

Treinamento de redes neurais

(1)

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 .

(2)

- h será composta de uma aplicação afim com uma função não linear a (função de ativação):

$$h(x; w, b) = a(wx + b).$$

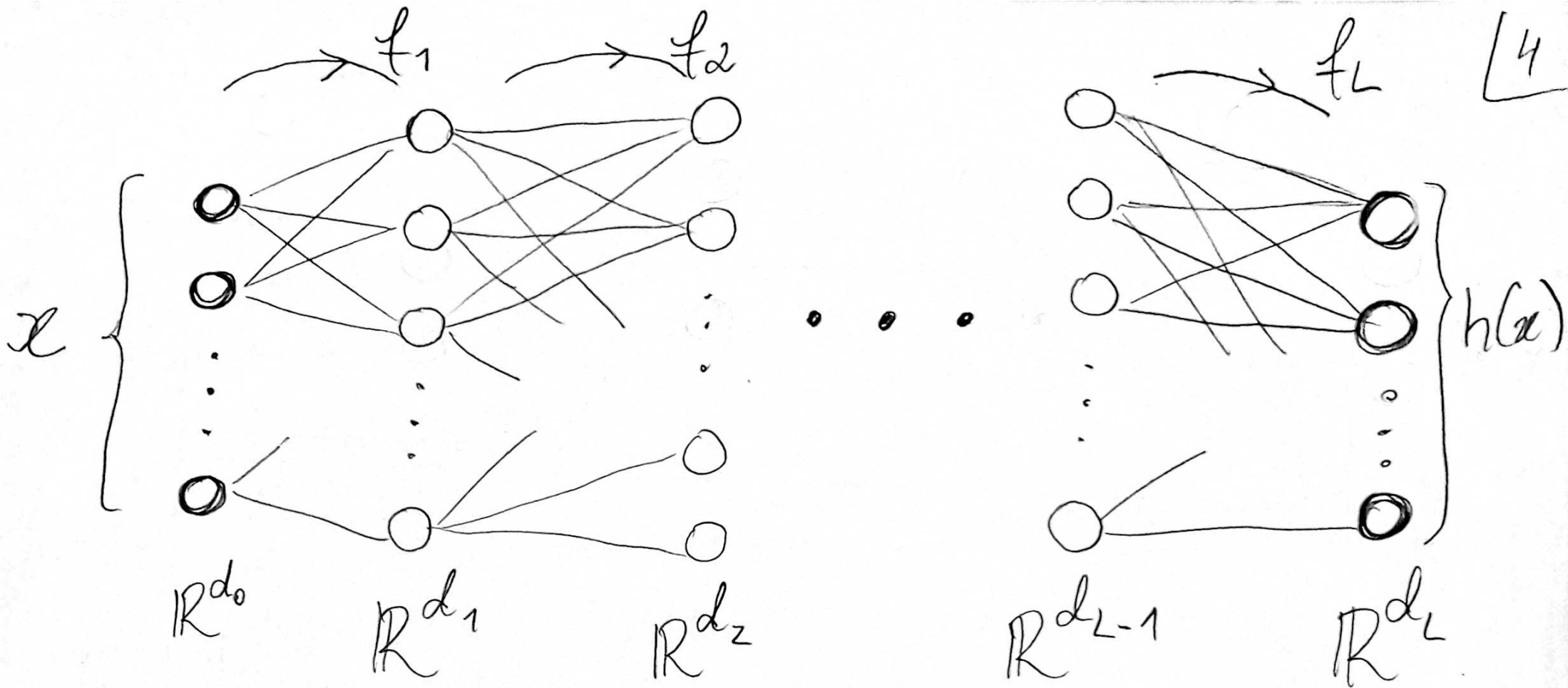
Um exemplo de função de ativação é a sigmoide: $a(z) = 1 / (1 + e^{-z})$.

Notação: $a(w) = \begin{bmatrix} a(w_1) \\ \vdots \\ a(w_m) \end{bmatrix}$, $w \in \mathbb{R}^m$.

