Aprendizado Profundo 1

Cross Entropy, Softmax e Grafo Computacional

Professor: Lucas Silveira Kupssinskü

MLP para tarefa classificação multiclasse

- Função de ativação: Softmax
- Função de Custo: Entropia Cruzada

X

0	0
0	1
1	0
1	1

y

1	0	0
0	1	0
0	1	0
0	0	1

$\theta^{(1)}$

-0.69	0.63
0.36	0.59

$\beta^{(1)}$

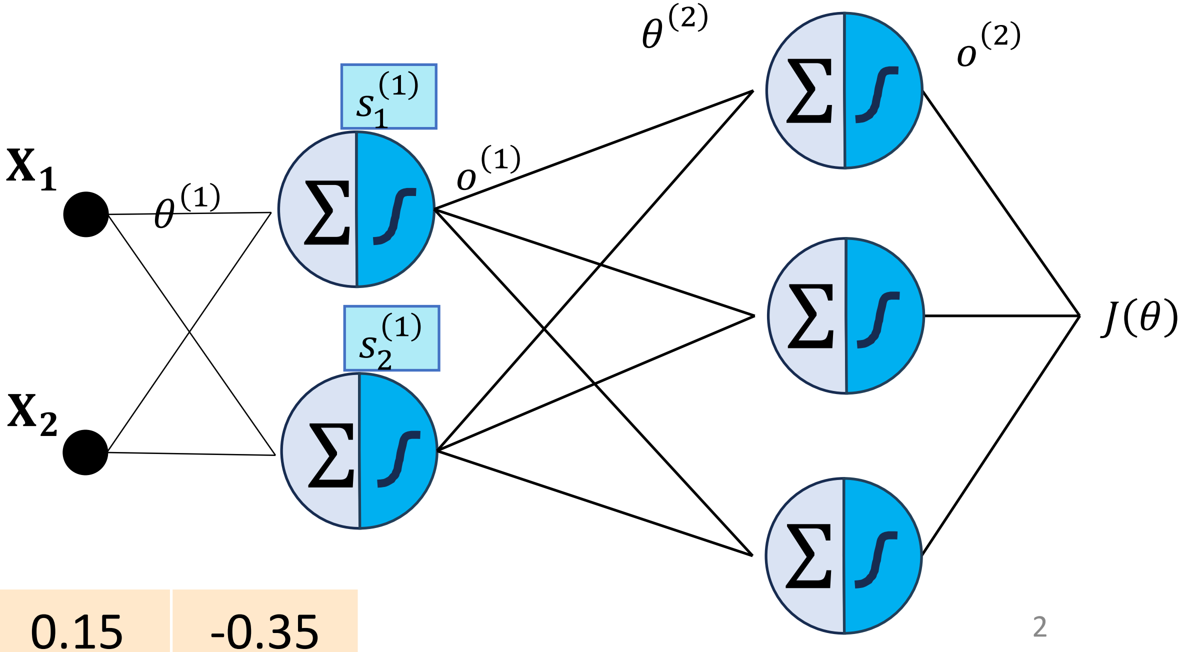
-0.41	-0.01

$\theta^{(2)}$

-0.31	0.02
0.48	-0.68
-0.41	-0.27

$\beta^{(2)}$

-0.22	0.15	-0.35



MLP para tarefa classificação multiclasse

- *Forward Pass* camada 1

$$X\left(\theta^{(1)}\right)^T + \beta^{(1)} =$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -0.69 & 0.63 \\ 0.36 & 0.59 \end{bmatrix}^T + \begin{bmatrix} -0.41 & -0.01 \end{bmatrix} = \begin{bmatrix} -0.41 & -0.01 \\ 0.22 & 0.58 \\ -1.10 & 0.35 \\ -0.47 & 0.94 \end{bmatrix}$$

```
z1 = np.matmul(x, theta_1.T) + bias_1
```

MLP para tarefa classificação multiclasse

- *Forward Pass* camada 1

$$sigmoid\left(s^{(1)}\right) =$$

$$sigmoid\begin{bmatrix} -0.41 & -0.01 \\ 0.22 & 0.58 \\ -1.10 & 0.35 \\ -0.47 & 0.94 \end{bmatrix} =$$

0.3989	0.4975
0.5548	0.6411
0.2497	0.5866
0.3846	0.7191

```
o1 = 1/(1+np.exp(-s1))
```

MLP para tarefa classificação multiclasse

- *Forward Pass* camada 2

$$o^{(1)}\left(\theta^{(2)}\right)^T + \beta^{(2)} =$$

$$\begin{bmatrix} 0.3989 & 0.4975 \\ 0.5548 & 0.6411 \\ 0.2497 & 0.5866 \\ 0.3846 & 0.7191 \end{bmatrix} \begin{bmatrix} -0.31 & 0.02 \\ 0.48 & -0.68 \\ -0.41 & -0.27 \end{bmatrix}^T + \begin{bmatrix} -0.22 & 0.15 & -0.35 \end{bmatrix} = \begin{bmatrix} -0.3337 & 0.0032 & -0.6479 \\ -0.3792 & -0.0196 & -0.7505 \\ -0.2857 & -0.1290 & -0.6108 \\ -0.3248 & -0.1544 & -0.7018 \end{bmatrix}$$

```
s2 = np.matmul(o1, theta_2.T) + bias_2
```

MLP para tarefa classificação multiclasse

- *Forward Pass* camada 2

$$o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$softmax \begin{bmatrix} -0.3337 & 0.0032 & -0.6479 \\ -0.3792 & -0.0196 & -0.7505 \\ -0.2857 & -0.1290 & -0.6108 \\ -0.3248 & -0.1544 & -0.7018 \end{bmatrix} = \begin{matrix} 0.3194 & 0.4473 & 0.2333 \\ 0.3203 & 0.4588 & 0.2209 \\ 0.3458 & 0.4044 & 0.2498 \\ 0.3482 & 0.4129 & 0.2388 \end{matrix}$$

```
o2 = np.exp(s2)/np.sum(np.exp(s2), axis=1, keepdims=True)
```

MLP para tarefa classificação multiclasse

- *Loss*

$$L(y,\hat{y}) = -\sum_{i=1}^{C} y_i ln\hat{y}$$

$$-\sum_{i=1}^{C}$$

0.3194	0.4473	0.2333
0.3203	0.4588	0.2209
0.3458	0.4044	0.2498
0.3482	0.4129	0.2388

1	0	0
0	1	0
0	1	0
0	0	1

$$= -\sum_{i=1}^{C}$$

-1.1413	0.	0.
0.	-0.7791	0.0
0.	-0.905	0.
0.	0.	-1.43

```
y*np.log(a2)
```

MLP para tarefa classificação multiclasse

- *Loss*

$$L(y,\hat{y}) = -\sum_{i=1}^{C} y_i ln\hat{y}$$

$$-\sum_{i=1}^{C} \quad -\sum_{i=1}^{C}$$

-1.1413	0.	0.
0.	-0.7791	0.0
0.	-0.905	0.
0.	0.	-1.43

$$= 1.06441$$

```
-np.sum(y*np.log(a2))/a2.shape[0]
```

MLP para tarefa classificação multiclasse

- *Backward Pass*

$$\delta^{(2)} = \frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\boxed{\partial o^{(L)}}} \frac{\partial\left(o^{(L)}\right)}{\boxed{\partial s^{(L)}}}$$

São vetores!

