Treinamento de redes neurais

Problema: $$\min_{w,b} R_m(w,b) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i; w,b); y_i)$$

onde

- ℓ é a função de perda, que mede o erro na classificação do dado (x_i, y_i).
  - ↳ p. ex., $\ell(h,y) = \frac{1}{2}\|h-y\|^2$;
    $\ell(h,y) = \log(1+e^{-h^T y})$.
- h é a função de predição, cuja imagem é

a resposta dada à entrada x.

- h será composta de uma aplicação afim com uma função não linear $\underline{a}$ (função de ativação):

$$h(x; W, b) = a(Wx + b).$$

Um exemplo de função de ativação é a sigmoid: $a(z) = 1/(1+e^{-z})$.

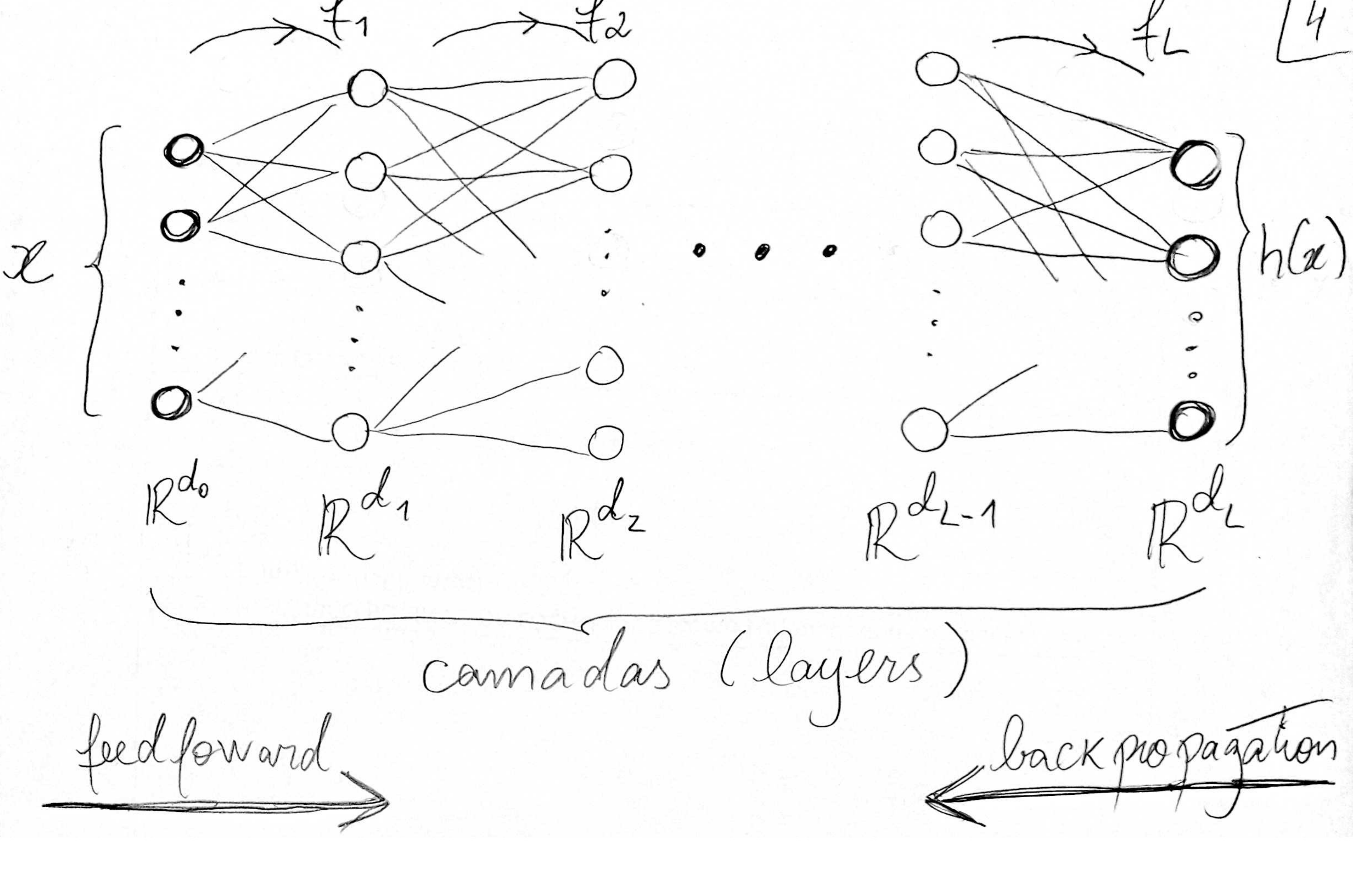
<u>Notação</u>: $a(v) = \begin{bmatrix} a(v_1) \\ \vdots \\ a(v_n) \end{bmatrix}$, $v \in \mathbb{R}^n$.

