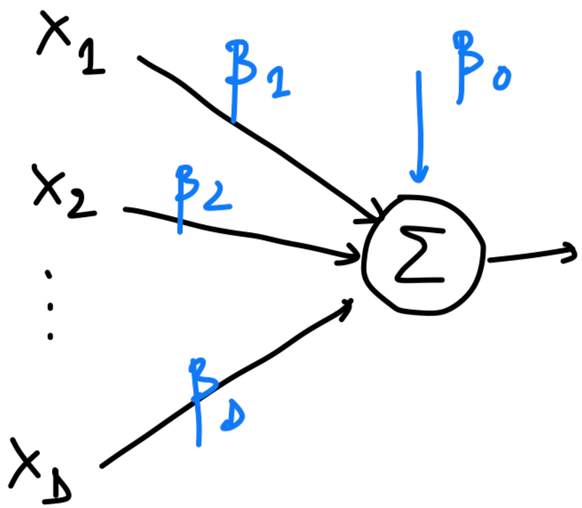
Rappels → supervisé

→ Régression linéaire

→ Classification linéaire

↳ K classes

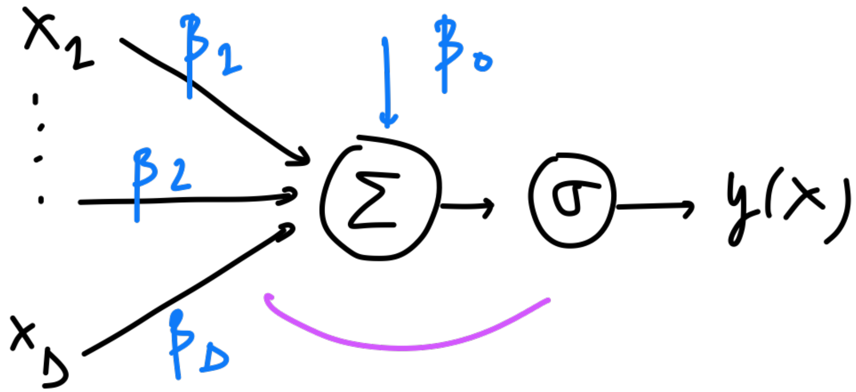
→ points adhérents → fonction d'activation

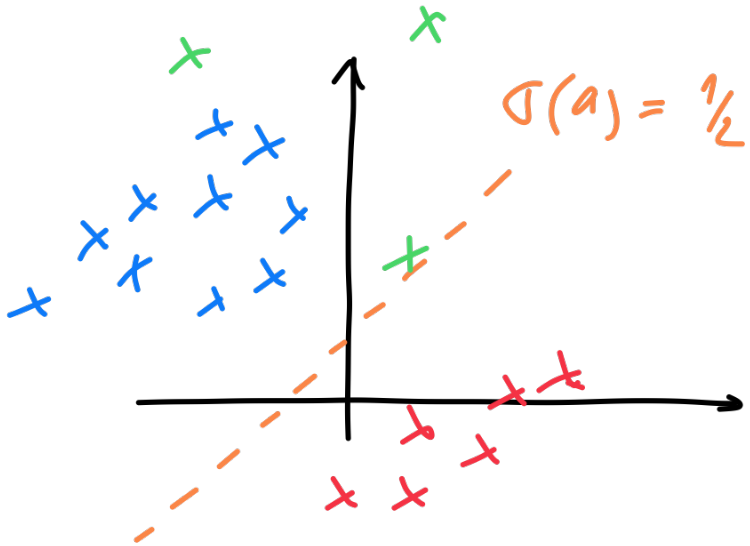


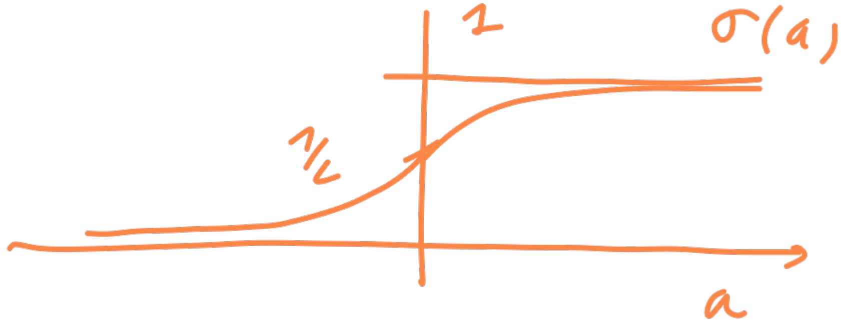
$$y(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$$

