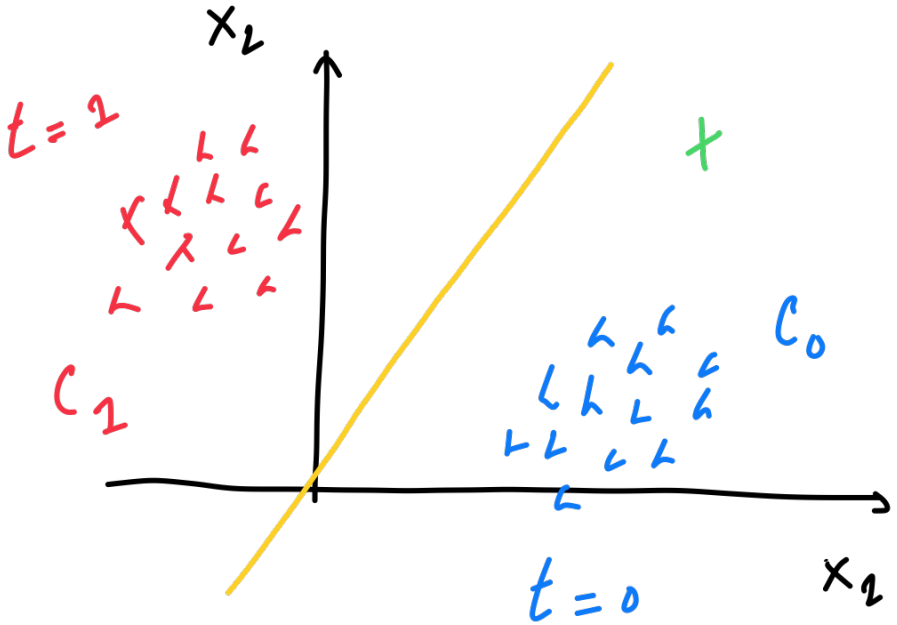
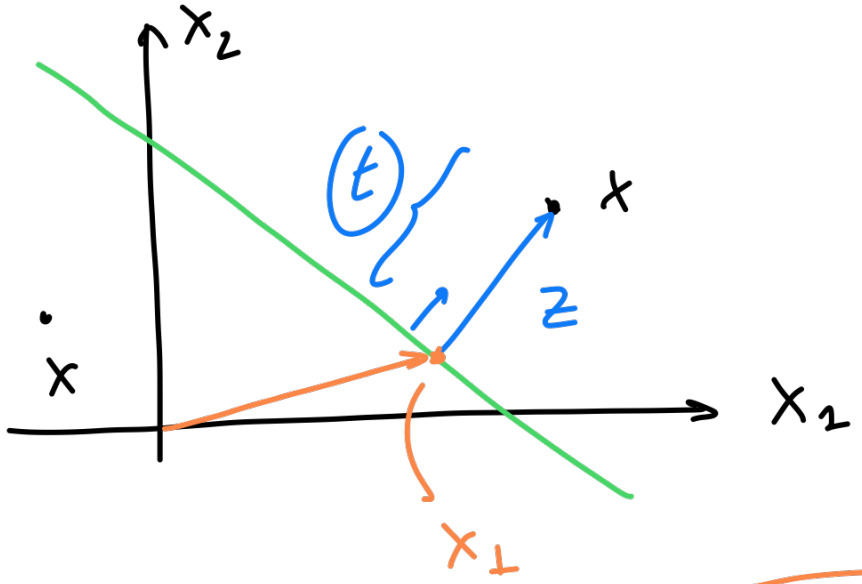
Apprentissage supervisé $\{x^{(i)}, t^{(i)}\}_{i=1}^{N}$

$\rightarrow h_\beta(x) = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}$

Partie 2 : Classification $t^{(i)} \in \{1, 2, .., K\}$



$t^{(i)} \in \{0, 1\}$



$$\vec{x} = \vec{x}_\perp + \vec{z}$$

$$(x_1, x_2) = ((x_\perp)_1, (x_\perp)_2) + (z_1, z_2)$$

$\rightarrow$ $x_\perp$ est sur le plan : $\beta_0 + \beta_1 (x_\perp)_1 + \beta_2 (x_\perp)_2 = 0$ $\leftarrow$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = \beta_0 + \beta_1 (x_\perp)_1 + \beta_2 (x_\perp)_2 + \beta_1 z_1 + \beta_2 z_2$$