EX: $\min_{w, b} R_m(w, b) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} \|a(wx_i + b) - y_i\|^2$ (3)

Queremos aplicar gradiente estocástico, logo precisamos de avaliar, dados w, b , os termos da soma e seus gradientes em relação a (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 forma $a(wx + b)$.



feed forward \rightarrow

\leftarrow back propagation

Feed forward (avaliar $R_m(w, b)$)

5

Para cada $i \in \{1, \dots, m\}$, escrevemos o termo $C = \frac{1}{2} \|a(wx_i + b) - y_i\|^2$ como

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

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

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

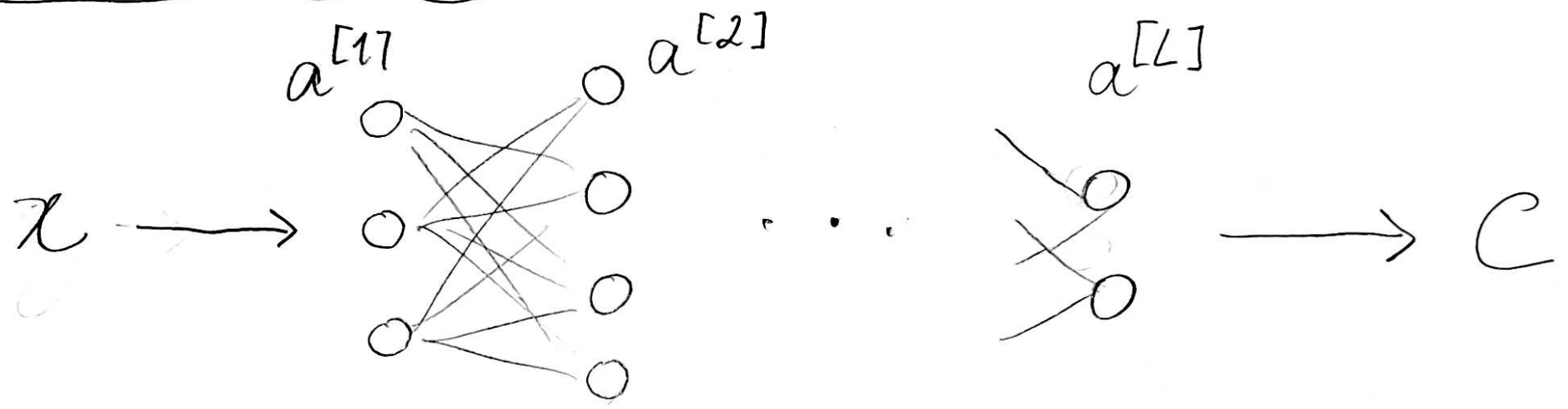
$$a^{[1]} = x$$

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

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

⋮

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



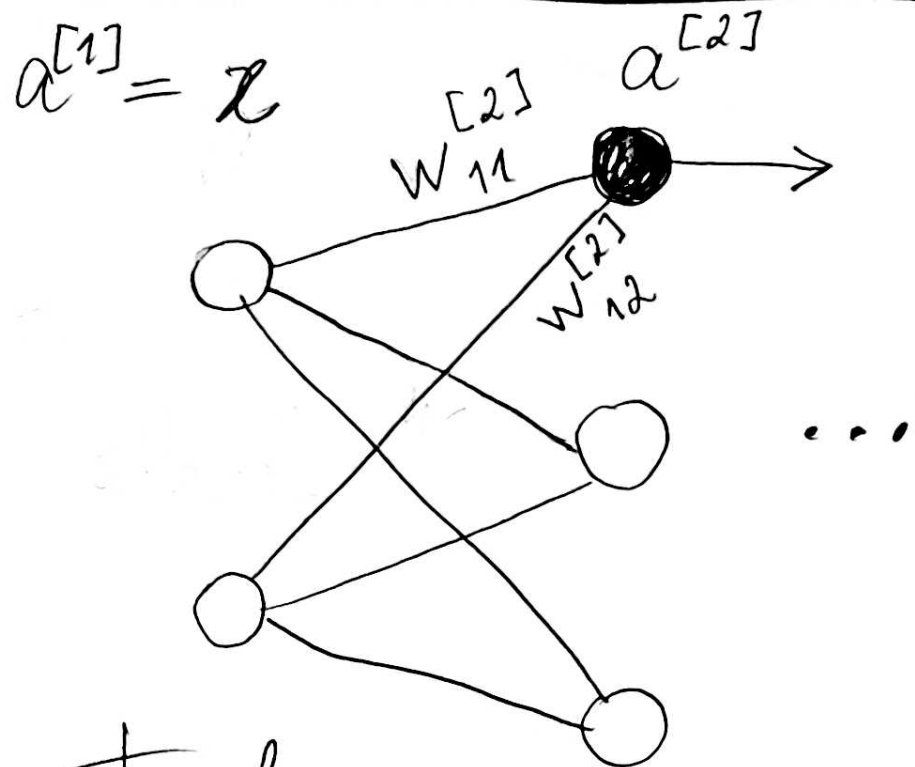