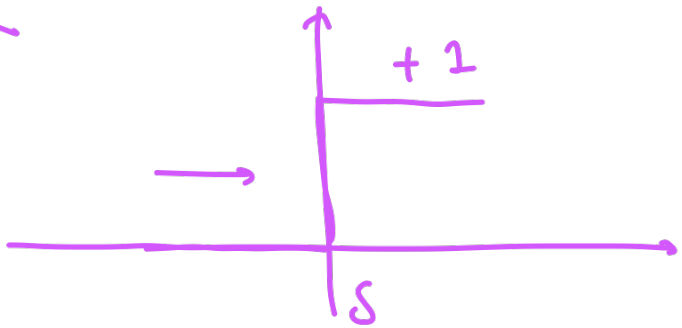
Regression logistique

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









$$H(a) = \begin{cases} +1 & a \geq 0 \\ -1 & a < 0 \end{cases}$$

$$p(t(x) = 1 \mid x ; \beta) = \sigma(\beta_0 + \beta_1 x_1 + \dots + \beta_D x_D)$$

$$p(t(x) = 0 \mid x ; \beta) = 1 - \sigma(\beta_0 + \beta_1 x_1 + \dots + \beta_D x_D)$$

→ Données indépendantes

$$p(\{t^{(i)}\}_{i=1}^{N} \mid \{x^{(i)}\}_{i=1}^{N} ; \beta) = \prod_{i=1}^{N} p(t(x) = t^{(i)} \mid x^{(i)}, \beta)$$

$$p(t(x^{(i)}) = t^{(i)} \mid x^{(i)} ; \beta) = \sigma(\beta_0 + \dots + \beta_D x_D^{(i)})^{t^{(i)}} (1 - \sigma(\beta_0 + \dots + \beta_D x_D^{(i)}))^{1-t^{(i)}}$$

$$\beta^* = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{N} \sigma(\beta_0 + \dots + \beta_D x_D^{(i)})^{t^{(i)}} \left(1 - \sigma(\beta_0 + \dots + \beta_D x_D^{(i)})\right)^{1-t^{(i)}}$$

log

$$\underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N} t^{(i)} \log(\sigma(\beta_0 + \dots + \beta_D x_D^{(i)})) + (1 - t^{(i)}) \log(1 - \sigma(\beta_0 + \dots + \beta_D x_D^{(i)}))$$

$$= \underset{\beta}{\operatorname{argmin}} - \sum_{i=1}^{N} t^{(i)} \log(\sigma(\beta_0 + \dots + \beta_D x_D^{(i)})) + (1 - t^{(i)}) \log(1 - \sigma(\beta_0 + \dots + \beta_D x_D^{(i)}))$$

ENTROPIE BINAIRE CROISÉE

$$L(\beta) = - \sum_{i=1}^{N} t^{(i)} \log(y(x^{(i)})) + (1 - t^{(i)}) \log(1 - y(x^{(i)}))$$

pour entraîner le modèle de regression logistique

→ Descente de gradient

$$\underset{\beta}{\mathrm{grad}}\, L(\beta) = -\sum_{i=1}^{N} t^{(i)} \frac{\underset{\beta}{\mathrm{grad}}\, \sigma(\beta^T \tilde{x}^{(i)})}{\sigma(\beta^T \tilde{x}^{(i)})} + (1-t^{(i)}) \frac{\underset{\beta}{\mathrm{grad}}\, (1-\sigma(\beta^T \tilde{x}^{(i)}))}{(1-\sigma(\beta^T \tilde{x}^{(i)}))}$$