$t \, \|(\beta_1, \beta_2)\| \geq 0$

0

$$\vec{z} = \frac{t(\beta_1, \beta_2)}{\|(\beta_1, \beta_2)\|} \leftarrow$$

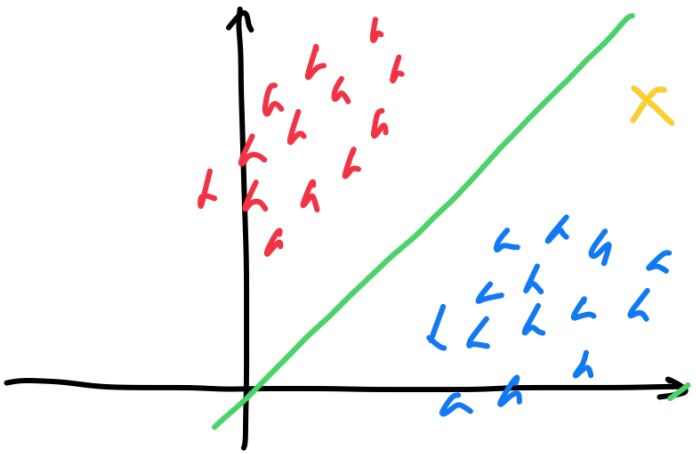
$\rightarrow$

$$\langle \vec{z}, (\vec{\beta}_1, \vec{\beta}_2) \rangle = \frac{t \, \|(\beta_1, \beta_2)\|^2}{\|(\beta_1, \beta_2)\|}$$

Fonction coût

$$\min_{\beta} \quad l(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} \right) \right)^2$$