$\boxed{\text{EX:}}$

$$\min_{w,b} R_m(w,b) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}\| a(Wx_i + b) - y_i \|^2$$

Queremos aplicar gradiente estocástico, logo precisamos de avaliar, dados W, b, os termos da soma e seus gradientes em relação à (w,b).

Lembre-se que $h(x; W, b) = a(Wx + b)$ é representado em um grafo, como aplicação sucessivas de funções:
$h(x) = f_L(f_{L-1}(\dots f_1(x)\dots))$, onde cada f_i é da



f1
f2
fL
x
h(x)
R^d0
R^d1
R^d2
R^dL-1
R^dL
camadas (layers)
feedforward
backpropagation

forma $a(Wx+b)$.

Feedforward (avaliar $R_m(w,b)$)

Para cada $i \in \{1,\dots,m\}$, escrevemos o termo

$C = \frac{1}{2}\|a(Wx_i+b) - y_i\|^2$ como

$$C = \frac{1}{2}\|a^{[L]} - y\|^2 \quad \text{(omitimos } i\text{)}$$

onde $a^{[L]}$ é a saída da rede neural.

Ele é obtido propagando a entrada x pela rede (W e b estão fixados):

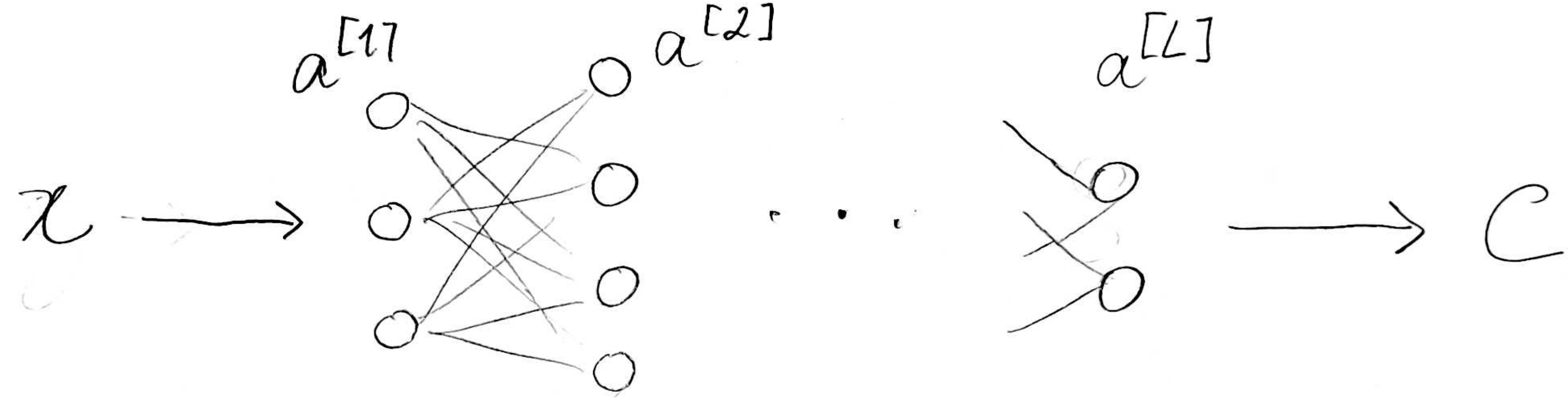
$$a^{[1]} = x$$

$$a^{[2]} = a(W^{[2]} a^{[1]} + b^{[2]})$$

$$a^{[3]} = a(W^{[3]} a^{[2]} + b^{[3]})$$

$$\vdots$$

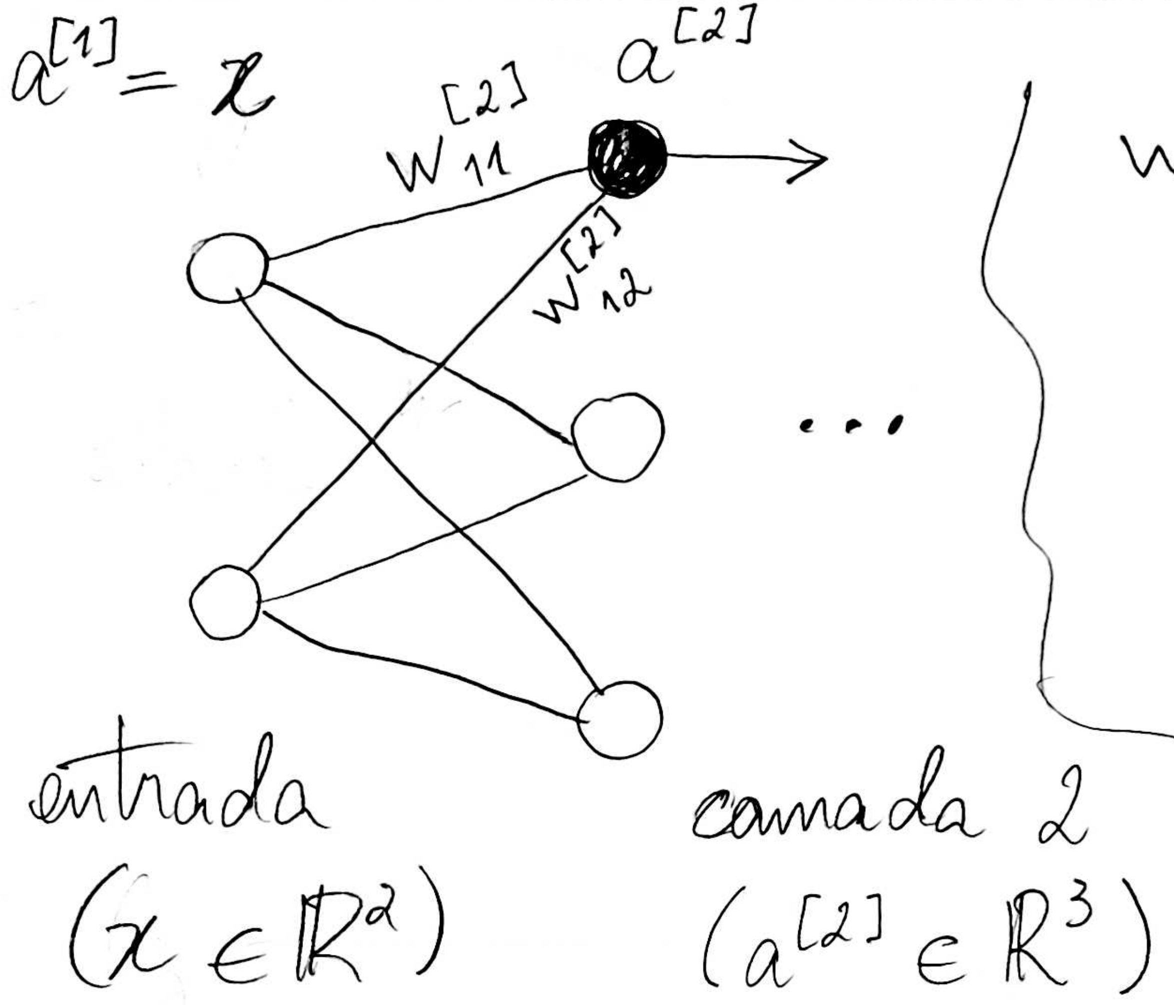
$$a^{[L]} = a(W^{[L]} a^{[L-1]} + b^{[L]})$$