$$\sigma = \sigma(\beta^T \tilde{x}^{(i)})$$

$$\underset{\beta}{\mathrm{grad}}\, \sigma(\beta^T \tilde{x}^{(i)}) = \sigma'(\beta^T \tilde{x}^{(i)}) \cdot \tilde{x}^{(i)} = \sigma(1-\sigma) \cdot \tilde{x}^{(i)}$$

$$\sigma'(a) = \frac{d}{da}\left(\frac{1}{1+e^{-a}}\right) = \frac{e^{-a}}{(1+e^{-a})^2} = (1-\sigma(a))\,\sigma(a)$$

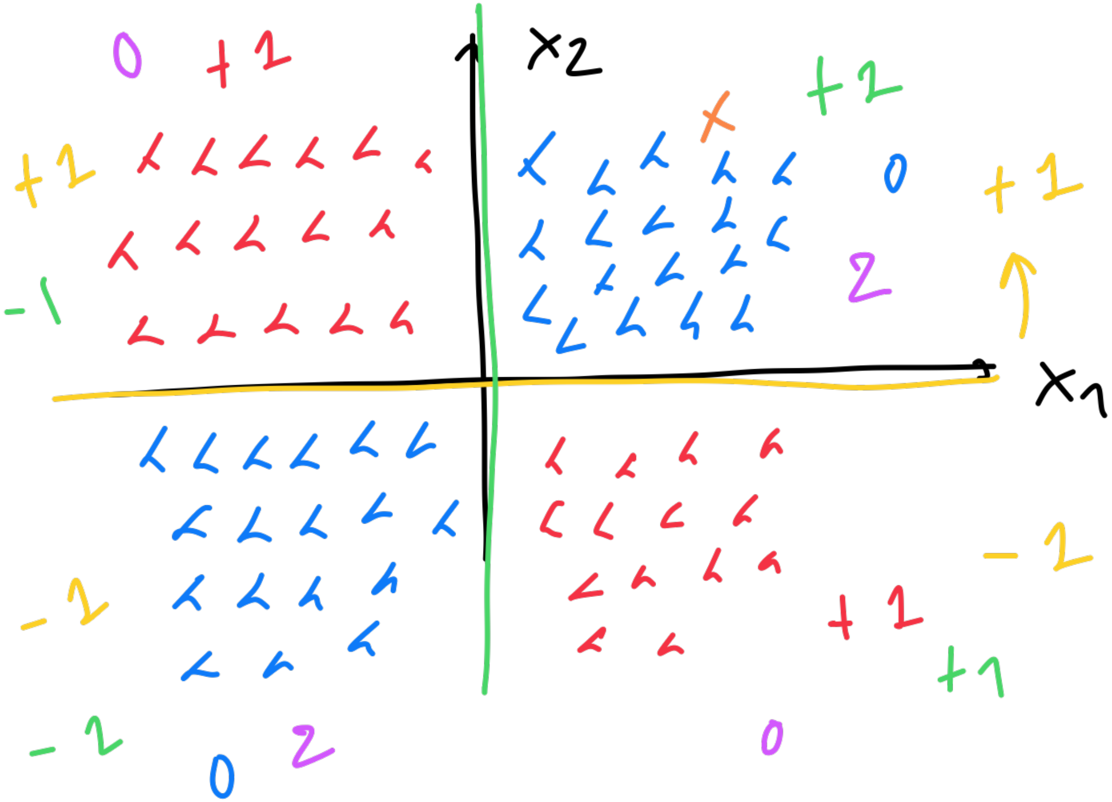
$$\operatorname{grad}_{\beta} L(\beta) = - \sum_{i=1}^{N} t^{(i)} \frac{\cancel{\sigma}(1-\sigma)\,\tilde{x}^{(i)}}{\cancel{\sigma}} - (1-t^{(i)}) \frac{\sigma\cancel{(1-\sigma)}\,\tilde{x}^{(i)}}{\cancel{1-\sigma}}$$