$t^{(i)} \in \{1, 0\}$

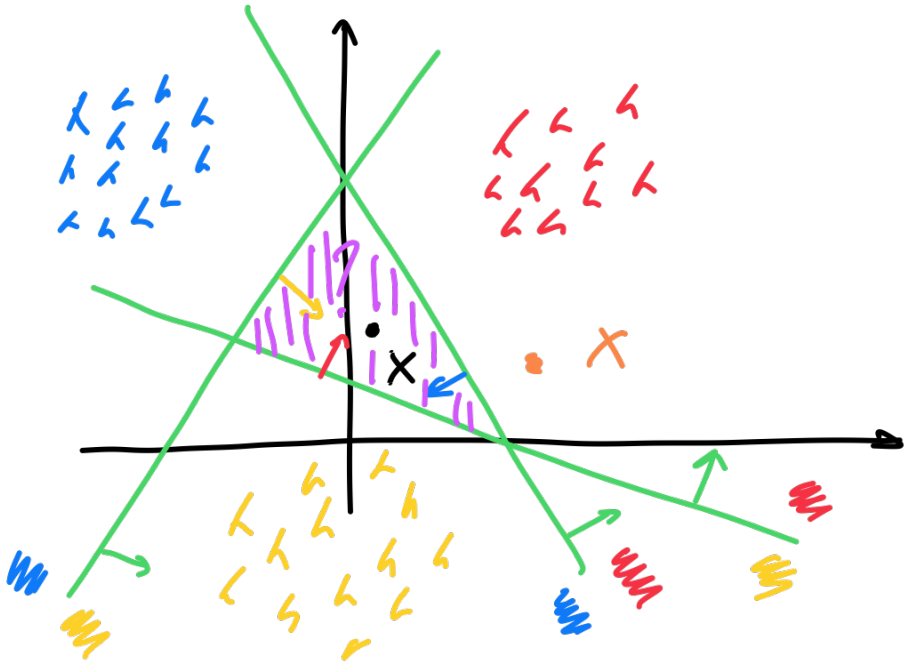


$$\text{prediction}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

C_1 C_0

if prediction $(x) > 1/2 \rightarrow C_1$

$< 1/2 \rightarrow C_0$

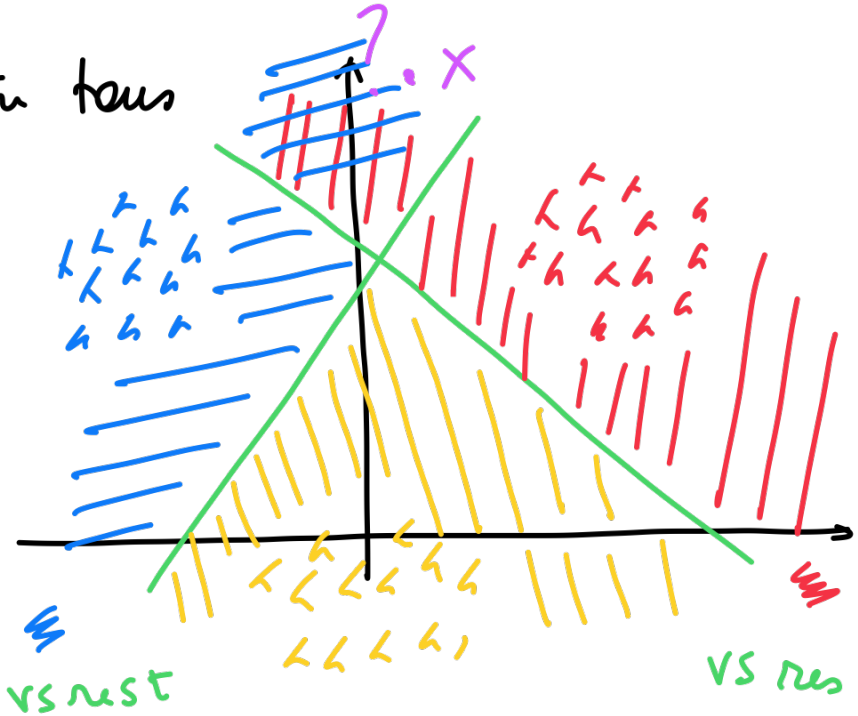


→ Approche 1 contre 1

$\frac{K(K-1)}{2}$ discriminants

Classe de X donnée par vote à la majorité

Approche 1 contre tous



$(K-1)$ discriminants

→ Solution #3 : Discriminant à K classes

$t^{(i)} \in \{0, 1\} \rightarrow \vec{t}^{(i)} = [0, 0, \dots, \underset{\uparrow}{1}, 0, 0 \dots 0]$ (one hot encoding)