Em geral,

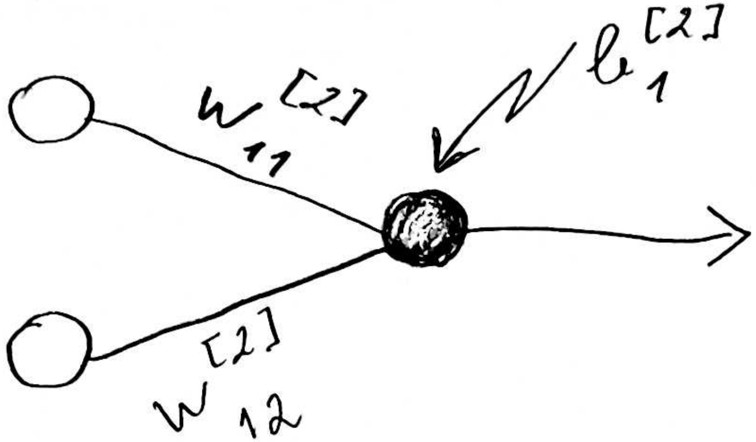
- $a^{[1]} = x$
- $a^{[l]} = \sigma\left(W^{[l]} a^{[l-1]} + b^{[l]}\right)$, $l = 2, \dots, L$.
- $C = \frac{1}{2} \| a^{[L]} - y \|^2$.

<u>Notação:</u> $W^{[l]}$, $b^{[l]}$ são os pesos e vieses dos neurônios na camada l. Note que a camada de entrada não possui w e b...



$w^{[l]}_{jk}$: peso associado ao neurônio j da camada l, fluxo do neurônio K da camada anterior

neurônio ● recebe o fluxo x, pondera com o vetor $w^{[2]}_1 \in \mathbb{R}^2$, adiciona o viés $b^{[2]}_1$ e passa o fluxo aplicando a ativação <u>a</u>.



$$a_1^{[2]} = a\left(w_1^{[2]^T} a^{[1]} + b_1^{[1]}\right).$$

$$a^{[1]} = x$$

Em geral, $a_j^{[l]} = a\left(\sum_k w_{jk}^{[l]} a_k^{[l-1]} + b_j^{[l]}\right)$

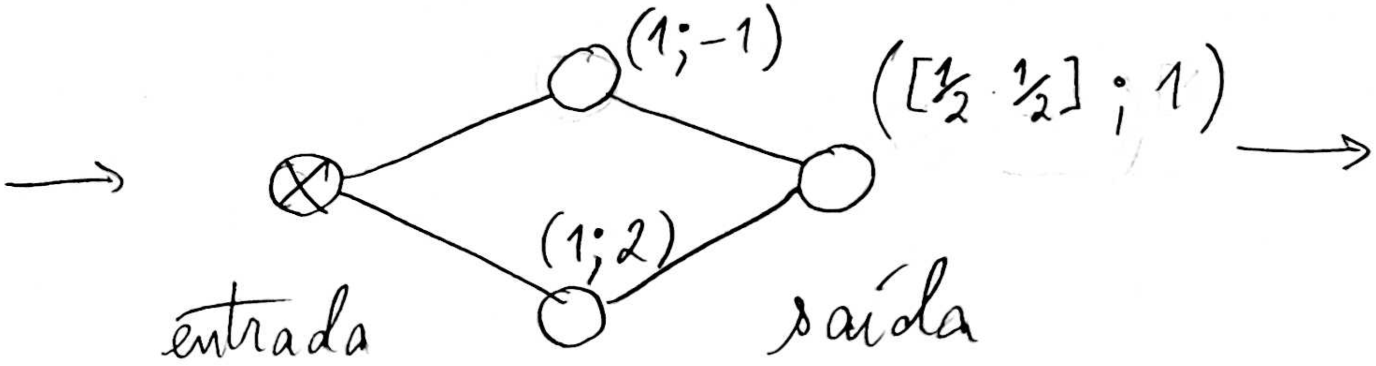
é a saída do neurônio j da camada l. Escrevendo em termos de matrizes,