Em geral,

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

- $a^{[l]} = a(w^{[l]} a^{[l-1]} + b^{[l]})$, $l = 2, \dots, L$.

- $C = \frac{1}{2} \|a^{[L]} - y\|^2$.

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

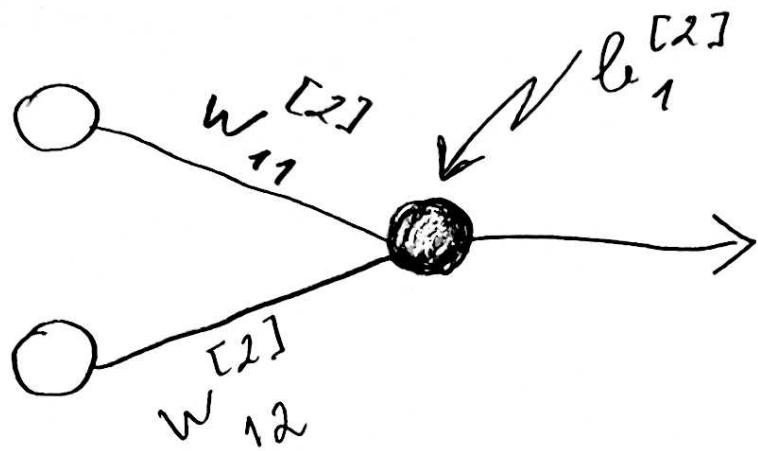


entrada
 $(x \in \mathbb{R}^2)$

camada 2
 $(a^{[2]} \in \mathbb{R}^3)$

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

neurônio \bullet recebe o fluxo x , pondera com o vetor $w_1^{[2]} \in \mathbb{R}^2$, adiciona o vies $b_1^{[2]}$ e passa o fluxo aplicando a ativação \underline{a} .



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

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

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,

$$a^{[l]} = a(W^{[l]} a^{[l-1]} + b^{[l]}), \text{ onde } \quad (10)$$

$W^{[l]}$ é matriz $m_l \times m_{l-1}$ e $b^{[l]} \in \mathbb{R}^{m_l}$,

m_l, m_{l-1} número de neurônios nas camadas l e $l-1$, respectivamente.

