$X \xrightarrow{\text{L couches à activation ReLu}} A_L \xrightarrow[\text{fonction de décision}]{\text{softmax} \; +} \hat{Y}$ à comparer avec Y

image vectorisée — prédiction — "vrais" labels

$ReLu(x) = max(0, x)$

$$softmax(X) = \left(\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \right)_{1 \leq i \leq n}$$

X : vecteur taille n

fonction de décision : ~~pred~~ $S(X) = (x_i = 1 \text{ si } x_i = max(x_j)_{1 \leq j \leq n} \; ; 0 \text{ sinon})$

Fonction de perte : $L = -\frac{1}{m} \sum_{k=1}^{m} \left({}^t y_k \log(\hat{y}_k) + {}^t(1 - y_k) \log(1 - \hat{y}_k) \right)$: (produit scalaire np. vdot.)

m le nombre d'exemples, (y_k) les labels, colonnes de Y

$(\hat{y}_k)$ les prédictions, colonnes de $\hat{Y}$

On note $z^1 = W_1 X + b_1$

$a_1 = relu(z_1)$; $a_0 = x$

$\forall \ell \in [\![2, L]\!]$: $z_\ell = W_\ell a_{\ell-1} + b_\ell$

$a_\ell = relu(z_\ell)$

$\hat{y} = softmax(a_L)$

$\boxed{\text{Objectif :}}$ calculer $\frac{\partial L}{\partial W_\ell}$ et $\frac{\partial L}{\partial b_\ell}$, pour $\ell \in [\![1, L]\!]$

étapes : calculer $\frac{\partial L}{\partial \hat{y}}$, $\frac{\partial L}{\partial a_L}$, $\frac{\partial L}{\partial z_L}$, $\frac{\partial L}{\partial z_\ell}$ et $\frac{\partial L}{\partial a_\ell}$, pour $\ell \in [\![1, L-1]\!]$

$\frac{\partial L}{\partial \hat{y}} = \frac{1}{m} \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} \right)$ (division coordonnée à coordonnée, np. divide)

$\frac{\partial L}{\partial a_L} = \frac{1}{m} \left((1-y) * \hat{y} - y * (1 - \hat{y}) \right)$ ($*$: multiplication coordonnée à coordonnée, np. multiply)

$\forall \ell \in [\![1, L]\!]$: $\frac{\partial L}{\partial z_\ell} = \frac{\partial L}{\partial a_\ell} * \mathbb{1}_{\mathbb{R}_+^*}(z_\ell)$

$\forall \ell \in [\![1, L-1]\!]$: $\frac{\partial L}{\partial a_\ell} = {}^T W_{\ell+1} \frac{\partial L}{\partial z_{\ell+1}}$ (produit matriciel np. dot)

d'où $\forall \ell \in [\![1, L]\!]$

$\frac{\partial L}{\partial W_\ell} = \frac{\partial L}{\partial z_\ell} \, {}^T a_{\ell-1}$

$\frac{\partial L}{\partial b_\ell} = $ np. sum $\left(\frac{\partial L}{\partial z_\ell}, \text{axis} = 1 \right)$

Démonstration :

- $\frac{\partial L}{\partial \hat{y}} = \left(\frac{\partial L}{\partial \hat{y}_{li}}\right)_{\substack{1 \leq l \leq m \\ 1 \leq i \leq 10}}$ $\hat{y}$ matrice de taille $m \times 10$, $\hat{y}_{li}$ ses coordonnées, de même pour y

$$\left(\frac{\partial L}{\partial \hat{y}_{li}}\right) = -\frac{1}{m} \sum_{k=1}^{m} \underbrace{\sum_{j=1}^{10} y_{kj} \frac{\partial \log \hat{y}_{kj}}{\partial \hat{y}_{li}}}_{\frac{y_{ki}}{\hat{y}_{li}} \delta_{kl}} + \underbrace{\sum_{j=1}^{10} (1 - y_{kj}) \frac{\partial \log(1-\hat{y}_{kj})}{\partial \hat{y}_{li}}}_{-\frac{1-y_{kj}}{1-\hat{y}_{li}} \delta_{kl}} = \frac{1}{m}\left(\frac{1-y_{li}}{1-\hat{y}_{li}} - \frac{y_{li}}{\hat{y}_{li}}\right)$$

d'où $\boxed{\frac{\partial L}{\partial \hat{y}} = \frac{1}{m}\left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right)}$ (np. divide)

- $\frac{\partial L}{\partial a_L} = \left(\frac{\partial L}{\partial a_{L_i}}\right)_{1 \leq i \leq 10}$

$$\frac{\partial L}{\partial a_{L_i}} = \sum_{k=1}^{10} \frac{\partial L}{\partial \hat{y}_k} \times \boxed{\frac{\partial \hat{y}_k}{\partial a_{L_i}}} = \frac{\partial L}{\partial \hat{y}_i} * \hat{y}_i * (1 - \hat{y}_i)$$

$$\frac{\partial}{\partial a_{L_i}} (\text{softmax } a_{L_k}) = \text{softmax } a_{L_i} * (1 - \text{softmax } a_{L_i})$$
$$= \hat{y}_i * (1 - \hat{y}_i)\, \delta_{ki}$$

d'où $\boxed{\frac{\partial L}{\partial a_L} = \frac{1}{m}\left((1-y) * \hat{y} - y * (1-\hat{y})\right)}$ (np. multiply)