$$a^{[\ell]} = a\left(W^{[\ell]} a^{[\ell-1]} + b^{[\ell]}\right), \text{ onde}$$

$W^{[\ell]}$ é matriz $n_\ell \times n_{\ell-1}$ e $b^{[\ell]} \in \mathbb{R}^{n_\ell}$, n_ℓ, $n_{\ell-1}$ número de neurônios nas camadas ℓ e $\ell-1$, respectivamente.

Exemplo: calcular $R_2(w,b)$ com 2 dados $(x_1, y_1) = (1,1)$ e $(x_2, y_2) = (2,0)$, no "ponto" $W^{[2]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $W^{[3]} = [2 \ 3]$, $b^{[2]} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$,

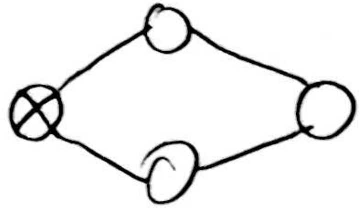
$b^{[3]} = [1]$, relativos à rede



(obs: (w, b) representa as matrizes $W^{[l]}$, $b^{[l]}$, por exemplo, empilhando as colunas em sequência).

Considere a função de ativação sigmoid $a(z) = 1/(1+e^{-z})$ em todos os neurônios:

Para o dado $(x_1, y_1) = (1,1)$:

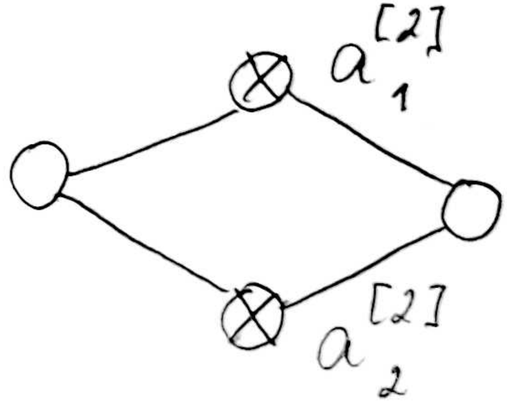


- $a^{[1]} = x_1 = [1]$

- $a^{[2]} = a(W^{[2]} a^{[1]} + b^{[2]})$

$$= a\left(\begin{bmatrix}1\\1\end{bmatrix}[1] + \begin{bmatrix}-1\\2\end{bmatrix}\right) = a\left(\begin{bmatrix}0\\3\end{bmatrix}\right)$$

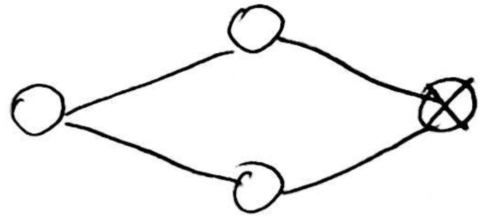
$$= \begin{bmatrix}\dfrac{1}{1+e^{0}}\\[2ex] \dfrac{1}{1+e^{-3}}\end{bmatrix} \approx \begin{bmatrix}0,5\\[2ex] -0,1398\end{bmatrix}$$