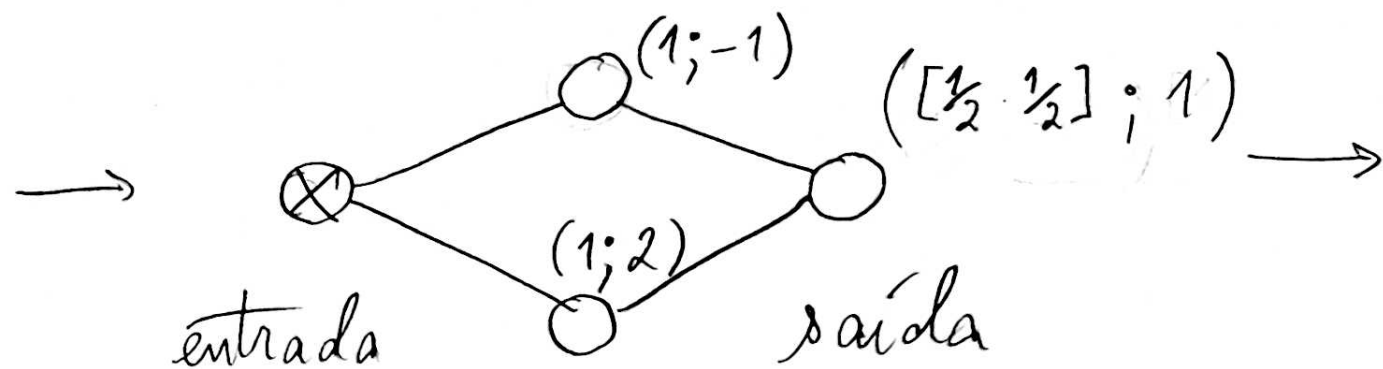
Exemplo: calcular $R_2(w, b)$ com 2 dados

$(x_1, y_1) = (1, 1)$ 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

(11)



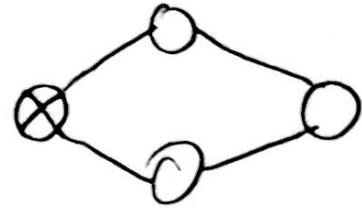
(obs: (w, b) representa as matrizes $W^{[l]}$, $b^{[l]}$, por exemplo, empilhando as colunas em sequência).

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

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

(12)

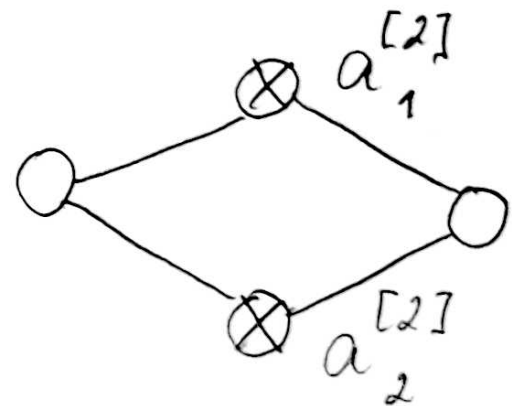
- $a^{[1]} = x_1 = [1]$



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

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

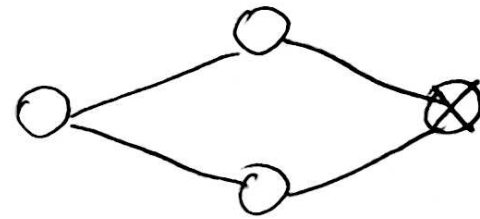
$$= \begin{bmatrix} \frac{1}{1+e^0} \\ \frac{1}{1+e^{-3}} \end{bmatrix} \approx \begin{bmatrix} 0,5 \\ -0,1398 \end{bmatrix}$$



$$\bullet a^{[3]} = a(w^{[3]} a^{[2]} + b^{[3]})$$

$$= a\left([2 \ 3] \begin{bmatrix} 0,5 \\ -0,1398 \end{bmatrix} + [1]\right) = a(1,5207)$$

$$\approx [-0,3033]$$



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

Para o dado $(x_2, y_2) = (2, 0)$: (14)

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 $\nabla R_m(w, b)$)

[15]

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

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

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

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

sendo $C = \frac{1}{2} \|a^{[L]} - y\|^2$, $a^{[L]}$ em posição de 16
funções, aplicamos a regra da cadeia.

