# MLP PER MUTUALLY EXCLUSIVE MULTI-CLASS CLASSIFICATION

Fin ora abbiamo risolto il problema di classificare più classi, ma non in modo esclusivo. Il che vuol dire che in uscita potevamo avere più di una classe predetta.

Esempio: Classificare gli oggetti di una singola immagine.

## SOFT-MAX

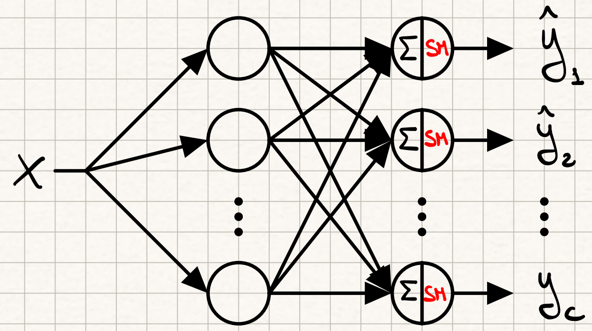
Come classificare in modo esclusivo? La risposta è usare una funzione di attivazione particolare, chiamata softmax, in grado di trasformare un vettore di ingresso in una distribuzione di probabilità (la somma degli output della softmax fa 1):

$$\text{Softmax}(\vec{z}) = \begin{bmatrix} \dfrac{e^{z_1}}{\sum_{j=1}^{c} e^{z_j}} \\ \vdots \\ \dfrac{e^{z_c}}{\sum_{j=1}^{c} e^{z_j}} \end{bmatrix}$$

Esempio

$$\begin{bmatrix} 1{,}3 \\ 5{,}1 \\ 2{,}2 \\ 0{,}7 \\ 1{,}1 \end{bmatrix} \xrightarrow{\text{softmax}} \begin{array}{cc} \begin{bmatrix} 0{,}02 \\ 0{,}90 \\ 0{,}05 \\ 0{,}01 \\ 0{,}02 \end{bmatrix} & \begin{array}{c} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{array} \end{array}$$

# Modello




# CATEGORICAL CROSS-ENTROPY

$$\mathcal{L}^{CCE} = \sum_{k=1}^{C} -y_k \log(\hat{y}_k)$$

È la loss usata per allenare modelli di classificazione esclusiva.

# DERIVATA SOFTMAX

Calcoliamo le derivate parziali della softmax in due dimensioni $(x, y)$:

$$\text{softmax}(x,y) = \left[ \frac{e^x}{e^x + e^y} \, ; \, \frac{e^y}{e^x + e^y} \right]$$

$$\text{softmax}_x(x,y) = \left[ \frac{e^x e^y}{(e^x + e^y)^2} \, ; \, -\frac{e^x e^y}{(e^x + e^y)^2} \right]$$

$$\text{softmax}_y(x,y) = \left[ -\frac{e^x e^y}{(e^x + e^y)^2} \, ; \, \frac{e^x e^y}{(e^x + e^y)^2} \right]$$

Per calcolare la derivata di un rapporto fra due funzioni abbiamo la seguente formula:

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)}{g(x)} - \frac{f(x)g'(x)}{g^2(x)}$$

Sostituendo $f(x) = e^x$ e $g(x) = e^x + e^y$ abbiamo che

$$\left(\frac{e^x}{e^x+e^y}\right)' = \frac{e^x}{e^x+e^y} - \left(\frac{e^x}{e^x+e^y}\right)^2 = \dots = \frac{e^x e^y}{(e^x+e^y)^2}$$

Considerando la derivata parziale di una sola componente della softmax rispetto alla prima variabile di preattivazione $z_1$:

$$\frac{\partial a_1}{\partial z_1} = \frac{e^{z_1}}{e^{z_1}+e^{z_2}} - \left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)^2 = a_1 - a_1^2 = a_1(1-a_1)$$

Rispetto alle altre preattivazioni, $z_2 \dots z_c$ abbiamo

$$\frac{\partial a_1}{\partial z_2} = -\frac{e^{z_1}e^{z_2}}{(e^{z_1}+e^{z_2})^2} = -a_1 a_2$$

## In generale

Possiamo dire che, la c-esima componente della softmax ha derivata, rispetto alla preattivazione $z_K$:

$$\frac{\partial a_c}{\partial z_K} = \begin{cases} a_c(1-a_c) & \text{se } c=K \\ -a_c a_K & \text{se } c \neq K \end{cases}$$

