Rete neurale 3→3→2→1: forward, MSE e backprop (SGD su 3 esempi)

1. Dati, normalizzazione e obiettivo

Input: altezza (cm), peso (kg), età (anni). Output: rischio obesità $y \in \{0, 1\}$ (0=no, 1=sì).

$$x_1 = \frac{\text{altezza}}{200}, \qquad x_2 = \frac{\text{peso}}{150}, \qquad x_3 = \frac{\text{età}}{100}$$

Tabella training (10 esempi). In questo documento useremo per SGD "vero" i primi 3 esempi (#1,#2,#3) in sequenza.

#	Altezza	Peso	Età	y	x_1	x_2	x_3
1	170	65	25	0	0.850	0.4333	0.25
2	160	80	45	1	0.800	0.5333	0.45
3	180	90	35	1	0.900	0.6000	0.35
4	175	70	30	0	0.875	0.4667	0.30
5	155	95	50	1	0.775	0.6333	0.50
6	165	55	22	0	0.825	0.3667	0.22
7	172	110	40	1	0.860	0.7333	0.40
8	168	60	28	0	0.840	0.4000	0.28
9	158	75	33	1	0.790	0.5000	0.33
10	185	78	27	0	0.925	0.5200	0.27

2. Architettura e funzioni

Architettura: 3 input → 3 neuroni → 2 neuroni → 1 neurone.

Attivazione (tutti i neuroni): sigmoide $\sigma(z) = \frac{1}{1+e^{-z}}$, con derivata $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.

Loss (MSE su 1 esempio):

$$L = (\hat{y} - y)^2$$

Learning rate: $\eta = 0.5$.

2.1 Schema

```
x1,x2,x3 -> [a1,a2,a3] -> [g1,g2] -> yhat

con bias in ogni neurone.
```

3. Forward (formule, senza matrici)

Strato 1:

$$z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1, \qquad a_1 = \sigma(z_1)$$

$$z_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2, \qquad a_2 = \sigma(z_2)$$

$$z_3 = w_{31}x_1 + w_{32}x_2 + w_{33}x_3 + b_3, \qquad a_3 = \sigma(z_3)$$

Strato 2:

$$s_1 = v_{11}a_1 + v_{12}a_2 + v_{13}a_3 + c_1, \qquad g_1 = \sigma(s_1)$$

$$s_2 = v_{21}a_1 + v_{22}a_2 + v_{23}a_3 + c_2, \qquad g_2 = \sigma(s_2)$$

Output:

$$t = u_1g_1 + u_2g_2 + d, \qquad \hat{y} = \sigma(t)$$

4. Backprop: dimostrazione della regola $\frac{dL}{dw_{11}} = \delta_{a1}x_1$ e $\frac{dL}{db_1} = \delta_{a1}$

Per il neurone a_1 (strato 1): $z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1$, $a_1 = \sigma(z_1)$.

$$\frac{dL}{dw_{11}} = \frac{dL}{dz_1} \cdot \frac{dz_1}{dw_{11}}$$

Definisco $\delta_{a1} \equiv \frac{dL}{dz_1}$. Poiché z_1 dipende linearmente da w_{11}: $\frac{dz_1}{dw_{11}} = x_1$. Quindi:

$$\boxed{\frac{dL}{dw_{11}} = \delta_{a1}\, x_1}$$

Analogamente $\frac{dz_1}{dw_{12}} = x_2$ e $\frac{dz_1}{dw_{13}} = x_3$, quindi $\frac{dL}{dw_{12}} = \delta_{a1}x_2$, $\frac{dL}{dw_{13}} = \delta_{a1}x_3$.
Per il bias: $\frac{dz_1}{db_1} = 1$, quindi:

$$\boxed{\frac{dL}{db_1} = \delta_{a1}}$$

5. Backprop (deltas e gradienti, formule generali)

Output:

$\mathrm{dL}_{\overline{d\hat{y}=2(\hat{y}-y)}}$
$\mathrm{d}_{\overline{dt=\hat{y}(1-\hat{y})}}$
$\delta_{out} = \frac{dL}{dt} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dt} = 2(\hat{y}-y)\hat{y}(1-\hat{y})$

$$\frac{dL}{du_1} = \delta_{out} g_1, \quad \frac{dL}{du_2} = \delta_{out} g_2, \quad \frac{dL}{dd} = \delta_{out}$$

