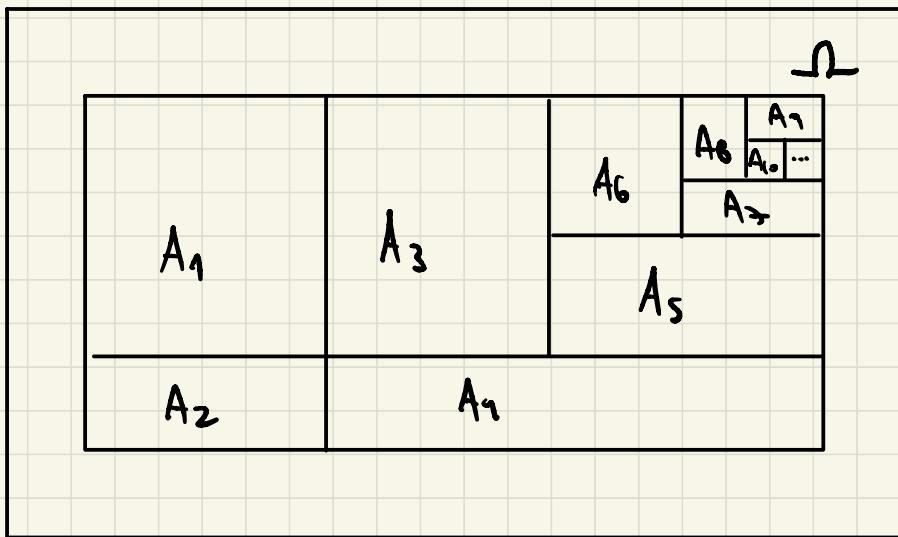
Definición 1 : Una ley de Probabilidad es una función $P: \Omega \rightarrow [0, 1]$ que mapea un evento $A \subseteq \Omega$ a un número real en $[0, 1]$.

Ω es un espacio de **eventos**. Esta función debe satisfacer los siguientes axiomas :

- i) No - negatividad : $P[A] \geq 0$, para cualquier $A \subseteq \Omega$
- ii) Normalización : $P[\Omega] = 1$
- iii) Aditividad : Para cualquier conjunto de conjuntos disjuntos $\{A_1, A_2, \dots\}$ debe ser verdadero que :

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i]$$

- * (i) Simplemente se refiere a que la función transforma eventos en números mayores o iguales de cero.
- * (ii) La probabilidad de que todos los eventos ocurran es igual a 1.
- * (iii) Ilustraremos:



- * La probabilidad de la unión de los A_i es la suma de las probabilidades individuales por causa de ser disjuntos!

Ejemplo : A_1 : sale cara en una moneda

A_2 : sale sello en una moneda

$$A_1 \cap A_2 = \emptyset \quad (\text{disjuntos} : \text{no pueden suceder simultáneamente})$$

* En este caso $\Omega = A_1 \cup A_2$

* Si ambas salidas tienen la misma probabilidad de ocurrir entonces :

$$P[A_1] = 1/2 ; \quad P[A_2] = 1/2$$

$$P[\Omega] = P[A_1 \cup A_2] = P[A_1] + P[A_2] = 1$$

2) Definición 2.1: Probabilidad Condicional

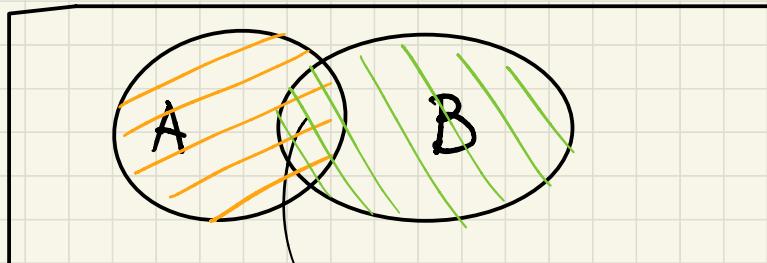
Es otra forma de calcular probabilidades.
Cumple con los axiomas de la definición (1).

Dados dos eventos A y B , asumiendo que $P[B] \neq 0$; La probabilidad condicional de A dado B es :

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

↑
A dado B

Ejemplos : Ilustraremos la definición 2



Al lanzar un dado : $\rightarrow A \cap B$ (Suceden conjuntamente algunas veces)

$$A = \{ \text{Obtener un } 3 \}$$

$$B = \{ \text{Obtener un número ímpar} \}$$

$$\begin{aligned} P[A|B] &= \frac{P[A \cap B]}{P[B]} = \frac{1/6}{3/6} \\ &= 1/3 \end{aligned} \quad \left. \begin{array}{l} \text{el } 3 \text{ es} \\ \text{la intersección} \end{array} \right\} \quad \left. \begin{array}{l} \text{hay } 3 \\ \text{números} \\ \text{ímpares} \end{array} \right\}$$

$$P[B|A] = \frac{1/6}{1/6} = 1 \quad \left(\begin{array}{l} \text{Dados que obtuvimos,} \\ \text{un } 3 \text{ eso siempre será} \\ \text{ímpar}. \end{array} \right)$$

Definición 2.2 : Dos eventos son estadísticamente independientes si : $P[A \cap B] = P[A] \cdot P[B]$

Esto se justifica por la definición 2.1 :

Si A no depende de B entonces :

$$P[A|B] = P[A]$$

“la probabilidad de A dado B ” = “es la probabilidad de que ocurra A por si sola”.