$$= - \sum_{i=1}^{N} t^{(i)}(1-\sigma)\,\tilde{x}^{(i)} - (1-t^{(i)})\,\sigma\,\tilde{x}^{(i)}$$

$$= - \sum_{i=1}^{N} \left(t^{(i)} - \sigma(\beta^T \tilde{x}^{(i)})\right) \tilde{x}^{(i)}$$

$$\beta^{(k+1)} \leftarrow \beta^{(k)} + \eta \sum_{i=1}^{N} \left(t^{(i)} - \sigma(\beta^T \tilde{x}^{(i)})\right) \tilde{x}^{(i)}$$

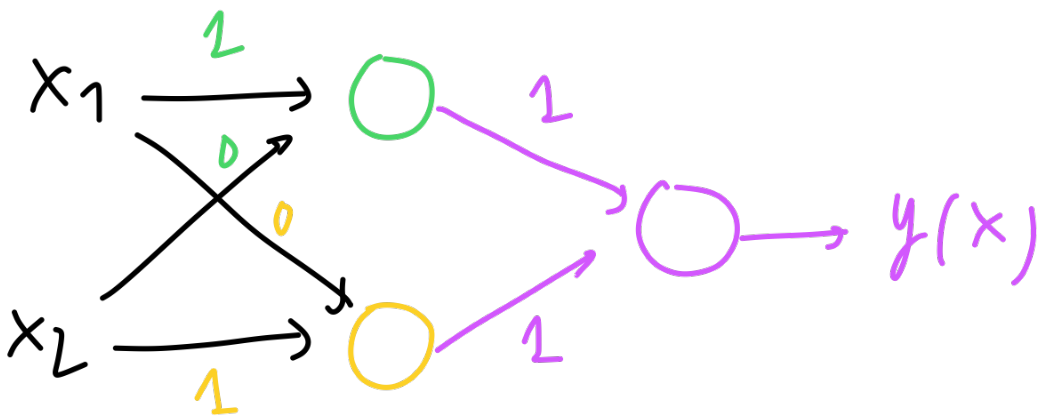
Données non linéairement séparables



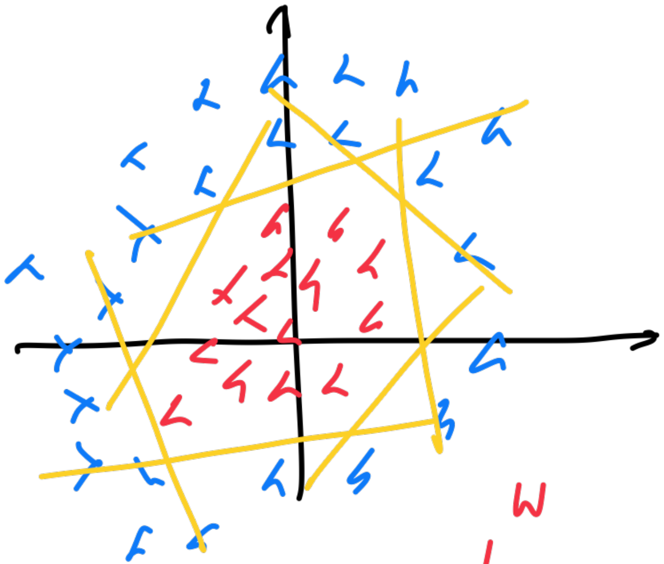
$$\sigma(x_2) = \begin{cases} +1 & x_2 \geq 0 \\ -1 & x_2 < 0 \end{cases}$$

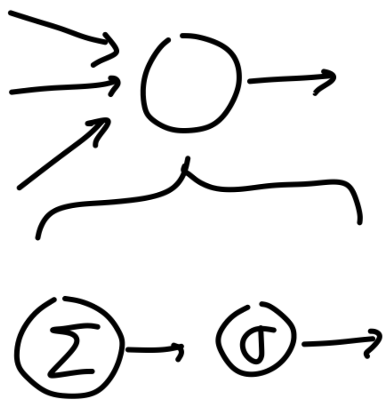
$$\sigma(x_1) = \begin{cases} +1 & x_1 \geq 0 \\ -1 & x_1 < 0 \end{cases}$$

