Linéaire

MSELoss

Forward

$MSE(y, \hat{y}) = ||\hat{y} - y||^2$

Backward

$\frac{\partial MSE}{\partial \hat{y}} = \frac{\partial \sum_i^N ||\hat{y}_i - y_i||^2}{\partial \hat{y}}$

On utilise le denominator layout

$$2 \begin{pmatrix} \hat{y}_{11} - y_{11} & \hat{y}_{12} - y_{12} & \dots & \hat{y}_{1d} - y_{1d} \\ \hat{y}_{21} - y_{21} & \hat{y}_{22} - y_{22} & \dots & \hat{y}_{2d} - y_{2d} \\ \dots & \dots & \dots & \dots \\ \hat{y}_{b1} - y_{b1} & \hat{y}_{b2} - y_{b2} & \dots & \hat{y}_{bd} - y_{bd} \end{pmatrix} = 2 \begin{pmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y2 \\ \dots \\ \hat{y}_b - y_b \end{pmatrix}$$

Donc:

$\frac{\partial MSE}{\partial \hat{y}} = -2(y - \hat{y})$

Module Linéaire

Backward update gradient

$\frac{\partial Loss}{\partial W^h} = \frac{\partial Loss}{\partial z^h} \frac{\partial z^h}{\partial W^h} = \delta^h \frac{\partial z^h}{\partial W^h}$

On utilise le denominator layout

$$z^h = \begin{pmatrix} z_1^h & z_2^h & \dots & z_{d'}^h \end{pmatrix} = \begin{pmatrix} \sum_i^d z_i^{h-1} w_{i1}^h & \sum_i^d z_i^{h-1} w_{i2}^h & \dots & \sum_i^d z_i^{h-1} w_{id'}^h \end{pmatrix}$$

$$W^h = \begin{pmatrix} w_1^h \\ w_2^h \\ \dots \\ w_{d'}^h \end{pmatrix} = \begin{pmatrix} w_{11}^h & w_{21}^h & \dots & w_{d1}^h \\ w_{12}^h & w_{22}^h & \dots & w_{d2}^h \\ \dots & \dots & \dots & \dots \\ w_{1d'}^h & w_{2d'}^h & \dots & w_{dd'}^h \end{pmatrix}$$

$$\frac{\partial z^h}{\partial W^h} = \begin{pmatrix} z_1^{h-1} & 0 & \dots & 0 & z_2^{h-1} & 0 & \dots & 0 & z_d^{h-1} & 0 & 0 & 0 \\ 0 & z_1^{h-1} & \dots & 0 & 0 & z_2^{h-1} & \dots & 0 & 0 & z_d^{h-1} & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z_1^{h-1} & 0 & 0 & \dots & z_2^{h-1} & 0 & 0 & 0 & z_d^{h-1} \end{pmatrix}$$

$\frac{\partial z^h}{\partial W^h} = \begin{pmatrix} z^{h-1} & 0 & \dots & 0 \\ 0 & z^{h-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z^{h-1} \end{pmatrix}$

Dimension (d', dd')

$\delta^h = \begin{pmatrix} \delta^h_1 \\ \delta^h_2 \\ \dots \\ \delta^h_{d'} \end{pmatrix}$

$\frac{\partial Loss}{\partial W^h} = \delta^h \frac{\partial z^h}{\partial W^h} = \begin{pmatrix} \delta^h_1 & \delta^h_2 & \dots & \delta^h_{d'} \end{pmatrix} \begin{pmatrix} z^{h-1} & 0 & \dots & 0 \\ 0 & z^{h-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z^{h-1} \end{pmatrix} = (\delta^h)^T z^{h-1}$

Dimension $(N, d') \times (d', dd') = (N, dd')$

Backward delta

$\frac{\partial Loss}{\partial z^{h-1}} = \frac{\partial Loss}{\partial z^h} \frac{\partial z^h}{\partial z^{h-1}} = \delta^h \frac{\partial z^h}{\partial z^{h-1}}$

$z^{h-1} = \begin{pmatrix} z^{h-1}_1 \\ z^{h-1}_2 \\ \dots \\ z^{h-1}_d \end{pmatrix}$

$z^h = \begin{pmatrix} z^h_1 & z^h_2 & \dots & z^h_{d'} \end{pmatrix} = \begin{pmatrix} \sum_i^d z^{h-1}_i w^h_{i1} & \sum_i^d z^{h-1}_i w^h_{i2} & \dots & \sum_i^d z^{h-1}_i w^h_{id'} \end{pmatrix}$