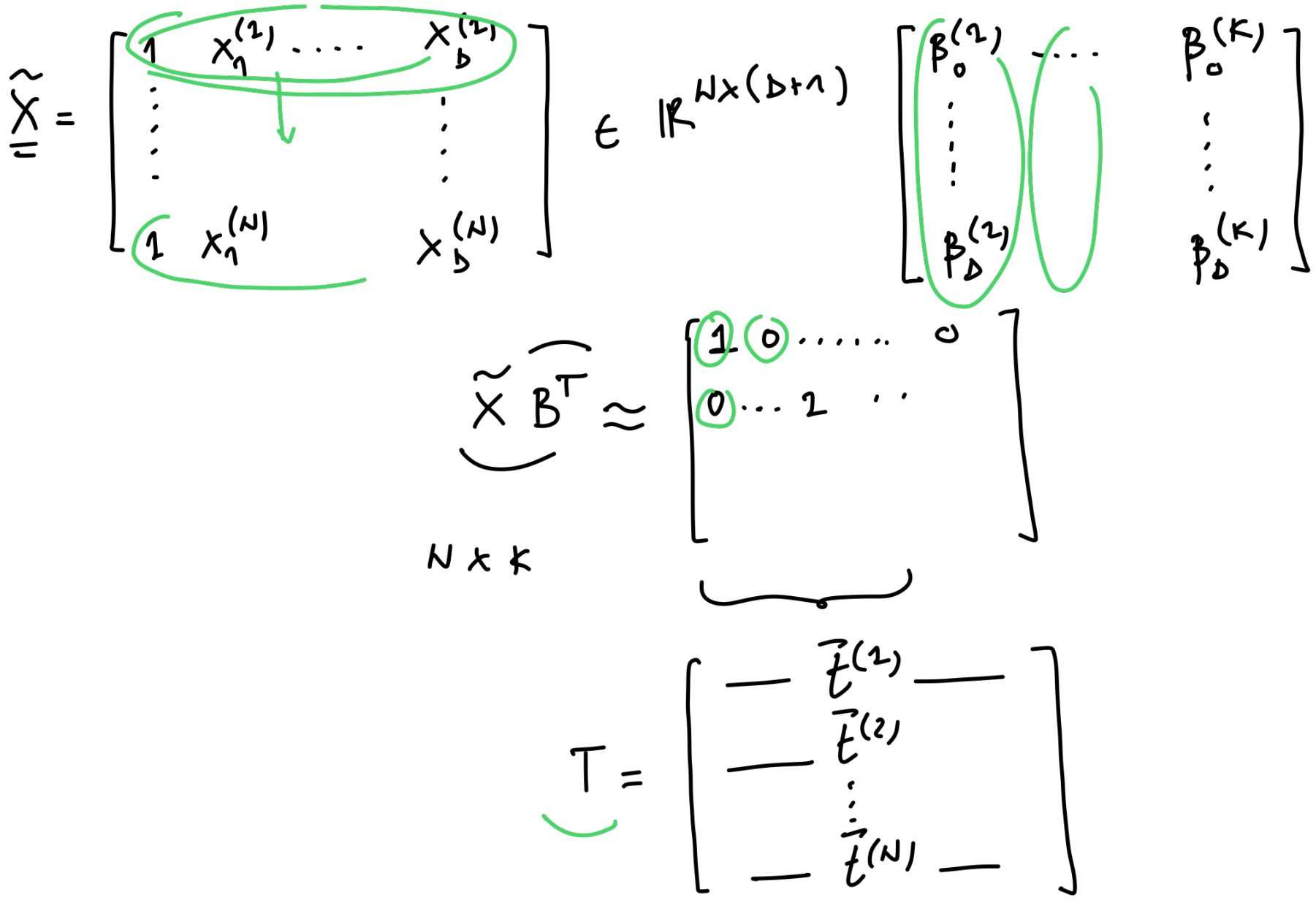
K

$$\vec{B} = \begin{bmatrix} \beta_0^{(1)} & \dots & \beta_D^{(1)} \\ \beta_0^{(2)} & \dots & \beta_D^{(2)} \\ \vdots & & \\ \beta_0^{(K)} & \dots & \beta_D^{(K)} \end{bmatrix} \in \mathbb{R}^{K \times (D+1)}$$

→ discriminant pour la classe K

si $x \in$ classe K $\Rightarrow$ $\beta_0^{(K)} + \beta_1^{(K)} x_1 + \dots + \beta_D^{(K)} x_D \approx 1$

si $x \notin$ classe K $\Rightarrow$ $\beta_0^{(K)} + \beta_1^{(K)} x_1 + \dots + \beta_D^{(K)} x_D \approx 0$


$\tilde{X} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_D^{(1)} \\ \vdots & & & \vdots \\ 1 & x_1^{(N)} & & x_D^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times (D+1)}$
$\begin{bmatrix} \beta_0^{(1)} & \cdots & \beta_0^{(K)} \\ \vdots & & \vdots \\ \beta_D^{(1)} & & \beta_D^{(K)} \end{bmatrix}$
$\tilde{X} B^T \approx \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \cdots & 2 & \cdots \end{bmatrix}$
$N \times K$
$T = \begin{bmatrix} — & \vec{t}^{(1)} & — \\ — & \vec{t}^{(2)} & \\ & \vdots & \\ — & \vec{t}^{(N)} & — \end{bmatrix}$

$$\underset{B}{\operatorname{argmin}} \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(t_k^{(i)} - \left(\beta_0^{(k)} + \beta_1^{(k)} x_1^{(i)} + \dots + \beta_D^{(k)} x_D^{(i)} \right) \right)^2$$

$$\underset{B}{\operatorname{argmin}} \frac{1}{NK} \left\langle T - \tilde{\underline{\underline{X}}} B^T, \; T - \tilde{\underline{\underline{X}}} B^T \right\rangle_F$$

$$\langle A, B \rangle_F = \sum_{ij} A_{ij} B_{ij} \qquad \langle B^T, \; \tilde{X}^T (T - \tilde{X} B^T) \rangle$$