- $a^{[3]} = a(W^{[3]} a^{[2]} + b^{[3]})$

$$= a\left(\begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 0{,}5 \\ -0{,}1398 \end{bmatrix} + \begin{bmatrix} 1 \end{bmatrix}\right) = a(1{,}5807)$$

$$\approx \begin{bmatrix} -0{,}3033 \end{bmatrix}$$



- $C_1 = \frac{1}{2} | a^{[3]} - y_1 |^2 \approx 0{,}8493$

Para o dado $(x_2, y_2) = (2, 0)$:

Faça as contas e conclua que $C_2 \approx 0{,}0969$. (exercício). Assim

$$R_2(w, b) \approx \frac{1}{2}(0{,}8493 + 0{,}0969) \approx 0{,}4731$$

Obs: note que ao avaliar $C_i = \frac{1}{2}\| a(x_i; w, b) - y_i\|^2$, computamos $a^{[l]}$, $1 \leq l \leq L$, e

$$z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}, \quad 2 \leq l \leq L.$$

Back propagation (avaliar $\nabla R_m(w, b)$)

No método do gradiente estocástico devemos avaliar um ou mais (poucos) gradientes de termos

$$C = C_i = \frac{1}{2} \| a(Wx_i + b) - y_i \|^2$$

em relação à (w, b). Vamos omitir o "i" do dado. Precisamos computar as parciais

$$\frac{\partial C}{\partial w_{jk}^{[\ell]}} \quad e \quad \frac{\partial C}{\partial b_j^{[\ell]}}, \; \forall j, k, \; 2 \leq \ell \leq L.$$

Sendo $C = \frac{1}{2}\|a^{[L]} - y\|^2$, $a^{[L]}$ composição de funções, aplicamos a regra da cadeia.

As quantidades

$$\delta_j^{[l]} = \frac{\partial C}{\partial z_j^{[l]}} \quad , \quad \forall j \; , \; 2 \leq l \leq L ,$$

serão úteis. Note que

$$z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]} \quad , \quad 2 \leq l \leq L ,$$

foram calculados no feed foward, assim como

$a^{[l]}$, $1 \leq l \leq L$.

- $a^{[L]} = a(z^{[L]}) \Rightarrow \dfrac{\partial a_j^{[L]}}{\partial z_j^{[L]}} = a'(z_j^{[L]})$

- $\dfrac{\partial C}{\partial a_j^{[L]}} = \dfrac{\partial}{\partial a_j^{[L]}} \tfrac{1}{2} \| a^{[L]} - y \|^2 = a_j^{[L]} - y_j$

- $\delta_j^{[L]} = \dfrac{\partial C}{\partial z_j^{[L]}} = \dfrac{\partial C}{\partial a_j^{[L]}} \cdot \dfrac{a_j^{[L]}}{\partial z_j^{[L]}} = (a_j^{[L]} - y_j)\, a'(z_j^{[L]})$

Agora tratamos os $\delta^{[l]}$ intermediários $(l < L)$.
A ideia é escrever $\delta^{[l]}$ em função de $\delta^{[l+1]}$

(lembre-se que no back propagation vamos da camada de saída p/ a de entrada). Para relacionar $\delta_j^{[l]} = \dfrac{\partial C}{\partial z_j^{[l]}}$ à $\delta_k^{[l+1]} = \dfrac{\partial C}{\partial z_k^{[l+1]}}$, utilizamos novamente a regra da cadeia:

$$\delta_j^{[l]} = \frac{\partial C}{\partial z_j^{[l]}} = \sum_k \frac{\partial C}{\partial z_k^{[l+1]}} \cdot \frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = \sum_k \delta_k^{[l+1]} \frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}}.$$