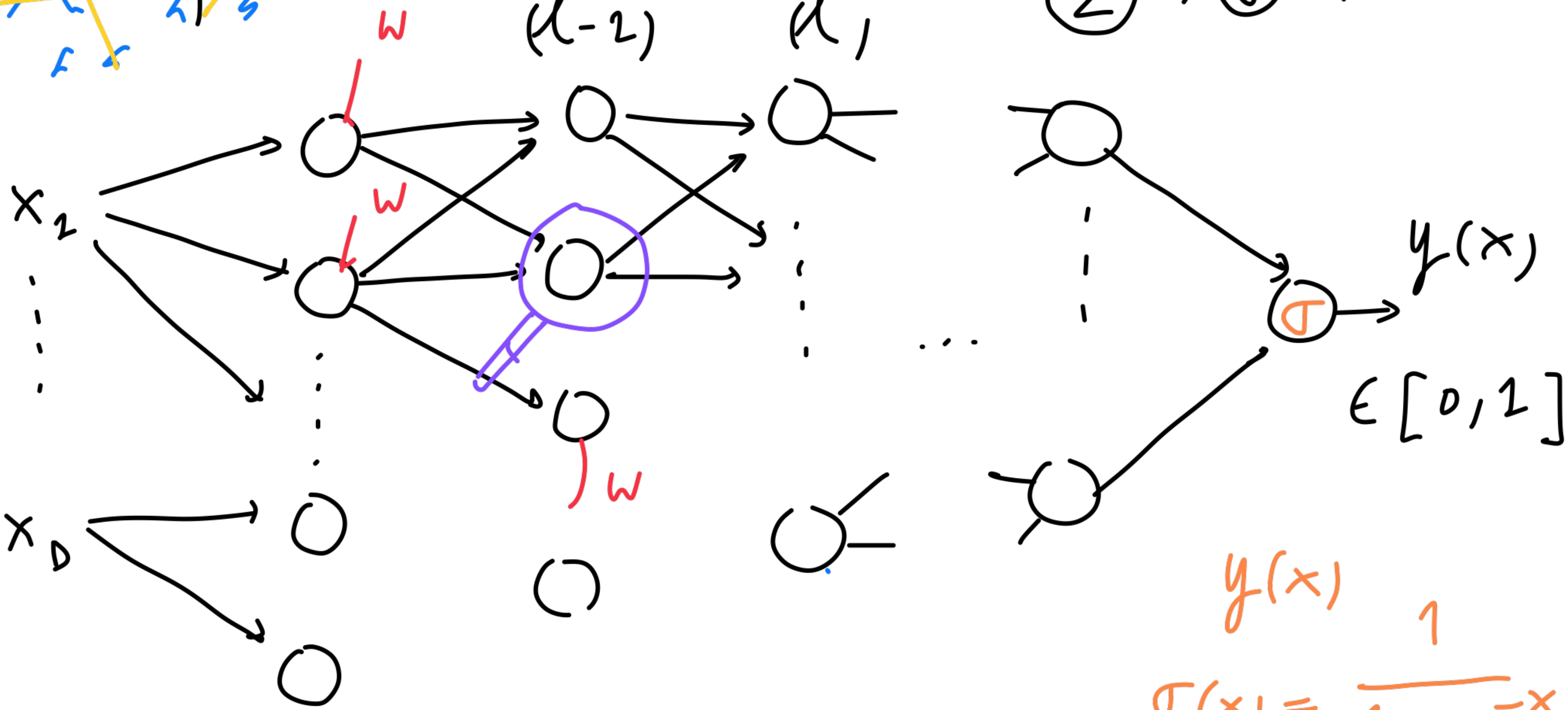
$$\sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$