$$\underline{\hat{B}} = (\tilde{X}^T \tilde{X})^{-2} \tilde{X}^T \underline{T}$$

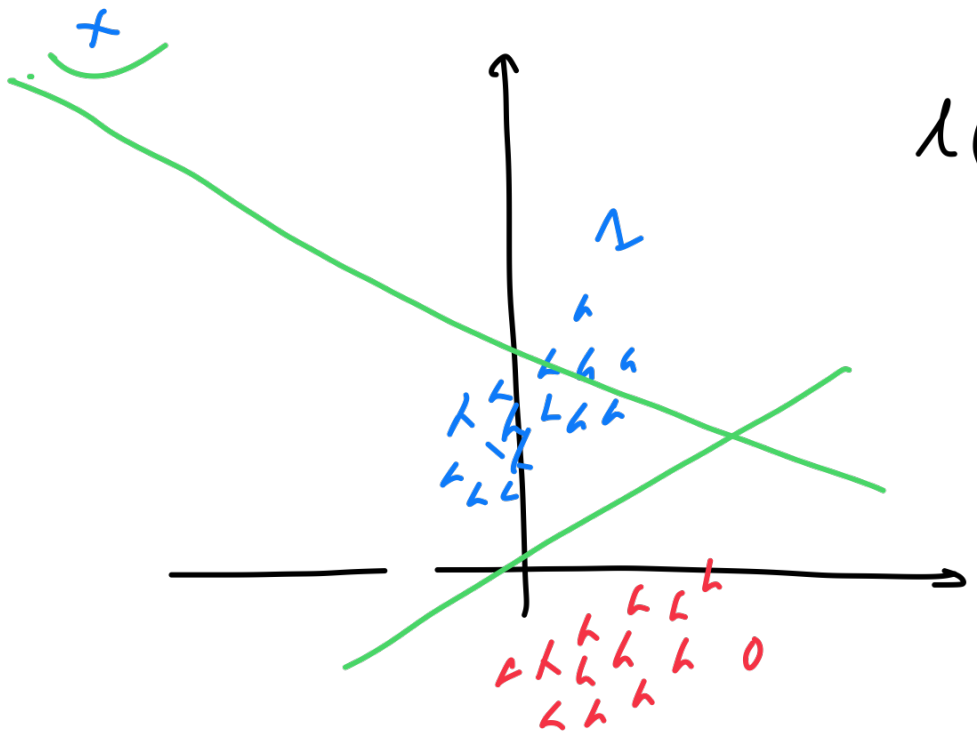
Soit x une donnée test

$\rightarrow$ la classe de x est calculée via

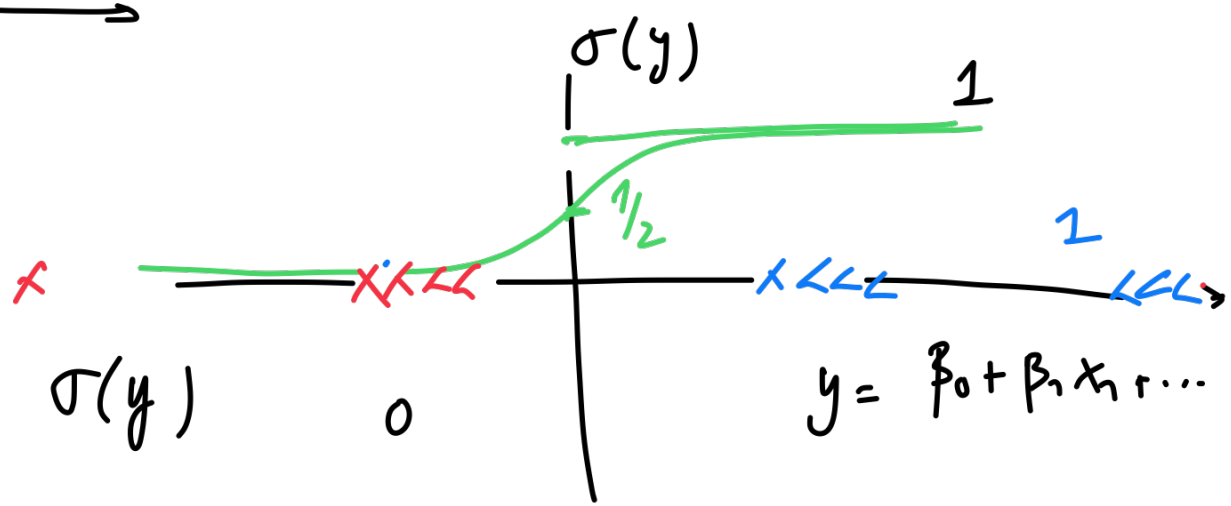
$$\begin{bmatrix} \beta_0^{(1)} \cdots & \beta_D^{(1)} \\ \vdots & \\ \beta_0^{(K)} \cdots & \beta_D^{(K)} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

$$= \begin{bmatrix} t_1(x) \\ \vdots \\ t_K(x) \end{bmatrix}$$

$$\text{Classe de } x = \underset{k}{\operatorname{argmax}} \; t_k(x)$$



$$\mathcal{L}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} \right) \right)$$



→ Fonction d'activation $\sigma(y)$

$$y(x) = \sigma(\beta_0 + \beta_1 x_1 + \dots + \beta_D x_D)$$

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

$$p(t = \text{blue} \mid x) = \sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D)$$

$$p(t = \text{red} \mid x) = 1 - \sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D)$$