Como $z_k^{[l+1]} = W_k^{[l+1]} a^{[l]} + b_k^{[l+1]} = \sum_p w_{kp}^{[l+1]} a(z_p^{[l]}) + b_k^{[l+1]}$, temos $\dfrac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = w_{kj}^{[l+1]} \, a'(z_j^{[l]})$. Assim

$$\delta_j^{[l]} = \sum_k \delta_k^{[l+1]} \, w_{kj}^{[l+1]} \, a'(z_j^{[l]})$$

$$\Rightarrow \quad \delta_j^{[l]} = a'(z_j^{[l]}) \left(\left(W^{[l+1]} \right)^T \delta^{[l+1]} \right)_j$$

Finalmente calculamos $\dfrac{\partial C}{\partial w_{jk}^{[l]}}$ e $\dfrac{\partial C}{\partial b_j^{[l]}}$:

201

- $$z_j^{[l]} = \left(W^{[l]} a^{[l-1]} + b^{[l]}\right)_j = \left(W^{[l]} a\left(z^{[l-1]}\right) + b^{[l]}\right)_j$$

$$\Rightarrow \frac{\partial z_j^{[l]}}{\partial b_j^{[l]}} = 1$$

- $$\frac{\partial C}{\partial b_j^{[l]}} = \frac{\partial C}{\partial z_j^{[l]}} \cdot \frac{\partial z_j^{[l]}}{\partial b_j^{[l]}} = \delta_j^{[l]}$$

- $$z_j^{[l]} = \left(W^{[l]} a^{[l-1]} + b^{[l]}\right)_j = \sum_k w_{jk}^{[l]} a_k^{[l-1]} + b_j^{[l]}$$

$$\Rightarrow \frac{\partial z_j^{[l]}}{\partial w_{jk}^{[l]}} = a_k^{[l-1]}$$

. Analizando a expressão de $z_p^{[l]}$, $p \neq j$, vemos que $w_{jk}^{[l]}$ não aparece na soma $\sum_k w_{pk}^{[l]}$, e logo $\frac{\partial z_p^{[l]}}{\partial w_{jk}^{[l]}} = 0$.

Daí,

$$\odot \quad \frac{\partial C}{\partial w_{jk}^{[l]}} = \sum_p \frac{\partial C}{\partial z_p^{[l]}} \cdot \frac{\partial z_p^{[l]}}{\partial w_{jk}^{[l]}} = \frac{\partial C}{\partial z_j^{[l]}} \cdot \frac{\partial z_j^{[l]}}{\partial w_{jk}^{[l]}} = \delta_j^{[l]} a_k^{[l-1]}.$$

Resumindo (itens ⊙) o cálculo de C':

1) $\delta_j^{[L]} = (a_j^{[L]} - y_j)\, a'(z_j^{[L]})$, $\forall j$ camada saída.

2) $\delta_j^{[l]} = a'(z_j^{[l]}) \left((w^{[l+1]})^T \delta^{[l+1]} \right)_j$, $\forall j$, $2 \leq l \leq L-1$

3) $\dfrac{\partial C}{\partial b_j^{[l]}} = \delta_j^{[l]}$, $\forall j$, $2 \leq l \leq L$

4) $\dfrac{\partial C}{\partial w_{jk}^{[l]}} = \delta_j^{[l]} a_k^{[l-1]}$, $\forall j,k$, $2 \leq l \leq L$.

Observações:

(i) somente 1) depende da escolha de C.
Logo, é possível adaptar facilmente este back propagation para outras funções $R_m(w,b)$.