As quantidades

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

serão úteis. Note que

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

foram calculados no feed forward, assim como

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

$$\bullet a^{[L]} = a(z_j^{[L]}) \Rightarrow \frac{\partial a_j^{[L]}}{\partial z_j^{[L]}} = a'(z_j^{[L]}) \quad (17)$$

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

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

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 118
 camada de saída p/ a de entrada). Para
 relacionar $\delta_j^{[l]} = \frac{\partial C}{\partial z_j^{[l]}}$ a $\delta_k^{[l+1]} = \frac{\partial C}{\partial z_k^{[l+1]}}$,

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 $z_k^{[l+1]} = w_k^{[l+1]} a_j^{[l]} + b_k^{[l+1]} = \sum_p w_{kp}^{[l+1]} a_p^{[l]} + b_k^{[l+1]}$,

temos

$$\frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = w_{kj}^{[l+1]} a'(z_j^{[l]}) \quad \text{Assim}$$

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

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

Finalmente calculamos

$$\frac{\partial C}{\partial w_{jk}^{[l]}} \quad \text{e} \quad \frac{\partial C}{\partial b_j^{[l]}}$$

201

$$\bullet z_j^{[l]} = (w^{[l]} a^{[l-1]} + b^{[l]})_j = (w^{[l]} a(z_j^{[l-1]}) + b^{[l]})_j$$

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

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

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

$$\Rightarrow \frac{\partial z_j^{[l]}}{\partial w_{jk}^{[l]}} = a_k^{[l-1]} \cdot \text{Qualizando a expres-} \quad [21]$$

são de $z_p^{[l]}$, $p \neq j$, menos que $w_{jk}^{[l]}$ não

aparece na soma $\sum_k w_{pk}^{[l]}$, e logo $\frac{\partial z_p^{[l]}}{\partial w_{jk}^{[l]}} = 0$.

Dai,

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

Resumindo (itens \odot) o cálculo de C' : (22)

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

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

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

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

Observações:

23

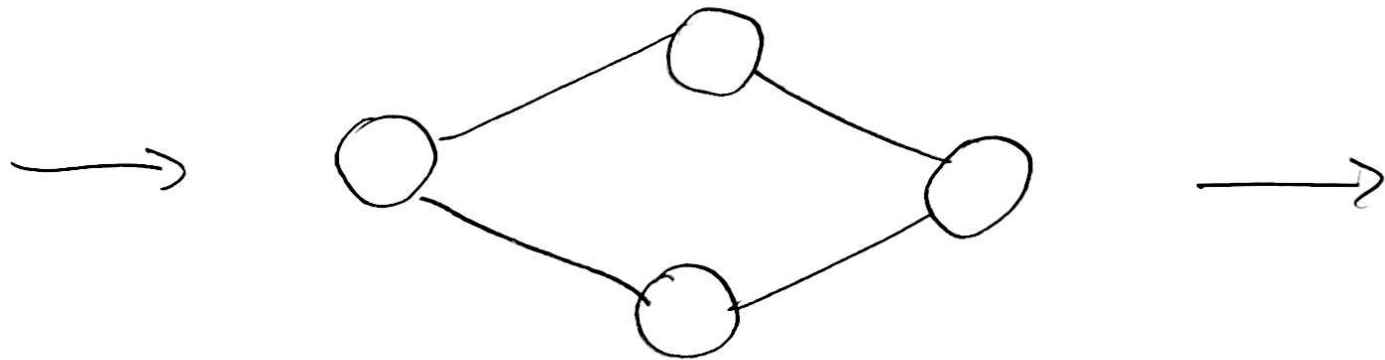
- (i) somente 1) depende da escolha de C .
Logo, é possível adaptar facilmente este
back propagation para outras funções $R_m(u, b)$.
-
- (ii) as contas independem da escolha da função
de ativação a (basta ser diferenciável nos
valores avaliados). A derivada a' deve ser
fornecido. Por exemplo, $a(z) = 1/(1 + e^{-z})$
 $\Rightarrow a'(z) = a(z)(1 - a(z))$.

Exemplo: considere o exemplo anterior: $R_2(w, b)$, (24)

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

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

sigmoide $a(z) = 1/(1 + e^{-z})$ 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.

25

Se aplicarmos feed forward, 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).

$$1) \delta^{[3]} = (a_1^{[3]} - y_1) a'(z_1^{[3]}) \approx (-1,3033) [a^{[3]}(1-a^{[3]})]^{26}$$

$$\approx [0,5152]$$

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

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

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

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

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

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

$$\nabla_b C(w, b).$$

27

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