- $\frac{\partial L}{\partial z_l} = \left(\frac{\partial L}{\partial z_{l_i}}\right)_{1 \leq i \leq \text{taille couche}}$ l fixé, $\in [\![1, L]\!]$

$$\frac{\partial L}{\partial z_{l_i}} = \sum_k \frac{\partial L}{\partial a_{l_k}} \boxed{\frac{\partial a_{l_k}}{\partial z_{l_i}}} = 1\!\!1_{\mathbb{R}_*^+}(z_{l_i}) * \frac{\partial L}{\partial a_{l_i}}$$

$$\frac{\partial}{\partial z_{l_i}} \text{relu}(z_{l_k}) = 1\!\!1_{\mathbb{R}_*^+}(z^l{}_i)\, \delta_{ik}$$

d'où $\boxed{\frac{\partial L}{\partial z_l} = \frac{\partial L}{\partial a_l} * 1\!\!1_{\mathbb{R}_*^+}(z_l)}$ (np. multiply) $\forall l \in [\![1, L]\!]$

- $\frac{\partial L}{\partial a_\ell} = \left(\frac{\partial L}{\partial a_{\ell\,i}}\right)_{1 \leq i \leq \text{taille couche}}$ $\quad$ ℓ fixé, $\in [\![1, L-1]\!]$

ligne i

$$\frac{\partial L}{\partial a_{\ell\,i}} = \sum_k \frac{\partial L}{\partial (z_{\ell+1})_k} \times \boxed{\frac{\partial (z_{\ell+1})_k}{\partial a_{\ell\,i}}} = \sum_k \left(\frac{\partial L}{\partial z_{\ell+1}}\right)_k W^{\ell+1}_{ki} = \sum_k ({}^tW^{\ell+1})_{ik} \left(\frac{\partial L}{\partial z_{\ell+1}}\right)_k = L_i\left({}^TW^{\ell+1} \frac{\partial L}{\partial z_{\ell+1}}\right)$$

$\downarrow$

$W^{\ell+1}_{ki}$ car $(z_{\ell+1})_k = \sum_{t=1}^{\dots} W^{\ell+1}_{kt} (a_\ell)_t + b_{\ell\,k}$

d'où $\boxed{\frac{\partial L}{\partial a_\ell} = {}^TW_{\ell+1} \frac{\partial L}{\partial z^{\ell+1}}}$ (np. dot) $\forall \ell \in [\![1, L-1]\!]$

Avec ces 4 resultats, on peut obtenir le $\frac{\partial L}{\partial W_\ell}$ et $\frac{\partial L}{\partial b_\ell}$

- $\frac{\partial L}{\partial W_\ell} = \left(\frac{\partial L}{\partial W^\ell_{ij}}\right)_{i,j}$, W_ℓ matrice dont la taille dépend de la taille des couches $\ell-1$ et ℓ $\quad \forall \ell \in [\![1, L]\!]$

$$\frac{\partial L}{\partial W^\ell_{ij}} = \sum_k \frac{\partial L}{\partial (z^\ell)_k} \times \boxed{\frac{\partial (z^\ell)_k}{\partial (W^\ell)_{ij}}} = \frac{\partial L}{\partial (z^\ell)_i} \times (a_{\ell-1})_j$$

$\downarrow$

$$\frac{\partial}{\partial W^\ell_{ij}}\left(\sum_u W_{ku} a^{\ell-1}_u\right) = \delta_{ik} a^{\ell-1}_j$$

donc $\boxed{\frac{\partial L}{\partial W_\ell} = \frac{\partial L}{\partial z_\ell} \cdot {}^ta_{\ell-1}}$ (produit matriciel np.dot) $\quad \forall \ell \in [\![1, L]\!]$

$a_0 = x$.

- $\frac{\partial L}{\partial b_\ell} = \left(\frac{\partial L}{\partial (b^\ell)_i}\right)_i$ $\quad$ b_ℓ vecteur de taille de la taille de la couche ℓ

$$\frac{\partial L}{\partial (b_\ell)_i} = \sum_k \frac{\partial L}{\partial (z_\ell)_k} \boxed{\frac{\partial (z_\ell)_k}{\partial (b_\ell)_i}} = \frac{\partial L}{\partial (z_\ell)_i}$$

$\downarrow$

$= \delta_{ki}$

$\rightarrow$ taille $[1 \times m]$: il faut sommer les lignes

$\rightarrow \sum_{j=1}^{m} \frac{\partial L}{\partial (z_\ell)_{ij}}$ car le biais s'ajoute à Wa par broadcasting (pté de la librairie numpy)

d'où $\frac{\partial L}{\partial b_\ell} = \sum_{k=1}^{m} \left(\frac{\partial L}{\partial z}\right)_k$ (vecteur de taille couche ℓ) encore noté $\boxed{\frac{\partial L}{\partial b_\ell} = \text{np.sum}\left(\frac{\partial L}{\partial z_\ell}, \text{axis} = 1\right)}$