Strato 2:

$$\delta_{g1} = \frac{dL}{ds_1} = \delta_{out} u_1 g_1 (1 - g_1), \qquad \delta_{g2} = \frac{dL}{ds_2} = \delta_{out} u_2 g_2 (1 - g_2)$$

$$\frac{dL}{dv_{1j}} = \delta_{g1} a_j, \ \frac{dL}{dc_1} = \delta_{g1} \qquad \frac{dL}{dv_{2j}} = \delta_{g2} a_j, \ \frac{dL}{dc_2} = \delta_{g2}$$

Strato 1:

$\delta_{a1} = \frac{dL}{dz_1} = (\delta_{g1} v_{11} + \delta_{g2} v_{21}) a_1 (1 - a_1)$

$\delta_{a2} = \frac{dL}{dz_2} = (\delta_{g1} v_{12} + \delta_{g2} v_{22}) a_2 (1 - a_2)$

$\delta_{a3} = \frac{dL}{dz_3} = (\delta_{g1} v_{13} + \delta_{g2} v_{23}) a_3 (1 - a_3)$

$$\frac{dL}{dw_{i1}} = \delta_{ai} x_1, \ \frac{dL}{dw_{i2}} = \delta_{ai} x_2, \ \frac{dL}{dw_{i3}} = \delta_{ai} x_3, \ \frac{dL}{db_i} = \delta_{ai}$$

Update SGD: $\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$ con $\eta = 0.5$.

6. Esecuzione numerica completa: 3 iterazioni SGD sui primi 3 esempi

6.1 Parametri iniziali (prima di iterazione 1)

Strato 1 (w, b):
$w_{11} = 0.100000,\ w_{12} = -0.200000,\ w_{13} = 0.100000,\ b_1 = 0$
$w_{21} = -0.100000,\ w_{22} = 0.100000,\ w_{23} = 0.200000,\ b_2 = 0$
$w_{31} = 0.050000,\ w_{32} = 0.100000,\ w_{33} = -0.100000,\ b_3 = 0$

Strato 2 (v, c):
$v_{11} = 0.100000,\ v_{12} = -0.100000,\ v_{13} = 0.050000,\ c_1 = 0$
$v_{21} = -0.050000,\ v_{22} = 0.100000,\ v_{23} = 0.100000,\ c_2 = 0$

Output (u, d):
$u_1 = 0.100000,\ u_2 = -0.100000,\ d = 0$

Iterazione 1: SGD su esempio #1

Input: $x_1 = 0.850000,\ x_2 = 0.433300,\ x_3 = 0.250000,\ y = 0$

Parametri all'inizio dell'iterazione

Strato 1 (w, b):
$w_{11} = 0.100000,\ w_{12} = -0.200000,\ w_{13} = 0.100000,\ b_1 = 0$
$w_{21} = -0.100000,\ w_{22} = 0.100000,\ w_{23} = 0.200000,\ b_2 = 0$
$w_{31} = 0.050000,\ w_{32} = 0.100000,\ w_{33} = -0.100000,\ b_3 = 0$

Strato 2 (v, c):
$v_{11} = 0.100000,\ v_{12} = -0.100000,\ v_{13} = 0.050000,\ c_1 = 0$
$v_{21} = -0.050000,\ v_{22} = 0.100000,\ v_{23} = 0.100000,\ c_2 = 0$

Output (u, d):
$u_1 = 0.100000,\ u_2 = -0.100000,\ d = 0$

Forward (tutti i valori intermedi)

Strato 1:

$z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1 = (0.100000)(0.850000) + (-0.200000)(0.433300) + (0.100000)(0.250000) + (0)$
$a_1 = \sigma(z_1) = \sigma(0.023340) = 0.505835$

$z_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2 = (-0.100000)(0.850000) + (0.100000)(0.433300) + (0.200000)(0.250000) + (0)$
$a_2 = \sigma(z_2) = \sigma(0.008330) = 0.502082$

$z_3 = w_{31}x_1 + w_{32}x_2 + w_{33}x_3 + b_3 = (0.050000)(0.850000) + (0.100000)(0.433300) + (-0.100000)(0.250000) + (0)$
$a_3 = \sigma(z_3) = \sigma(0.060830) = 0.515203$