$\frac{\partial z^h}{\partial z^{h-1}} = \begin{pmatrix} w^h_{11} & w^h_{12} & \dots & w^h_{1d'} \\ w^h_{21} & w^h_{22} & \dots & w^h_{2d'} \\ \dots & \dots & \dots & \dots \\ w^h_{d1} & w^h_{d2} & \dots & w^h_{dd'} \end{pmatrix} = (W^h)^T$

$\frac{\partial Loss}{\partial z^{h-1}} = \delta^h \frac{\partial z^h}{\partial z^{h-1}} = \delta^h (W^h)^T$

Comme $z =< x.w >$, donc on utilise la transition de W. Or:

$\frac{\partial Loss}{\partial z^{h-1}} = \delta^h \frac{\partial z^h}{\partial z^{h-1}} = \delta^h W^h$

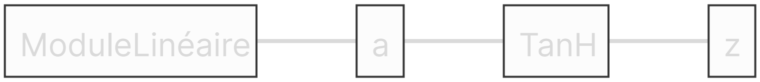
Non-linéaire

Tangente hyperbolique

Forward

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

Backward



$\frac{\partial Loss}{\partial a^h} = \frac{\partial Loss}{\partial z^h} \frac{\partial z^h}{\partial a^h} = \delta^h \frac{\partial z^h}{\partial a^h}$

$\frac{\partial z^h}{\partial a^h} = \frac{\partial \frac{e^{a^h} - e^{-a^h}}{e^{a^h} + e^{-a^h}}}{\partial a^h} = \frac{(e^{a^h} - e^{-a^h})'(e^{a^h} + e^{-a^h}) - (e^{a^h} + e^{-a^h})'(e^{a^h} - e^{-a^h})}{(e^{a^h} + e^{-a^h})^2} = \frac{(e^{a^h} + e^{-a^h})^2 - (e^{a^h} - e^{-a^h})^2}{(e^{a^h} + e^{-a^h})^2} = 1 - \frac{(e^{a^h} - e^{-a^h})^2}{(e^{a^h} + e^{-a^h})^2}$

Or:
$\frac{\partial z^h}{\partial a^h} = 1 - tanH^2(a^h)$

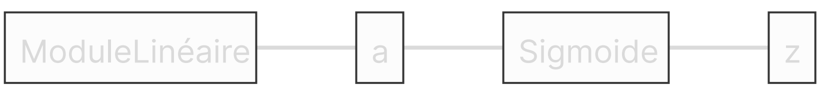
Donc:
$\frac{\partial Loss}{\partial a^h} = \delta^h \frac{\partial z^h}{\partial a^h} = \delta^h(1 - tanH^2(a^h))$

Sigmoide

Forward

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

Backward



$\frac{\partial Loss}{\partial a^h} = \frac{\partial Loss}{\partial z^h} \frac{\partial z^h}{\partial a^h} = \delta^h \frac{\partial z^h}{\partial a^h}$

$\frac{\partial z^h}{\partial a^h} = \frac{\partial \frac{1}{1+e^{-a^h}}}{\partial a^h} = \frac{-(1+e^{-a^h})'}{(1+e^{-a^h})^2} = \frac{e^{-a^h}}{(1+e^{-a^h})^2} = \frac{1}{1+e^{-a^h}} \frac{e^{-a^h}}{1+e^{-a^h}} = \frac{1}{1+e^{-a^h}} \frac{1+e^{-a^h}-1}{1+e^{-a^h}} = \frac{1}{1+e^{-a^h}}(1 - \frac{1}{1+e^{-a^h}}) = \sigma(a^h)($

Softmax et Cout Entropique

Soft-max

$Softmax(x) = \frac{e^x}{\sum_i^d e^i}$

Coût Cross-Entropique

Soit y le vecteur supervision codé en one-hot.
Par exemple, si y_i est de 3ème classe.

$y_i = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$

Et $\hat{y}_i$ est la vecteur de prédiction

$\hat{y}_i = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.3 \\ 0.1 \end{pmatrix}$

$CE(y_i, \hat{y}_i) = - < y_i . \hat{y}_i >= -(0 * 0.2 + 0 * 0.4 + 1 * 0.3 + 0 * 0.1) = -0.3$

Nous allons noter y comme l'indice de la classe à prédire.

