RDFIA TME 5-6
Introduction aux réseaux de neurones
Notes additionnelles

Les batch seront stockés dans des matrices avec 1 exemple par ligne. Par exemple pour l'entrée on a une matrice X de taille $N \times n_x$ pour N exemples chacun de dimension n_x. On fera de même pour les résultats intermédiaires. Ainsi, la matrice de poids W_h sera par exemple de taille $n_h \times n_x$.

$\mathrm{sum}_{\textsf{line}}(X)$ avec X de taille $N \times n_x$ fait une somme par ligne et retourne un vecteur colonne de taille N.

Forward

Elementwise

$$\begin{cases} \tilde{h}_i = \sum_j W_{h,ij} x_j + b_{h,i} \\ h_i = \tanh(\tilde{h}_i) \\ \tilde{y}_i = \sum_j W_{y,ij} h_j + b_{y,i} \\ y_i = \mathrm{SoftMax}(\tilde{y}_i) = \dfrac{e^{\tilde{y}_i}}{\sum_j e^{\tilde{y}_j}} \end{cases}$$

Vectoriel

$$\begin{cases} \tilde{h} = W_h x + b_h \\ h = \tanh(\tilde{h}) \\ \tilde{y} = W_y h + b_y \\ y = \mathrm{SoftMax}(\tilde{y}) \end{cases}$$

Vectoriel par batch

$$\begin{cases} \tilde{H} = X W_h^\top + \mathrm{repmat}_{n \textsf{ lines}}(b_h^\top) \\ H = \tanh(\tilde{H}) \\ \tilde{Y} = H W_y^\top + \mathrm{repmat}_{n \textsf{ lines}}(b_y^\top) \\ Y = \mathrm{SoftMax}_{\textsf{line}}(\tilde{Y}) \end{cases}$$

Loss

$$\begin{cases} \ell(y, \tilde{y}) = -\sum_i y_i \log \hat{y}_i = -\sum_i y_i \tilde{y}_i + \log \sum_j e^{\tilde{y}_j} \\ \mathcal{L}(Y, \hat{Y}) = \dfrac{1}{N} \sum_k \sum_i Y_{k,i} \log \hat{Y}_{k,i} = \mathrm{mean}_{\textsf{col}}(\mathrm{sum}_{\textsf{line}}(Y \log \hat{Y})) \end{cases}$$

Backward

Elementwise

$$\begin{cases} \delta_{y,i} = \dfrac{\partial \ell}{\partial \tilde{y}_i} = \hat{y}_i - y_i \\ \dfrac{\partial \ell}{\partial W_{y,ij}} = \delta_{y,i} h_j \\ \dfrac{\partial \ell}{\partial b_{y,i}} = \delta_{y,i} \\ \delta_{h,i} = \dfrac{\partial \ell}{\partial \tilde{h}_i} = (1 - h_i^2) \sum_j \delta_{y,j} W_{y,ji} \\ \dfrac{\partial \ell}{\partial W_{h,ij}} = \delta_{h,i} x_j \\ \dfrac{\partial \ell}{\partial b_{h,i}} = \delta_i^h \end{cases}$$

Vectoriel

$$\begin{cases} \nabla_{\tilde{y}} = \hat{y} - y \\ \nabla_{W_y} = \nabla_{\tilde{y}} h^\top \\ \nabla_{b_y} = \nabla_{\tilde{y}} \\ \nabla_{\tilde{h}} = W_y^\top \nabla_{\tilde{y}} \odot (1 - h^2) \\ \nabla_{W_h} = \nabla_{\tilde{h}} x^\top \\ \nabla_{b_h} = \nabla_{\tilde{h}} \end{cases}$$

Vectoriel par batch

$$\begin{cases} \nabla_{\tilde{Y}} = \hat{Y} - Y \\ \nabla_{W_y} = \nabla_{\tilde{Y}}^\top H \\ \nabla_{b_y} = \mathrm{sum}_{\textsf{col}}(\nabla_{\tilde{Y}})^\top \\ \nabla_{\tilde{H}} = \nabla_{\tilde{Y}} W_y \odot (1 - H^2) \\ \nabla_{W_h} = \nabla_{\tilde{H}}^\top X \\ \nabla_{b_h} = \mathrm{sum}_{\textsf{col}}(\nabla_{\tilde{H}})^\top \end{cases}$$