Strato 2:

$s_1 = v_{11}a_1 + v_{12}a_2 + v_{13}a_3 + c_1 = (0.100000)(0.505835) + (-0.100000)(0.502082) + (0.050000)(0.515203) + (0) =$
$g_1 = \sigma(s_1) = \sigma(0.026135) = 0.506533$

$s_2 = v_{21}a_1 + v_{22}a_2 + v_{23}a_3 + c_2 = (-0.050000)(0.505835) + (0.100000)(0.502082) + (0.100000)(0.515203) + (0) =$
$g_2 = \sigma(s_2) = \sigma(0.076437) = 0.519100$

Output:

$t = u_1g_1 + u_2g_2 + d = (0.100000)(0.506533) + (-0.100000)(0.519100) + (0) = -0.001257$
$\hat{y} = \sigma(t) = \sigma(-0.001257) = 0.499686$

Loss (MSE su 1 esempio)

$$L = (\hat{y} - y)^2 = (0.499686 - 0)^2 = 0.249686$$

Backprop: deltas (error signals)

$$\frac{dL}{d\hat{y}} = 2(\hat{y} - y) = 2(0.499686 - 0) = 0.999372$$

$$\frac{d\hat{y}}{dt} = \hat{y}(1 - \hat{y}) = (0.499686)(1 - 0.499686) = 0.250000$$

$$\delta_{out} = \frac{dL}{dt} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dt} = 0.999372 \cdot 0.250000 = 0.249843$$

$$\delta_{g1} = \frac{dL}{ds_1} = \delta_{out} u_1 g_1(1 - g_1) = 0.249843 \cdot 0.100000 \cdot (0.506533)(1 - 0.506533) = 0.006245$$

$$\delta_{g2} = \frac{dL}{ds_2} = \delta_{out} u_2 g_2(1 - g_2) = 0.249843 \cdot -0.100000 \cdot (0.519100)(1 - 0.519100) = -0.006237$$

$$\delta_{a1} = \frac{dL}{dz_1} = (\delta_{g1} v_{11} + \delta_{g2} v_{21})\, a_1(1 - a_1) = (0.006245 \cdot 0.100000 + -0.006237 \cdot -0.050000) \cdot (0.505835)(1 - 0.5058$$

$$\delta_{a2} = \frac{dL}{dz_2} = (\delta_{g1} v_{12} + \delta_{g2} v_{22})\, a_2(1 - a_2) = (0.006245 \cdot -0.100000 + -0.006237 \cdot 0.100000) \cdot (0.502082)(1 - 0.5020$$

$$\delta_{a3} = \frac{dL}{dz_3} = (\delta_{g1} v_{13} + \delta_{g2} v_{23})\, a_3(1 - a_3) = (0.006245 \cdot 0.050000 + -0.006237 \cdot 0.100000) \cdot (0.515203)(1 - 0.51520$$

Tutti i gradienti (23 valori)

$$\frac{dL}{du_1} = \delta_{out} g_1 = 0.249843 \cdot 0.506533 = 0.126554$$
$$\frac{dL}{du_2} = \delta_{out} g_2 = 0.249843 \cdot 0.519100 = 0.129693$$
$$\frac{dL}{dd} = \delta_{out} = 0.249843$$

$$\frac{dL}{dv_{11}} = \delta_{g1} a_1 = 0.006245 \cdot 0.505835 = 0.003159$$
$$\frac{dL}{dv_{12}} = \delta_{g1} a_2 = 0.006245 \cdot 0.502082 = 0.003136$$
$$\frac{dL}{dv_{13}} = \delta_{g1} a_3 = 0.006245 \cdot 0.515203 = 0.003217$$
$$\frac{dL}{dc_1} = \delta_{g1} = 0.006245$$

$$\frac{dL}{dv_{21}} = \delta_{g2} a_1 = -0.006237 \cdot 0.505835 = -0.003155$$
$$\frac{dL}{dv_{22}} = \delta_{g2} a_2 = -0.006237 \cdot 0.502082 = -0.003131$$
$$\frac{dL}{dv_{23}} = \delta_{g2} a_3 = -0.006237 \cdot 0.515203 = -0.003213$$
$$\frac{dL}{dc_2} = \delta_{g2} = -0.006237$$