MLP para tarefa classificação multiclasse

- *Backward Pass*

$$\delta^{(2)} = \frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\boxed{\partial o^{(L)}}} \frac{\partial\left(o^{(L)}\right)}{\boxed{\partial s^{(L)}}}$$

São vetores!

$$\frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\partial o_i^{(L)}} = \frac{\partial\left(-\sum_{j=1}^{C} y_j ln o_j^{(L)}\right)}{\partial o_i^{(L)}} = \frac{-y_i}{o_i^{(L)}}$$

MLP para tarefa classificação multiclasse

- *Backward Pass*

$$\delta^{(2)} = \frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\boxed{\partial o^{(L)}}} \frac{\partial\left(o^{(L)}\right)}{\boxed{\partial s^{(L)}}}$$

São vetores!

$$\frac{\partial\left(o_i^{(L)}\right)}{\partial s_j^{(L)}} = \begin{cases} Se\ i = j, o_i^{(L)}(1 - o_i^{(L)}) \\ Se\ i \neq j, \qquad -o_i^{(L)} o_j^{(L)} \end{cases}$$

MLP para tarefa classificação multiclasse

- *Backward Pass*

$$\delta^{(2)} = \frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\boxed{\partial o^{(L)}}} \frac{\partial\left(o^{(L)}\right)}{\partial s^{(L)}}$$

$$\frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\partial o_i^{(L)}} = \frac{-y_i}{o_i^{(L)}}$$

-3.1310	0.	0.
0.	-2.1795	0.0
0.	-2.4727	0.
0.	0.	-4.1868

MLP para tarefa classificação multiclasse

- *Backward Pass*

$$\delta^{(2)} = \frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\partial o^{(L)}} \frac{\partial\left(o^{(L)}\right)}{\boxed{\partial s^{(L)}}}$$

$$\frac{\partial\left(o_i^{(L)}\right)}{\partial s_j^{(L)}} = \begin{cases} Se\ i = j, o_i^{(L)}\left(1 - o_i^{(L)}\right) \\ Se\ i \neq j, \quad -o_i^{(L)} o_j^{(L)} \end{cases}$$

```
a = np.eye(a2.shape[-1])
temp1 = np.zeros((a2.shape[0], a2.shape[1],
a2.shape[1]),dtype=np.float32)
temp2 = np.zeros((a2.shape[0], a2.shape[1],
a2.shape[1]),dtype=np.float32)
temp1 = np.einsum('ij,jk->ijk',a2,a)
temp2 = np.einsum('ij,ik->ijk',a2,a2)
temp1-temp2
```

0.2174	-0.1429	-0.0745
-0.1429	0.2472	-0.1044
-0.0745	-0.1044	0.1789

0.2177	-0.1469	-0.0707
-0.1469	0.2483	-0.1014
-0.0707	-0.1014	0.1721

0.2262	-0.1398	-0.0864
-0.1398	0.2409	-0.1010
-0.0863	-0.1010	0.1874

0.2269	-0.1438	-0.0831
-0.1438	0.2424	-0.0986
-0.0832	-0.0986	0.1818

$$\delta^{(2)} = \frac{\partial\left(L\left(y, o^{(L)}\right)\right)}{\partial o^{(L)}} \frac{\partial\left(o^{(L)}\right)}{\partial s^{(L)}}$$

-3.1310	0.	0.
0.	-2.1795	0.0
0.	-2.4727	0.
0.	0.	-4.1868

0.2174	-0.1429	-0.0745
-0.1429	0.2472	-0.1044
-0.0745	-0.1044	0.1789

0.2177	-0.1469	-0.0707
-0.1469	0.2483	-0.1014
-0.0707	-0.1014	0.1721

0.2262	-0.1398	-0.0864
-0.1398	0.2409	-0.1010
-0.0863	-0.1010	0.1874

0.2269	-0.1438	-0.0831
-0.1438	0.2424	-0.0986
-0.0832	-0.0986	0.1818

-0.6806	0.4473	0.2332
0.3203	-0.5412	0.2209
0.3458	-0.5956	0.2498
0.3482	0.4129	-0.7612

$$\frac{\partial\left(L\left(y, o^{(2)}\right)\right)}{\partial \theta^{(2)}} = o^{(1)} \delta^{(2)}$$

0.3989	0.4975
0.5548	0.6411
0.2497	0.5866
0.3846	0.7191

-0.6806	0.4473	0.2332
0.3203	-0.5412	0.2209
0.3458	-0.5956	0.2498
0.3482	0.4129	-0.7612

-0.6806
0.4473
0.2332

0.3989	0.4975

-0.2715	-0.3386
0.1784	0.2225
0.0930	0.1161

$$\frac{\partial\left(L\left(y, o^{(2)}\right)\right)}{\partial \theta^{(2)}} = o^{(1)}\delta^{(2)}$$

0.3989	0.4975	-0.6806	0.4473	0.2332
0.5548	0.6411	0.3203	-0.5412	0.2209
0.2497	0.5866	0.3458	-0.5956	0.2498
0.3846	0.7191	0.3482	0.4129	-0.7612

-0.6806
0.4473
0.2332

0.3989	0.4975

-0.2715	-0.3386
0.1784	0.2225
0.0930	0.1161

-0.3203
-0.5412
0.2209

0.5548	0.6411

0.1777	0.2053
-0.3002	-0.3469
0.1226	0.1416

$$\frac{\partial\left(L\left(y,o^{(2)}\right)\right)}{\partial\theta^{(2)}}=o^{(1)}\delta^{(2)}$$