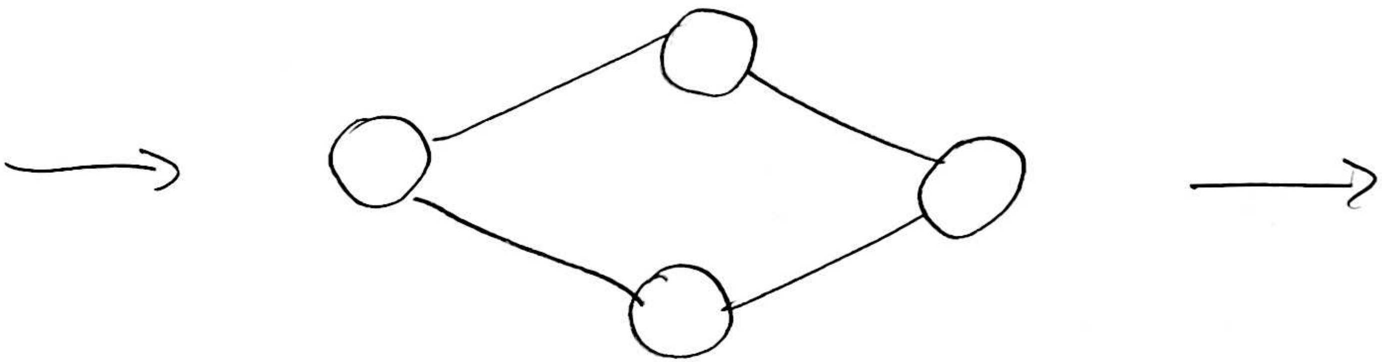
(ii) as contas independem da escolha da função de ativação $\underline{a}$ (basta ser diferenciável nos valores avaliados). A derivada a' deve ser fornecido. Por exemplo, $a(z) = 1/(1+e^{-z})$
$\Rightarrow a'(z) = a(z)(1-a(z))$.

<u>Exemplo</u>: Considere o exemplo anterior: $R_2(w,b)$,

$(x_1, y_1) = (1,1)$, $(x_2, y_2) = (2,0)$,

$W^{[2]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $W^{[3]} = \begin{bmatrix} 2 & 3 \end{bmatrix}$, $b^{[2]} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$, $b^{[3]} = [1]$,

sigmoid $a(z) = 1/(1+e^{-z})$ em todos os neurônios.



Vamos calcular $\nabla\left\{\frac{1}{2}\|a^{[L]} - y\|^2\right\}$ para o dado 1 no ponto (w, b) fornecido.

Ao aplicarmos feed foward, encontramos

- $a^{[1]} = [1]$
- $a^{[2]} = \begin{bmatrix} 0,5 \\ -0,1398 \end{bmatrix}$, $z^{[2]} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$
- $a^{[3]} = [-0,3033]$, $z^{[3]} = [1,5807]$

(reveja as contas do exemplo anterior).

1) $$\delta^{[3]} = (a_1^{[3]} - y_1)\, a'(z_1^{[3]}) \approx (-1{,}3033)\left[a^{[3]}(1-a^{[3]})\right]$$

$$\approx \left[0{,}5152\right]$$

2) $$(W^{[3]})^T \delta^{[3]} \approx \begin{bmatrix} 2 \\ 3 \end{bmatrix} \left[0{,}5152\right] = \begin{bmatrix} 1{,}0304 \\ 1{,}5456 \end{bmatrix}$$

$$\delta^{[2]} \approx \begin{bmatrix} a'(z_1^{[2]})\,(1{,}0304) \\ a'(z_2^{[2]})\,(1{,}5456) \end{bmatrix} = \begin{bmatrix} 1{,}0304 \cdot a_1^{[2]}(1-a_1^{[2]}) \\ 1{,}5456 \cdot a_2^{[2]}(1-a_2^{[2]}) \end{bmatrix}$$

$$\approx \begin{bmatrix} 0{,}2576 \\ -0{,}2463 \end{bmatrix}$$

3) $$\frac{\partial C}{\partial b_1^{[2]}} = \delta_1^{[2]} = 0{,}2576$$

$$\frac{\partial C}{\partial b_2^{[2]}} = \delta_2^{[2]} = -0{,}2463$$

$$\frac{\partial C}{\partial b_1^{[3]}} = \delta_1^{[3]} = 0{,}5152$$

$\nabla_b C(w,b)$.

4) Faça as contas e calcule $\nabla_w C(w,b)$.