$$\sigma(a) = |a|$$





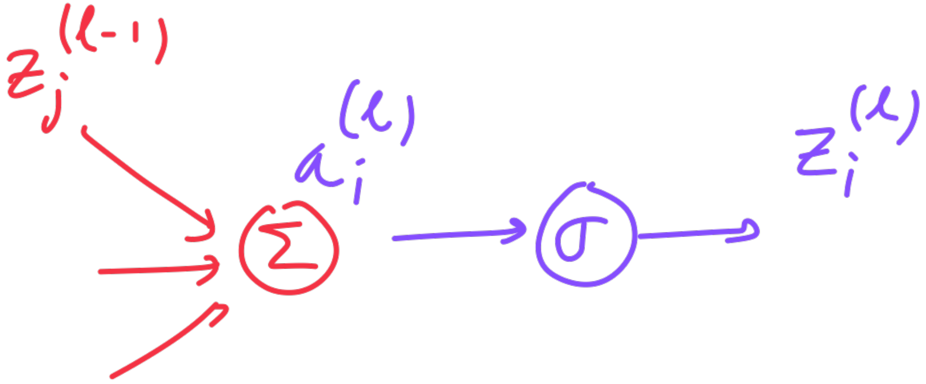


$$y(x)$$

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

fully connected feed forward

$$L(\beta) = -\,t^{(i)}\log\left(y(x^{(i)})\right) - (1-t^{(i)})\log\left(1-y(x^{(i)})\right)$$



$a_i^{(\ell)}$ = préactivation du neurone i de la couche ℓ

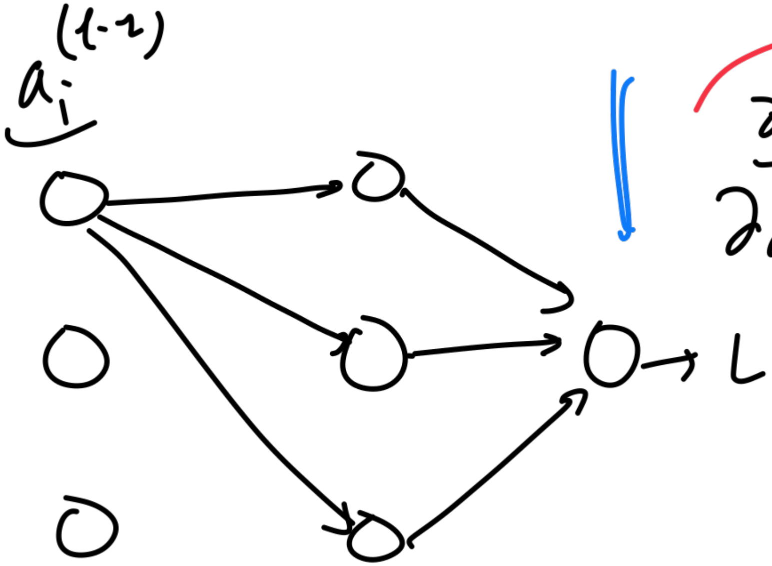
$z_i^{(\ell)}$ = post activation du neurone i de la couche ℓ

$$z_i^{(\ell)} = \sigma(a_i^{(\ell)}) \qquad \longrightarrow \qquad a_i^{(\ell)} = w_{i0} + \sum_{j=1}^{N_{\ell-1}} w_{ij}^{(\ell)} z_j^{(\ell-1)}$$