0.3989	0.4975
0.5548	0.6411
0.2497	0.5866
0.3846	0.7191

-0.6806	0.4473	0.2332
0.3203	-0.5412	0.2209
0.3458	-0.5956	0.2498
0.3482	0.4129	-0.7612

-0.6806
0.4473
0.2332

0.3989	0.4975

-0.2715	-0.3386
0.1784	0.2225
0.0930	0.1161

0.3458
-0.5956
0.2498

0.2497	0.5866

0.0863	0.2028
-0.1487	-0.3493
0.0624	0.1465

-0.3203
-0.5412
0.2209

0.5548	0.6411

0.1777	0.2053
-0.3002	-0.3469
0.1226	0.1416

$$\frac{\partial\left(L\left(y, o^{(2)}\right)\right)}{\partial\theta^{(2)}} = o^{(1)}\delta^{(2)}$$

0.3989	0.4975
0.5548	0.6411
0.2497	0.5866
0.3846	0.7191

-0.6806	0.4473	0.2332
0.3203	-0.5412	0.2209
0.3458	-0.5956	0.2498
0.3482	0.4129	-0.7612

-0.6806
0.4473
0.2332

0.3989	0.4975

-0.2715	-0.3386
0.1784	0.2225
0.0930	0.1161

0.3458
-0.5956
0.2498

0.2497	0.5866

0.0863	0.2028
-0.1487	-0.3493
0.0624	0.1465

-0.3203
-0.5412
0.2209

0.5548	0.6411

0.1777	0.2053
-0.3002	-0.3469
0.1226	0.1416

0.3482
0.4129
-0.7612

0.3846	0.7191

0.1339	0.2504
0.1588	0.2969
-0.2928	-0.5473

$$\frac{\partial\left(L\left(y, o^{(2)}\right)\right)}{\partial\theta^{(2)}} = o^{(1)}\delta^{(2)}$$

0.3989	0.4975
0.5548	0.6411
0.2497	0.5866
0.3846	0.7191

-0.6806	0.4473	0.2332
0.3203	-0.5412	0.2209
0.3458	-0.5956	0.2498
0.3482	0.4129	-0.7612

O gradiente é acumulado pela média (mas poderia ser soma...)

$$\frac{1}{4}\left[\begin{matrix}-0.2715 & -0.3386\\ 0.1784 & 0.2225\\ 0.0930 & 0.1161\end{matrix} + \begin{matrix}0.1777 & 0.2053\\ -0.3002 & -0.3469\\ 0.1226 & 0.1416\end{matrix} + \begin{matrix}0.0863 & 0.2028\\ -0.1487 & -0.3493\\ 0.0624 & 0.1465\end{matrix} + \begin{matrix}0.1339 & 0.2504\\ 0.1588 & 0.2969\\ -0.2928 & -0.5473\end{matrix}\right] = \begin{matrix}0.0316 & 0.0800\\ -0.0279 & -0.0442\\ -0.0037 & -0.0358\end{matrix}$$

MLP para classificação multiclasse

- Essa forma de trabalhar é propensa a erros e gera algumas multiplicações de matrizes que podem ter um tamanho grande
  - Podemos fazer melhor
- Considere a Loss abaixo que recebe como entrada os logits

$$L(y, s) = -\sum_{i=1}^{C} y_i ln \frac{e^{s_i}}{\sum_j e^{s_j}}$$

MLP para classificação multiclasse

$$L(y,s) = -\sum_{i=1}^{C} y_i ln \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$\frac{\partial L(y,s)}{\partial s_k} = \frac{\partial \left(-\sum_{i=1}^{C} y_i ln \frac{e^{s_i}}{\sum_j e^{s_j}} \right)}{\partial s_k}$$

$$= -\sum_{i=1}^{C} y_i \frac{\partial \left(ln \frac{e^{s_i}}{\sum_j e^{s_j}} \right)}{\partial s_k}$$

MLP para classificação multiclasse

$$= -\sum_{i=1}^{C} y_i \frac{\partial\left(ln\frac{e^{s_i}}{\sum_j e^{s_j}}\right)}{\partial s_k}$$

$$= -\sum_{i=1}^{C} y_i \frac{\partial(lno_i)}{\partial o_i}\frac{\partial(o_i)}{\partial s_k}, o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$= -\sum_{i=1}^{C} \frac{y_i}{o_i}\frac{\partial(o_i)}{\partial s_k}, o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

MLP para classificação multiclasse

$$= -\sum_{i=1}^{C} \frac{y_i}{o_i} \frac{\partial(o_i)}{\partial s_k}, o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$= -\left[\sum_{i \neq k} \frac{y_i}{o_i}(-o_i o_k) + \frac{y_k}{o_k}\big(o_k(1 - o_k)\big)\right], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$= -\left[\sum_{i \neq k}(-y_i o_k) + y_k(1 - o_k),\right], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

MLP para classificação multiclasse

$$= -\left[\sum_{i \neq k}(-y_i o_k) + y_k(1 - o_k)\right], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$= -\left[-o_k \sum_{i \neq k} y_i + y_k(1 - o_k)\right], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$= -[-o_k(1 - y_k) + y_k(1 - o_k)], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

MLP para classificação multiclasse

$$= -[-o_k(1 - y_k) + y_k(1 - o_k)], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$= -[-o_k + o_k y_k + y_k - o_k y_k], o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

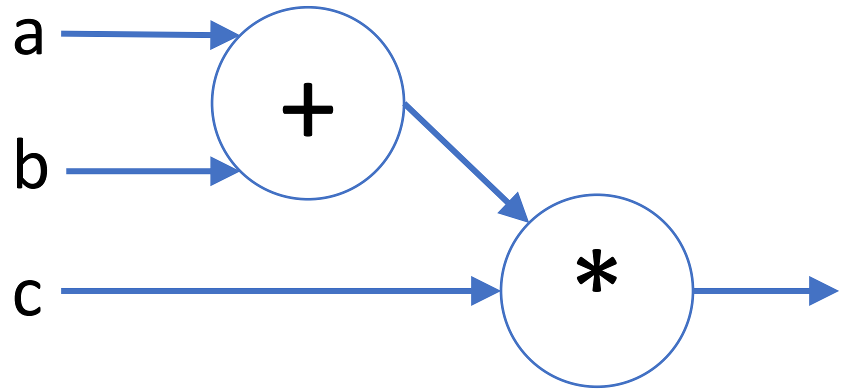
$$\boxed{\frac{\partial L(y,s)}{\partial s_k} = o_k - y_k, \qquad o_i = \frac{e^{s_i}}{\sum_j e^{s_j}}}$$