Análogamente :

$$P[B|A] = P[B]$$

Ejemplo : Lanzas 2 dados.

$$A = \{ \text{el primer dado es } 3 \}$$

$$B = \{ \text{La suma es } 7 \}$$

Cálcula tú mismo las probabilidades $P[A \cap B]$, $P[A]$, $P[B]$.

Definición 3 : Variables Aleatorias

Una variable aleatoria X es una función $X: \Omega \rightarrow \mathbb{R}$ que mapea un evento $\xi \in \Omega$ a un número real $X(\xi)$.

Ejemplo :

- Sea $\Omega = \{\Delta, \diamondsuit, \heartsuit, \circlearrowleft\}$
- Supongamos que la probabilidad de cada evento es dada por :
- $P[\Delta] = 1/6, P[\diamondsuit] = 2/6, P[\heartsuit] = 2/6, P[\circlearrowleft] = 1/6$
- Quisiéramos dejar de utilizar figuras para nuestros eventos. Los vamos a mapear :

$$X(\Delta) = 1 ; X(\diamondsuit) = 2 ; X(\heartsuit) = 3 ; X(\circlearrowleft) = 4$$

Ahora podemos anotar :

$$P[X=1] = 1/6 \quad P[X=2] = 2/6 \quad P[X=3] = 2/6$$

$$P[X=4] = 1/6.$$

$$P[X=1] = 1/6$$

↑ ↑ ↑ ↑
 Proba Variable Aleatoria Estado de la variable Probabilidad que ocurren

X : Definición del evento

Otro ejemplo : Tirar dos dados

$$\Omega = \{(1,1), (1,2), \dots, (6,6)\}$$

todos los eventos

$$\xi_1 = (1,1) \quad \xi_2 = (1,2) \quad \dots \quad \xi_{36} = (6,6)$$

↑
un evento

$X = \text{Suma de los dos números}$

Variable Aleatoria

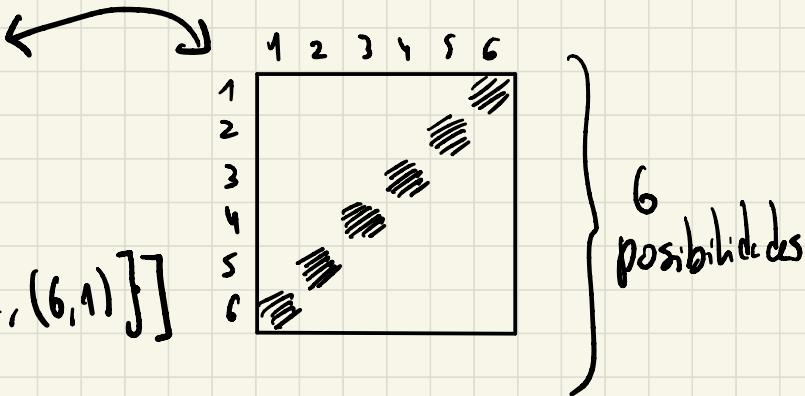
$$P[X = 7]$$

||

$$P[\{(1,6), (2,5), \dots, (6,1)\}]$$

||

$$P[(1,6)] + \dots + P[(6,1)] = 6 \cdot 1/36 = 1/6$$



Definición 4 : Función de distribución de probabilidad

Caso discreto

$$p_X(x) = P[X = x]$$

función del estado

Caso continuo

$$P[a \leq X \leq b] = \int_a^b f_X(x) dx$$

- Observe que el caso discreto obtiene directamente un valor de probabilidad (Hay finitos estados).
- En el caso continuo Ω es infinito y f_X ahora representa una densidad de probabilidad. Debemos acumular valores entre $[a, b]$ para medir la probabilidad.
- Las funciones de densidad f_X son nuestros modelos para caracterizar / modelar variables aleatorias.

Ejemplos :

$$1) f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{en otro caso} \end{cases} \quad (\text{Variable Aleatoria Uniforme})$$

Definición 5

2) Si $Y \sim \text{Bernoulli}(p)$ ("Si Y sigue una distribución de Bernoulli con parámetro p)

Entonces :

$$\begin{aligned} p_Y(0) &= 1 - p & p_Y(1) &= p \quad \text{con } 0 < p < 1 \\ \downarrow & & \downarrow & \\ P[Y=0|p] & & P[Y=1|p] & \end{aligned}$$

O en una sola expresión :

$$P[Y=y] = \begin{cases} p^y (1-p)^{1-y} & y \in \{0,1\} \\ 0 & \text{en otro caso} \end{cases}$$

Este modelo se utiliza para generar datos dado un p o ajustar p dado un conjunto de datos.

Por ejemplo, si $p=1/2$ modelamos la probabilidad de obtener cara $X=1$ o Sello $X=0$ para una moneda no cargada o "justa".

Construcción : Asumimos que los ejemplos de entrenamiento son independientes entre sí y que queremos ajustar el parámetro p .

Bajo nuestra notación $p \equiv \sigma$. Dado un conjunto de ejemplos (x_i, y_i) con $i=1, \dots, m$ independientes entonces podemos definir una probabilidad condicional dada por :

$$P[y_i | \sigma(x_i)] = \prod_{i=1}^m \sigma(x_i)^{y_i} (1-\sigma(x_i))^{(1-y_i)}$$

$$\log(P[y_i | \sigma(x_i)]) = \sum_{i=1}^m y_i \log(\sigma(x_i)) + (1-y_i) \log(1-\sigma(x_i))$$

Usualmente se considera $-\log$, para mantener la convención de minimizar. De ahí obtenemos:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m y_i \log(\sigma(x_i)) + (1 - y_i) \log(1 - \sigma(x_i))$$

↓
media