$CE(y, \hat{y}) = -\hat{y}_y$

Combinaison de Softmax et coût cross-entropique

Afin d'éviter des instabilités numériques, on enchaîne un $Softmax$ passé au logarithme (l $ogSoftMax$) et un coût cross entropique comme une combinaison.

Forward

$CE(y, \hat{y}) = -\log \frac{e^{\hat{y}_y}}{\sum_{i=1}^{K} e^{\hat{y}_i}} = -\hat{y}_y + \log \sum_{i=1}^{K} e^{\hat{y}_i}$

Backward

$\frac{\partial Loss}{\partial \hat{y}} = \frac{\partial \sum_{i=1}^{N} -\hat{y}_{i,y_i} + log \sum_{j=1}^{K} e^{\hat{y}_{i,j}}}{\partial \hat{y}}$

On utilise le denominator layout

$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_N \end{pmatrix} = \begin{pmatrix} \hat{y}_{11} & \hat{y}_{12} & \dots & \hat{y}_{1K} \\ \hat{y}_{21} & \hat{y}_{22} & \dots & \hat{y}_{2K} \\ \dots \\ \hat{y}_{N1} & \hat{y}_{N2} & \dots & \hat{y}_{NK} \end{pmatrix}$

$\frac{\partial Loss}{\partial \hat{y}_{kf}} = \frac{\partial \sum_{i=1}^{N} -\hat{y}_{i,y_i} + \log \sum_{j=1}^{K} e^{\hat{y}_{i,j}}}{\partial \hat{y}_{kf}} = \frac{\partial (-\hat{y}_{k,y_k} + \log \sum_{j=1}^{K} e^{\hat{y}_{k,j}})}{\partial \hat{y}_{kf}}$

Si $y_k = f$:

$\frac{\partial Loss}{\partial \hat{y}_{kf}} = -1 + \frac{e^{\hat{y}_{kf}}}{\sum_{j=1}^{K} e^{\hat{y}_{k,j}}}$

Si $y_k \neq f$:

$\frac{\partial Loss}{\partial \hat{y}_{kf}} = \frac{e^{\hat{y}_{kf}}}{\sum_{j=1}^{K} e^{\hat{y}_{k,j}}}$

Donc:

$$\frac{\partial Loss}{\partial \hat{y}} = \begin{pmatrix} -y_{11} + \frac{e^{\hat{y}_{11}}}{\sum_{j=1}^{K} e^{\hat{y}_{1,j}}} & -y_{12} + \frac{e^{\hat{y}_{12}}}{\sum_{j=1}^{K} e^{\hat{y}_{1,j}}} & \dots & -y_{1K} + \frac{e^{\hat{y}_{1K}}}{\sum_{j=1}^{K} e^{\hat{y}_{1,j}}} \\ -y_{21} + \frac{e^{\hat{y}_{21}}}{\sum_{j=1}^{K} e^{\hat{y}_{2,j}}} & -y_{22} + \frac{e^{\hat{y}_{22}}}{\sum_{j=1}^{K} e^{\hat{y}_{2,j}}} & \dots & -y_{2K} + \frac{e^{\hat{y}_{2K}}}{\sum_{j=1}^{K} e^{\hat{y}_{2,j}}} \\ \dots & \dots & \dots & \dots \\ -y_{N1} + \frac{e^{\hat{y}_{N1}}}{\sum_{j=1}^{K} e^{\hat{y}_{N,j}}} & -y_{N2} + \frac{e^{\hat{y}_{N2}}}{\sum_{j=1}^{K} e^{\hat{y}_{N,j}}} & \dots & -y_{NK} + \frac{e^{\hat{y}_{NK}}}{\sum_{j=1}^{K} e^{\hat{y}_{N,j}}} \end{pmatrix}$$

$$\frac{\partial Loss}{\partial \hat{y}} = -y + Softmax(\hat{y})$$

Se compresser

Encodage

Linéaire.256_100 — TanH — Linéaire.100_10 — Tanh

Décodage

Linéaire.10_100 — TanH — Linéaire.100_256 — Sigmoide

Coût Cross-entropique binaire

Forward

$$BCE(y, \hat{y}) = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

Backward

$$\frac{\partial Loss}{\partial \hat{y}} = -(\frac{y}{\hat{y}} + \frac{y-1}{1-\hat{y}})$$