Grafo Computacional

- Considere a seguinte função

$$f(a, b, c) = (a + b) * c$$

- Podemos representar essa expressão com o seguinte grafo computacional direcionado acíclico



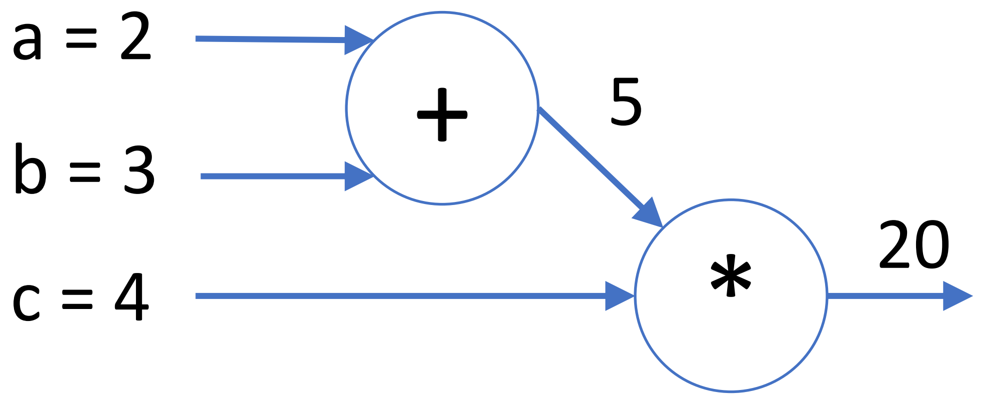
Essa parte é baseada no material de Fei-Fei Li & Andrej Karpathy & Justin Johnson CS231

Grafo Computacional

- Considere a seguinte função

$$f(a, b, c) = (a + b) * c$$

- Podemos representar essa expressão com o seguinte grafo computacional direcionado acíclico



Grafo Computacional

- Essa função pode ser derivada em relação a cada uma das variáveis

Derivando as operações nos nodos:

$$\frac{\partial(a+b)}{\partial a} = 1$$

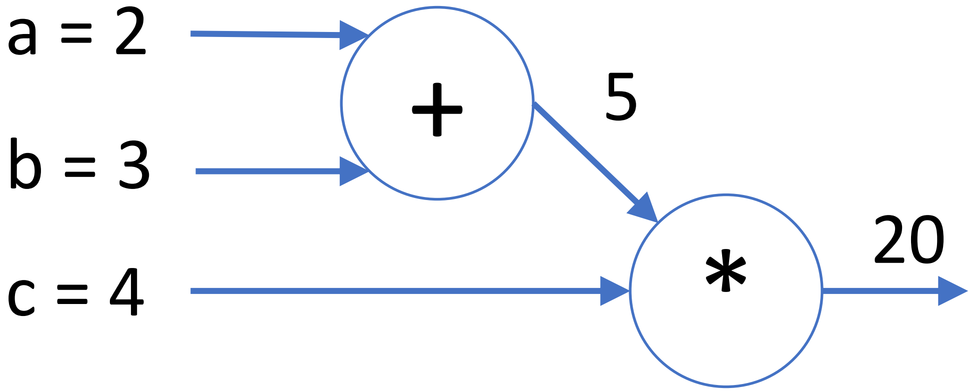
$$\frac{\partial(a+b)}{\partial b} = 1$$

$$\frac{\partial(d*c)}{\partial d} = c$$

$$\frac{\partial(d*c)}{\partial c} = d$$

$$f(a,b,c) = (a+b)*c$$

$$\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}$$



Grafo Computacional

$$f(a,b,c) = (a+b) * c$$

$$\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}$$

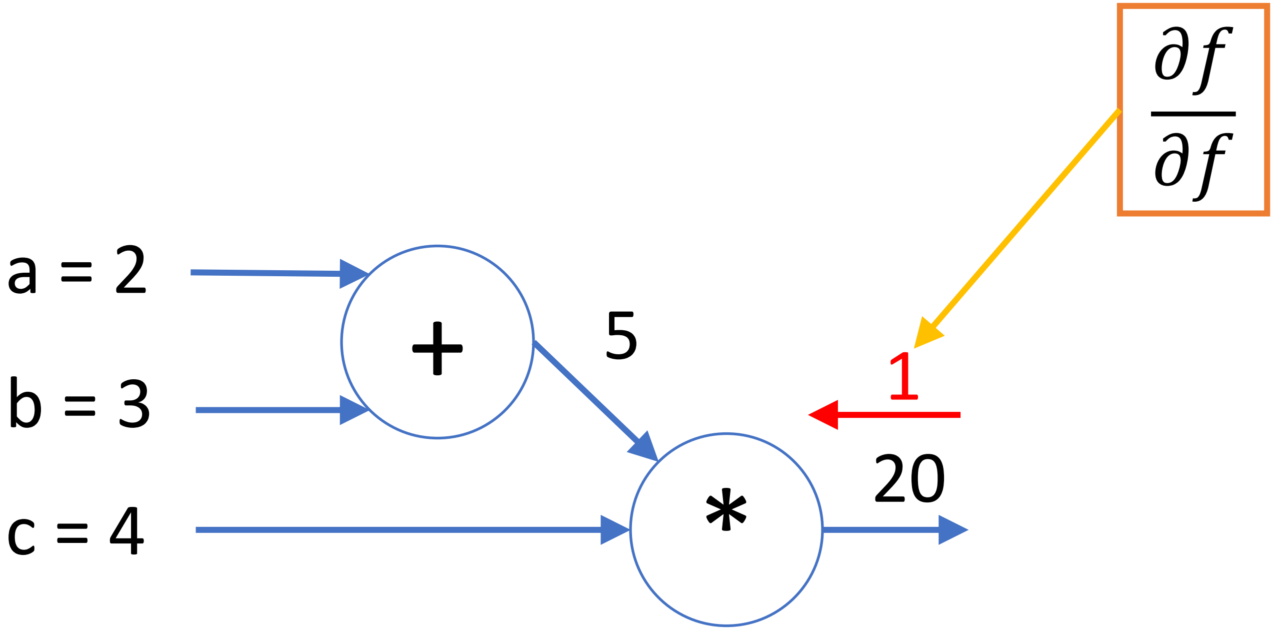
Derivando as operações nos nodos:

$$\frac{\partial(a+b)}{\partial a} = 1$$

$$\frac{\partial(a+b)}{\partial b} = 1$$

$$\frac{\partial(d*c)}{\partial d} = c$$

$$\frac{\partial(d*c)}{\partial c} = d$$



Grafo Computacional

$$f(a, b, c) = (a + b) * c$$

$$\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}$$

Derivando as operações nos nodos:

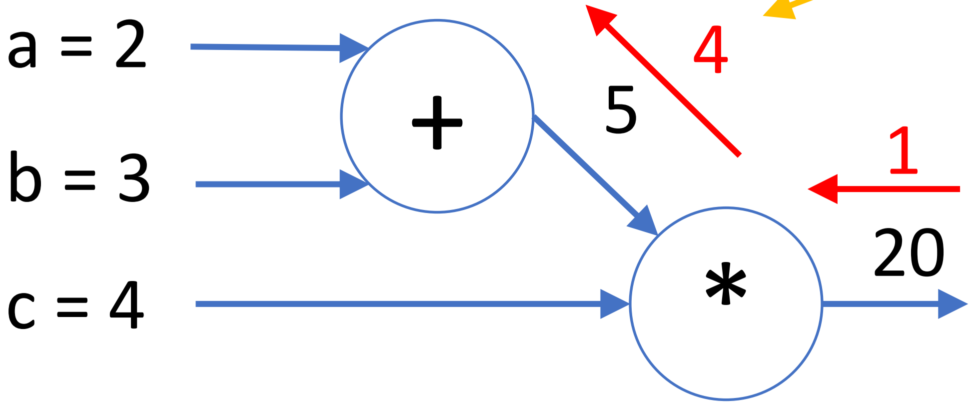
$$\frac{\partial (a+b)}{\partial a} = 1$$

$$\frac{\partial (a+b)}{\partial b} = 1$$

$$\frac{\partial (d*c)}{\partial d} = c$$

$$\frac{\partial (d*c)}{\partial c} = d$$

$$\boxed{\frac{\partial f}{\partial d}}$$



Grafo Computacional

$$f(a, b, c) = (a + b) * c$$

$$\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}$$

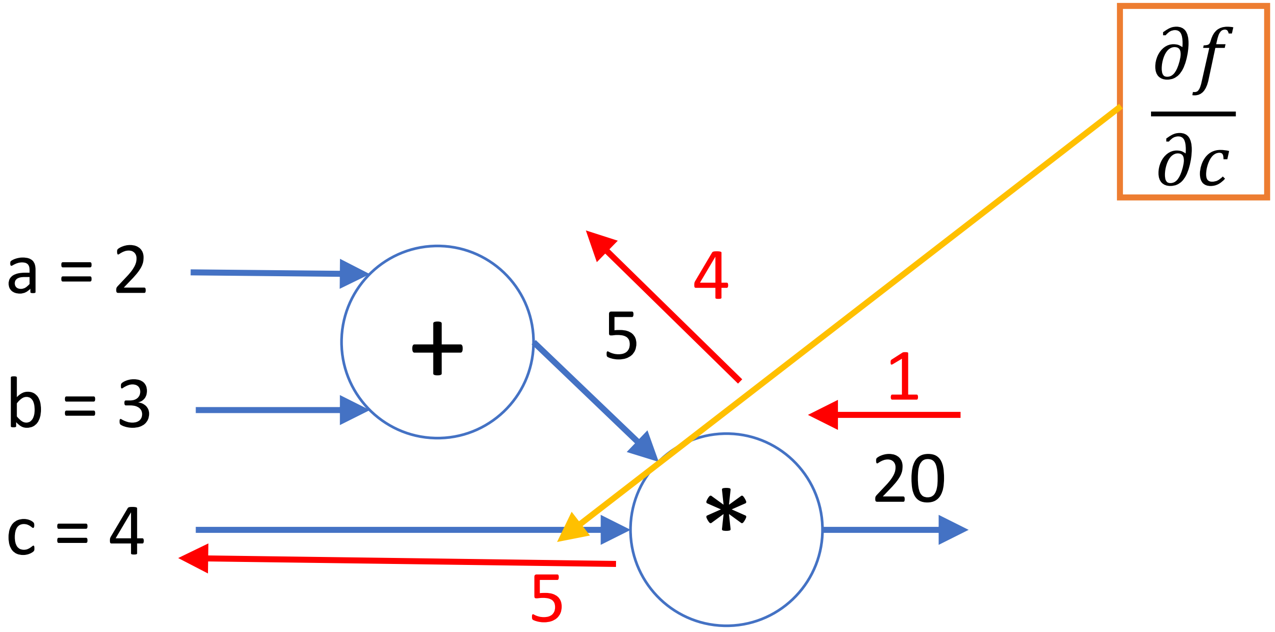
Derivando as operações nos nodos:

$$\frac{\partial(a+b)}{\partial a} = 1$$

$$\frac{\partial(a+b)}{\partial b} = 1$$

$$\frac{\partial(d*c)}{\partial d} = c$$

$$\frac{\partial(d*c)}{\partial c} = d$$



Grafo Computacional

$$f(a,b,c) = (a+b) * c$$

$$\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}$$

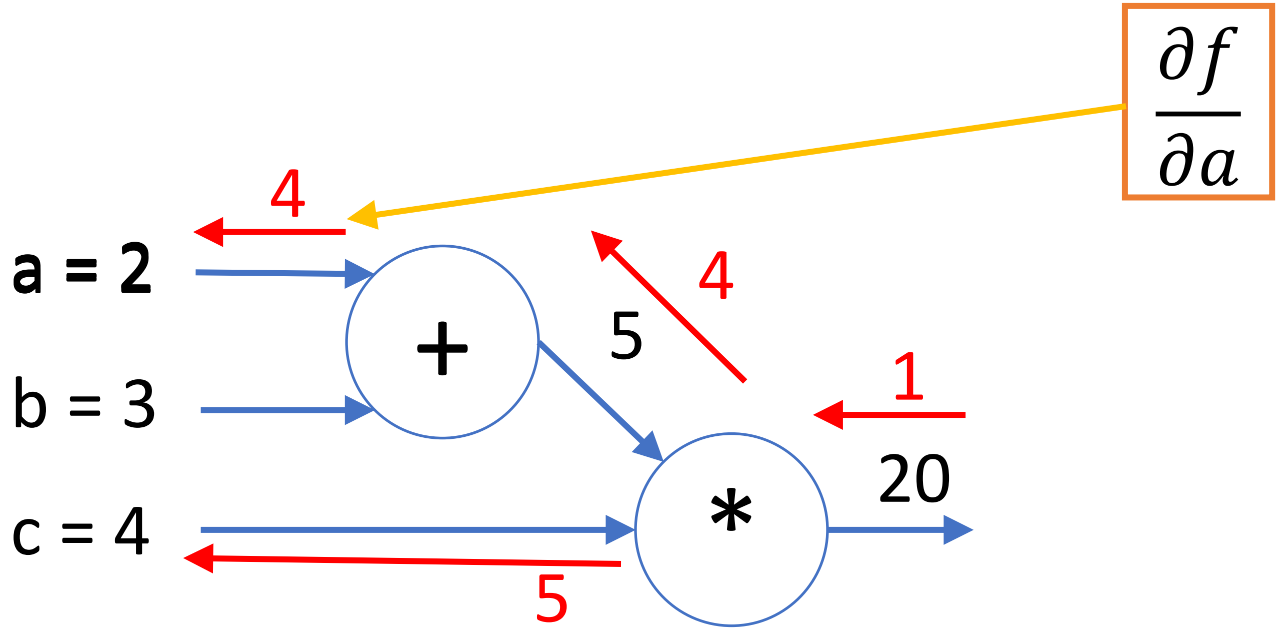
Derivando as operações nos nodos:

$$\frac{\partial(a+b)}{\partial a} = 1$$

$$\frac{\partial(a+b)}{\partial b} = 1$$

$$\frac{\partial(d*c)}{\partial d} = c$$

$$\frac{\partial(d*c)}{\partial c} = d$$



Grafo Computacional

$$f(a,b,c) = (a+b)*c$$

$$\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}$$

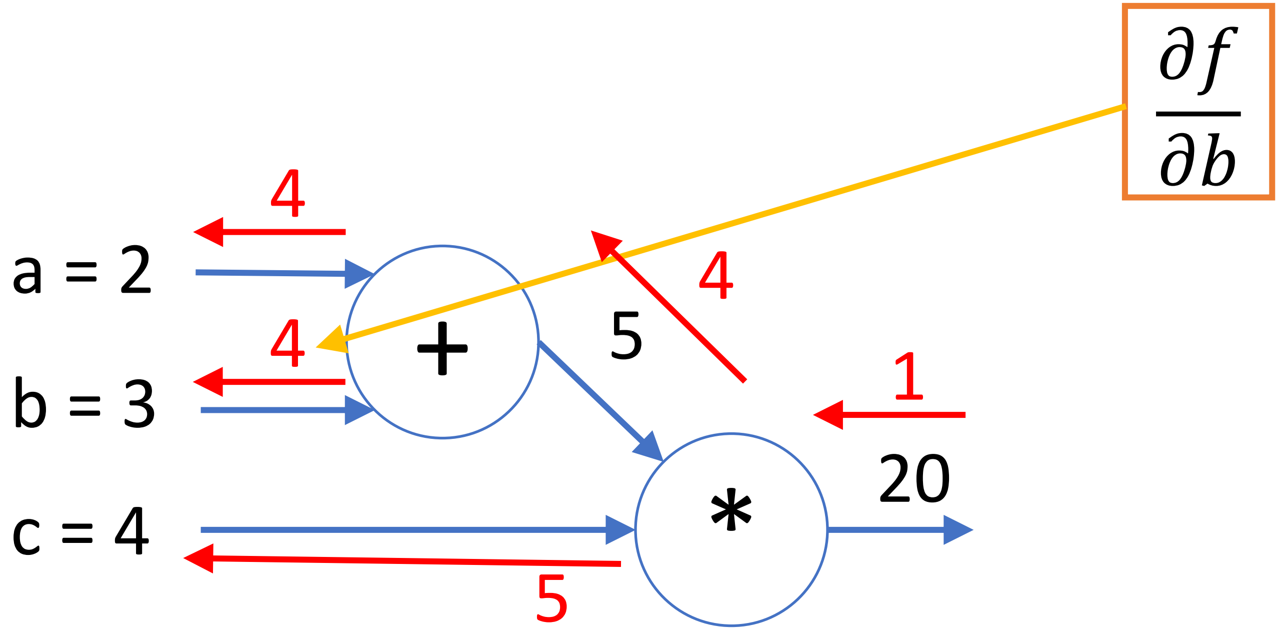
Derivando as operações nos nodos:

$$\frac{\partial(a+b)}{\partial a} = 1$$

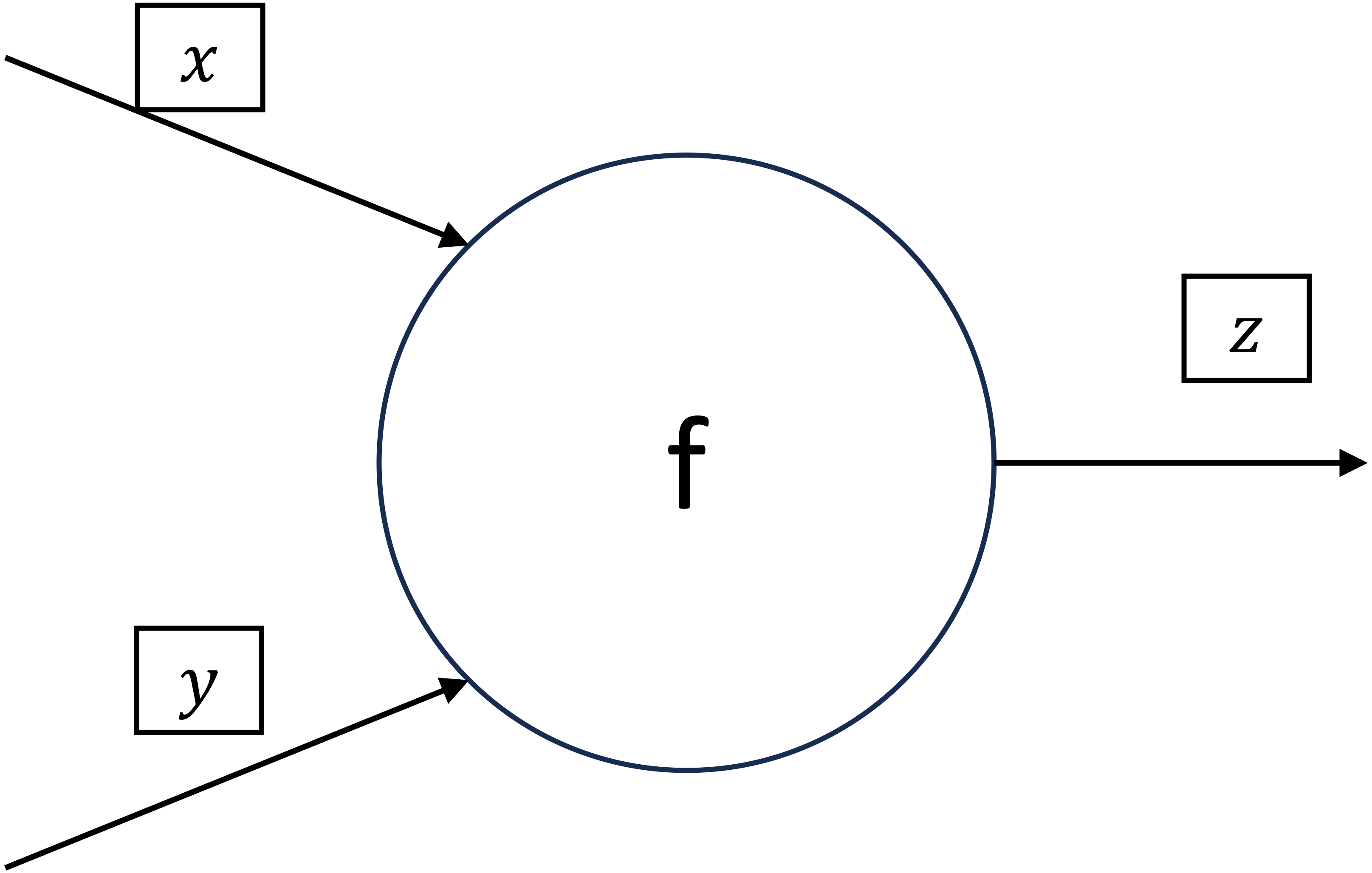
$$\frac{\partial(a+b)}{\partial b} = 1$$