## Backpropagation

Usando la $\mathcal{L}^{CCE}$, calcoliamo le derivate:

$$\frac{\partial \mathcal{L}^{CCE}}{\partial W_1} \quad \text{e} \quad \frac{\partial \mathcal{L}^{CCE}}{\partial W_2}$$

Considerando i seguenti layer finali:

$$\tau Z_2 = W_2 A_1 \qquad \tau Z_2 \in \mathbb{R}^{C \times B}$$
$$\tau A_2 = \text{softmax}(\tau Z_2) \qquad \tau A_2 \in \mathbb{R}^{C \times B}$$

## Calcoliamo $\partial\mathcal{L}/\partial W_2$:

Dato un mini-Batch B, la loss diventa:

$$\mathcal{L}^{CCE} = \sum_{k=1}^{C}\sum_{b=1}^{B} -y_k \log(\hat{y}_k)$$

Considerando B=1 e la classe c-esima:

$$\frac{\partial \mathcal{L}^{CCE}}{\partial \hat{y}_c} = -\frac{y_c}{\hat{y}_c} \xRightarrow{\hat{y}_c = {}_zA_2^c} -\frac{y}{{}_zA_2^c}$$

Per calcolare $\partial\mathcal{L}/\partial W_2$ sviluppiamo la seguente chain rule:

$$\frac{\partial\mathcal{L}}{\partial W_2} = \underbrace{\frac{\partial \mathcal{L}}{\partial {}_zZ_2^c}}\;\underbrace{\frac{\partial {}_zZ_2^c}{\partial W_2}}_{A_1^T}$$

$$\frac{\partial\mathcal{L}}{\partial {}_zZ_2^c} = \sum_{k=1}^{Ce} \frac{\partial\mathcal{L}}{\partial {}_zA_2^k}\frac{\partial {}_zA_2^k}{\partial {}_zZ_2^c} = \underbrace{\frac{\partial\mathcal{L}}{\partial {}_zA_2^c}}\;\underbrace{\frac{\partial {}_zA_2^c}{\partial {}_zZ_2^c}} + \sum_{\substack{k=1\\k\neq c}}^{Ce} \underbrace{\frac{\partial\mathcal{L}}{\partial {}_zA_2^k}}\;\underbrace{\frac{\partial {}_zA_2^k}{\partial {}_zZ_2^c}} =$$

Sommatoria delle derivate di ogni componente della softmax

$$= \underbrace{-\frac{y_c}{\cancel{{}_zA_2^c}}}_{\partial\mathcal{L}/\partial\hat{y}_c} \cdot \underbrace{\cancel{{}_zA_2^c}\left(1-{}_zA_2^c\right)}_{\text{Derivata Softmax } k=c} + \sum_{k\neq c}^{Ce} \underbrace{\left(-\frac{y_k}{\cancel{{}_zA_2^k}}\right)}_{\partial\mathcal{L}/\partial\hat{y}_k}\underbrace{\left(-{}_zA_2^c\;\cancel{{}_zA_2^k}\right)}_{\text{Derivata della softmax } k\neq c}$$

$$= -y_c\left(1 - zA_2^c\right) + \sum_{k\neq c}^{ce} y_k \cdot zA_2^c =$$

$$= \underbrace{-y_c + y_c \cdot zA_2^c + zA_2^c \cdot \sum_{k\neq c}^{ce} y_k}_{\text{Mettiamo in evidenza } zA_2^c} =$$

$$= -y_c + zA_2^c \underbrace{\sum_{k=1}^{ce} y_k}_{=1}$$

$= 1$ per la definizione di softmax

$$\frac{\partial \mathcal{L}}{\partial zZ_2^c} = \boxed{zA_2^c - y_c}$$

Quindi, mettendo insieme:

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial zZ_2^c}\,\frac{\partial zZ_2^c}{\partial W_2} = \left(zA_2^c - y_c\right) A_1^T$$

Calcoliamo $\partial \mathcal{L} / \partial W_1$

$$\frac{\partial \mathcal{L}}{\partial W_1} = \underbrace{\frac{\partial \mathcal{L}}{\partial Z_2^c}}_{zA_2^c - y_c}\,\underbrace{\frac{\partial Z_2^c}{\partial A_1}}_{W_2^T}\,\underbrace{\frac{\partial A_1}{\partial zZ_1}}_{\sigma'(zZ_1)}\,\underbrace{\frac{\partial zZ_1}{\partial W_1}}_{A_0^T}$$

Nota: La semplificazione $\partial \mathcal{L} / \partial Z_2 = zA_2^c - y_c$ vale solo usando la coppia $\mathcal{L}^{CCE}$ e softmax.