Regressão logística:

$$x \in \mathbb{R}^d \longmapsto p(x) \in [0,1].$$

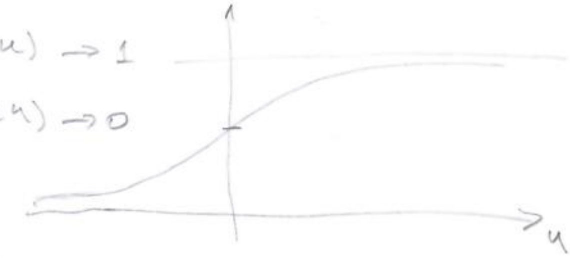
↑ probabilidade

$$\sigma(u) = \frac{e^u}{e^u + e^{-u}} = \frac{1}{1+e^{-2u}}$$

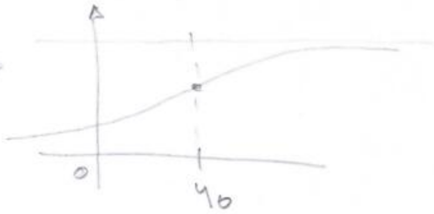
$u \to +\infty$, $\sigma(u) \to 1$

$u \to -\infty$, $\sigma(u) \to 0$

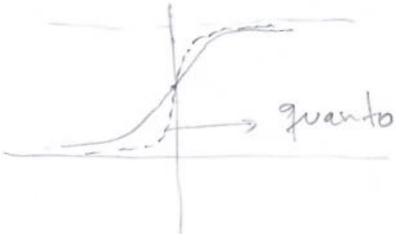
$\sigma(0) = 1/2$



$\sigma(u - u_0)$ →



$\sigma(tu)$



→ quanto maior o valor de t.

→ função aleatoria de X.

$$Y = \begin{cases} 1 \\ -1 \end{cases} \qquad \mathbb{P}(Y=1|X) = \frac{1}{1+e^{-s(X)}}, \quad s(X) = \tilde{w}_0 + \sum_{i=1}^{d} \tilde{w}_i X(i)$$

$$\tilde{w}_i = 2w_i$$

$$\mathbb{P}(Y=-1|X) = \frac{e^{s(X)}}{1+e^{s(X)}}$$

$$\boxed{\mathbb{P}(\underset{\pm 1}{Y = y} | X) = \frac{1}{1+e^{-y\, s(X)}}}$$

Dada amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$ queremos encontrar $w = (w_0, \ldots, w_d)$ que <u>maximize</u> a verosimilhança da amostra.

Maximizar a verosimilhança

$$\frac{1}{n} \log \mathbb{P}\left(Y_1 = y_1, \dots, Y_n = y_n \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\right)$$

$$= \frac{1}{n} \log \prod_{m=1}^{n} \frac{1}{1 + e^{-y_m s(x_m)}} = -\frac{1}{n} \sum_{m=1}^{n} \log\left(1 + e^{-y_m s(x_m)}\right)$$

onde $s(x) = w_0 + \sum_{j=1}^{d} w_j x(j)$

é o mesmo que minimizar

$$\mathcal{E}(w) = \frac{1}{n} \sum_{m=1}^{n} \log\left(1 + e^{-y_m s(x_m)}\right)$$

"error" "cross-entropy error"

Problema: Encontrar $\hat{w}$ que minimize $\mathcal{E}(w)$.

Exemplo: $d = 1$, $w = (\underset{a}{\underset{\parallel}{w_0}}, \underset{b}{\underset{\parallel}{w_1}})$

$$\mathcal{E}(a, b) = \frac{1}{n} \sum_{m=1}^{n} \log\left[1 + e^{-y_m(a + b x_m)}\right]$$

Gradient Descent

(Descida através do gradiante)

$$\mathcal{E}(a + \eta r_a, b + \eta r_b) - \mathcal{E}(a, b)$$

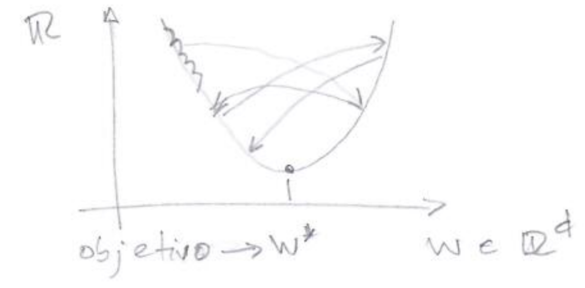
η: comprimento do paso

$r_a = \pm 1$, $r_b = \pm 1$.

Em geral:

Parto de $w(0) = (w_0(0), \dots, w_d(0))$ e paso $w(1), w(2), \dots$ de cada vez seguindo a direçao que disminui o erro.

Suponhamos que o erro tenha esta forma



η muito pequenho : Não chego nunca

η grande demais : oscilo.

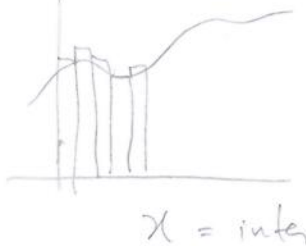
Redes Neurais

$x \in \mathbb{R}^d$, $\eta(x) \in [0,1]$

1 única camada

Posso aproximar → a aproximaçao é tanto melhor quanto maior for o número de sub-intervalos considerados.

$x \longmapsto \eta(x)$ por uma função escada (constante por intervalo)



η_K , K ← no de intervalo

$\text{distancia}(\eta, \eta_K) \xrightarrow[K \to \infty]{} 0$

Podemos aproximar funçoes contínuas de $\mathcal{X}$ en $[0,1]$ por :

- funçoes em escada
- polinômios
- funçoes trigonométricas.

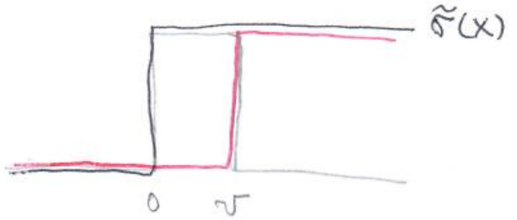
Teorema de Stone Weiestras

Aprendizagem Estatístico

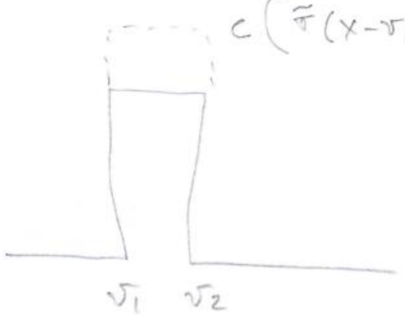
1. $\eta(x) = \mathbb{P}(Y=1 \mid X=x)$

2. Função escada é uma rede neural

$$\tilde{\sigma}(x) = \begin{cases} 1, & \text{se } x \geq 0 \\ 0, & \text{se } x < 0 \end{cases}$$



$c\left(\tilde{\sigma}(x-v_1) - \tilde{\sigma}(x-v_2)\right)$



$$\tilde{\sigma}(x) - \tilde{\sigma}(x-v) = \mathbb{1}_{[0,v]}(x)$$

$$\tilde{\sigma}(x-v_1) - \tilde{\sigma}(x-v_2)$$

$$f_K(x) = \sum_{i=0}^{K} c_i \left[\tilde{\sigma}(x-s_i^k) - \tilde{\sigma}(x-s_{i+1}^k)\right]$$

K = nº de "neuronios" da camada.