$$\frac{\partial(d*c)}{\partial d} = c$$

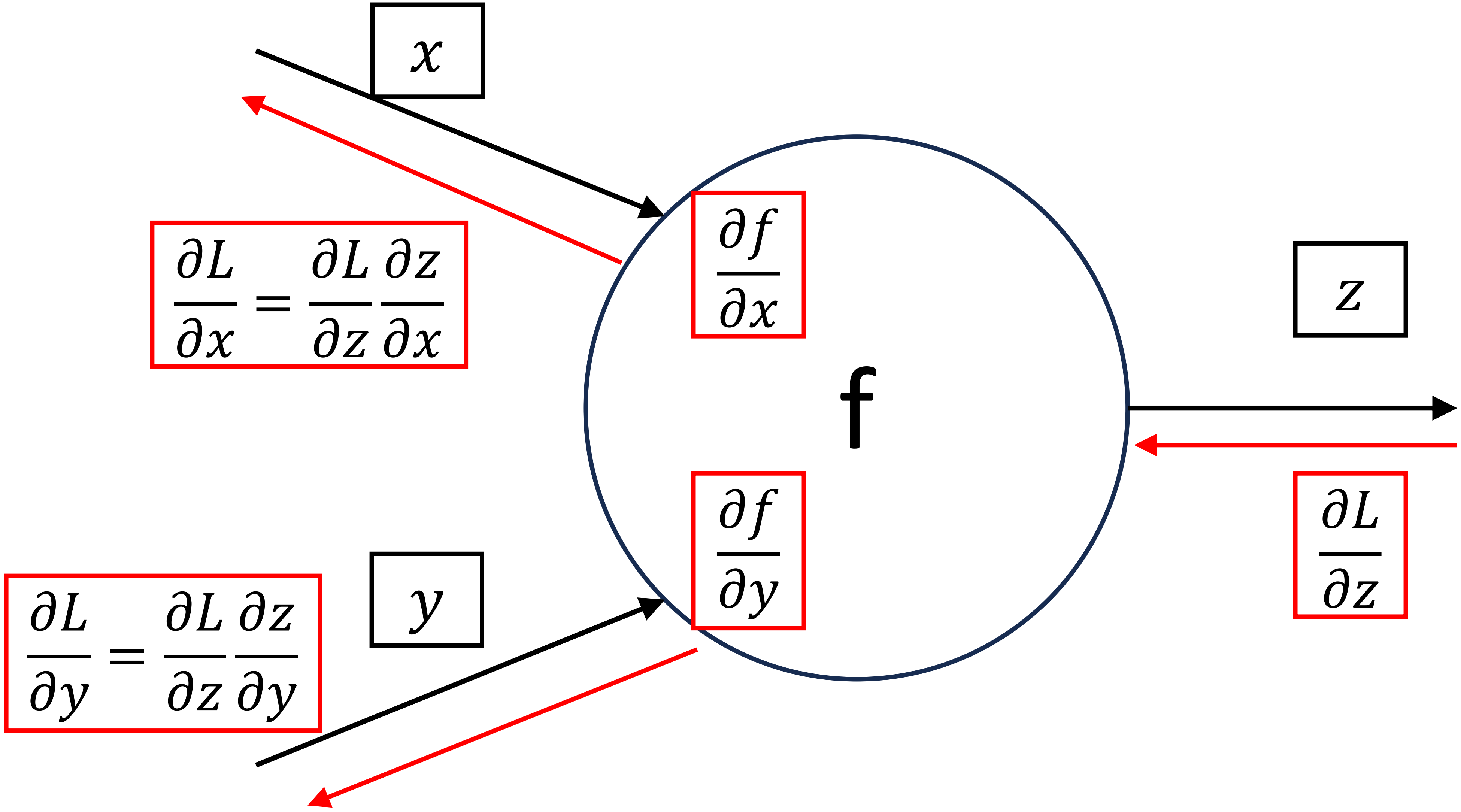
$$\frac{\partial(d*c)}{\partial c} = d$$