$w_{ij}^{(\ell)}$ = $j^{\text{ème}}$ coeff du neurone i de la couche ℓ

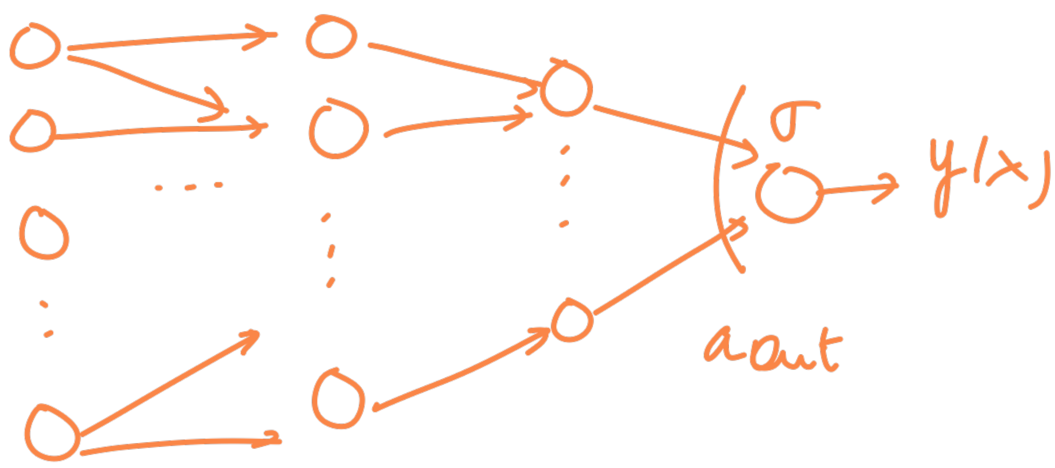
$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \left(\underbrace{\frac{\partial L}{\partial a_i^{(l)}}}_{\delta_i^{(l)}} \cdot \underbrace{\frac{\partial a_i^{(l)}}{\partial w_{ij}^{(l)}}}_{z_j^{(l-1)}} \right.$$

N_l = nombre de neurones de la couche l



$$\overbrace{\frac{\partial L}{\partial a_i^{(l-1)}}}^{\delta_i^{(l-1)}} = \sum_{j=1}^{N_l} \underbrace{\frac{\partial L}{\partial a_j^{(l)}}}_{\delta_j^{(l)}} \frac{\partial a_j^{(l)}}{\partial a_i^{(l-1)}} \qquad w_{ji}^{(l)} \sigma'(a_i^{(l-1)})$$

$$a_j^{(l)} = w_{j0}^{(l)} + \sum_{i=1}^{N_l} w_{ji}^{(l)} z_i^{(l-1)} = w_{j0} + \sum_{i=1}^{N_l} w_{ji}^{(l)} \sigma(a_i^{(l-1)})$$



$$L(w) = -t^{(i)} \log\left(y(x^{(i)})\right) - (1-t^{(i)}) \log\left(1-y(x^{(i)})\right)$$

$$= -t^{(i)} \log\left(\sigma(a_{out})\right) - (1-t^{(i)}) \log\left(1-\sigma(a_{out})\right)$$

$$\delta_{out} = \frac{\partial L}{\partial a_{out}} = -t^{(i)} \frac{\cancel{\sigma}(1-\sigma)}{\cancel{\sigma}} + (1-t^{(i)}) \frac{\sigma\cancel{(1-\sigma)}}{\cancel{1-\sigma}}$$

$$= -t^{(i)}(1-\sigma) + (1-t^{(i)})\sigma = \sigma - t^{(i)} \equiv \delta_{out}$$

BACKPROPAGATION pour le calcul du gradient dans les réseaux de neurones

Etant donné $\{x^{(i)}, t^{(i)}\}$

1) FORWARD PROPAGATE + Calculate every $a_i^{(\ell)}$ $z_i^{(\ell)}$ pour tous i, ℓ

2) Calculer $\delta_{out} = y(x^{(i)}) - t^{(i)}$

3) Rétropager δ_{out} et calculer tous les $\delta_i^{(\ell)}$ pour tous i, ℓ

$$\delta_i^{(\ell-1)} = \sum_{j=1}^{N_\ell} \delta_j^{(\ell)} w_{ji}^{(\ell)} \sigma'(a_i^{(\ell-1)})$$

4) Calculer les gradients

$$\frac{\partial L}{\partial w_{ij}^{(\ell)}} = \delta_i^{(\ell)} z_j^{(\ell-1)}$$

5) $$w_{ij}^{(\ell)} \leftarrow w_{ij}^{(\ell)} - \eta \frac{\partial L}{\partial w_{ij}^{(\ell)}}$$