Grafo Computacional - forward



Grafo Computacional - *Backward*



Grafo Computacional

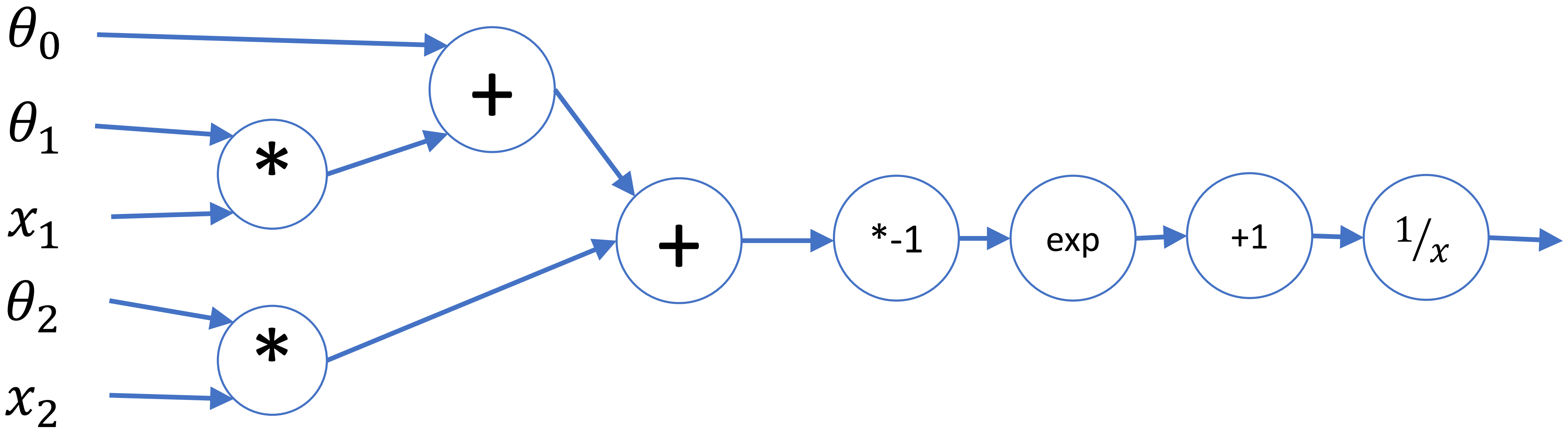
- Vamos ver outro exemplo:

$$f(\theta, x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$$

Grafo Computacional

- Vamos ver outro exemplo:

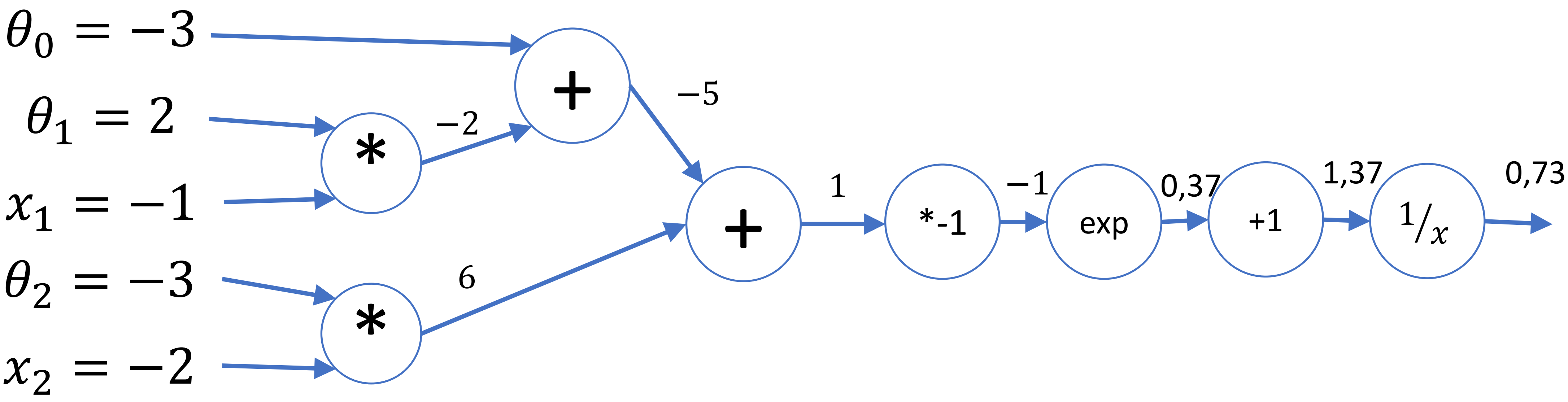
$$f(\theta, x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$$



Grafo Computacional

- Vamos ver outro exemplo:

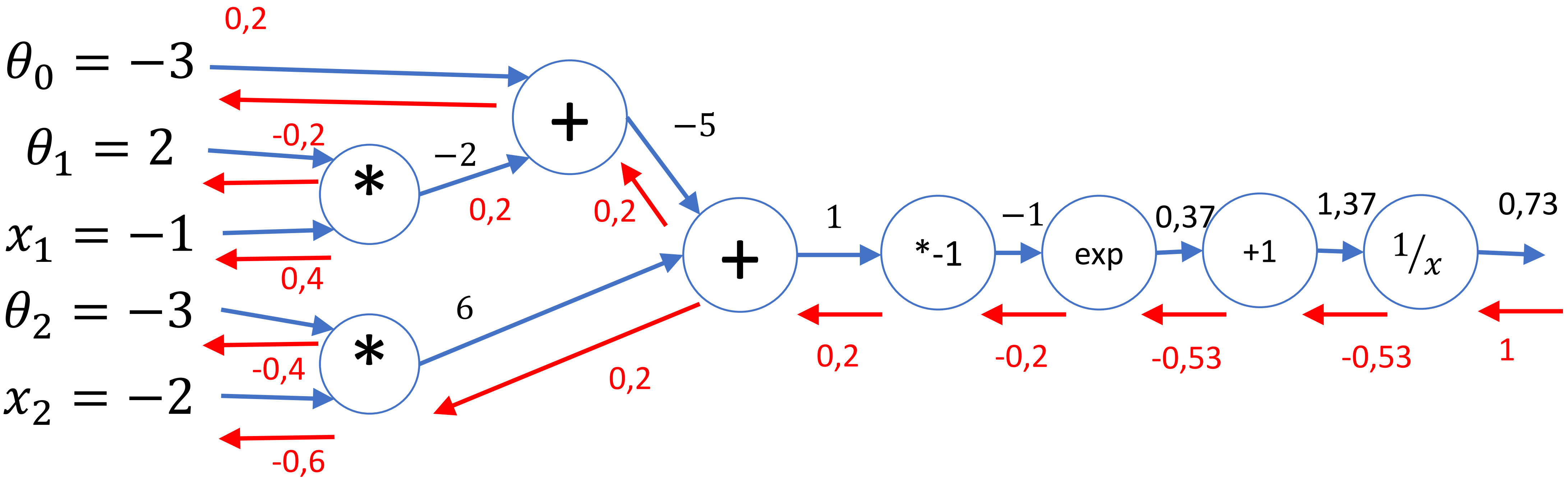
$$f(\theta, x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$$



Grafo Computacional

- Vamos ver outro exemplo:

$$f(\theta, x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$$



Grafo Computacional

- Vamos ver outro exemplo:

$$f(\theta, x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$$