$$\frac{dL}{dw_{11}} = \delta_{a1} x_1 = 0.000234 \cdot 0.850000 = 0.000199$$
$$\frac{dL}{dw_{12}} = \delta_{a1} x_2 = 0.000234 \cdot 0.433300 = 0.000101$$
$$\frac{dL}{dw_{13}} = \delta_{a1} x_3 = 0.000234 \cdot 0.250000 = 5.851379e-05$$
$$\frac{dL}{db_1} = \delta_{a1} = 0.000234$$

$$\frac{dL}{dw_{21}} = \delta_{a2} x_1 = -0.000312 \cdot 0.850000 = -0.000265$$
$$\frac{dL}{dw_{22}} = \delta_{a2} x_2 = -0.000312 \cdot 0.433300 = -0.000135$$
$$\frac{dL}{dw_{23}} = \delta_{a2} x_3 = -0.000312 \cdot 0.250000 = -7.801090e-05$$
$$\frac{dL}{db_2} = \delta_{a2} = -0.000312$$

$$\frac{dL}{dw_{31}} = \delta_{a3} x_1 = -7.778937e-05 \cdot 0.850000 = -6.612096e-05$$
$$\frac{dL}{dw_{32}} = \delta_{a3} x_2 = -7.778937e-05 \cdot 0.433300 = -3.370613e-05$$
$$\frac{dL}{dw_{33}} = \delta_{a3} x_3 = -7.778937e-05 \cdot 0.250000 = -1.944734e-05$$
$$\frac{dL}{db_3} = \delta_{a3} = -7.778937e-05$$

Update (SGD)

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta} \qquad (\eta = 0.5)$$

Parametri dopo l'update (fine iterazione):
Strato 1 (w, b):
$w_{11} = 0.099901,\ w_{12} = -0.200051,\ w_{13} = 0.099971,\ b_1 = -0.000117$
$w_{21} = -0.099867,\ w_{22} = 0.100068,\ w_{23} = 0.200039,\ b_2 = 0.000156$
$w_{31} = 0.050033,\ w_{32} = 0.100017,\ w_{33} = -0.099990,\ b_3 = 3.889468e-05$

Strato 2 (v, c):
$v_{11} = 0.098421,\ v_{12} = -0.101568,\ v_{13} = 0.048391,\ c_1 = -0.003123$
$v_{21} = -0.048423,\ v_{22} = 0.101566,\ v_{23} = 0.101607,\ c_2 = 0.003118$

Output (u, d):
$u_1 = 0.036723,\ u_2 = -0.164847,\ d = -0.124921$

Calcolo esplicito degli aggiornamenti (Iterazione 1)

Regola teorica di update (SGD su 1 esempio). Per ogni parametro scalare θ:

$$\theta^{(1)} = \theta^{(0)} - \eta \left.\frac{\partial L}{\partial \theta}\right|_{\theta^{(0)},\,\text{esempio } \#1}$$

dove:

$$\theta^{(0)} = \text{valore prima dell'update}, \quad \eta = 0.5 = \text{learning rate},$$

$\left.\frac{\partial L}{\partial \theta}\right|_{\theta^{(0)},\#1}$ = gradiente calcolato nella backprop dell'iterazione 1.

Strato 1. Esempio esplicito su w_{11}:

$$w_{11}^{(1)} = w_{11}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{11}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot 0.000199 = 0.0999005 \approx 0.099901$$

$$w_{12}^{(1)} = w_{12}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{12}}\right|_{\theta^{(0)},\#1} = -0.200000 - 0.5 \cdot 0.000101 = -0.2000505 \approx -0.200051$$

$$w_{13}^{(1)} = w_{13}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{13}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot 5.851379 \times 10^{-5} = 0.0999707 \approx 0.099971$$

$$b_1^{(1)} = b_1^{(0)} - \eta \left.\frac{\partial L}{\partial b_1}\right|_{\theta^{(0)},\#1} = 0 - 0.5 \cdot 0.000234 = -0.000117$$

$$w_{21}^{(1)} = w_{21}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{21}}\right|_{\theta^{(0)},\#1} = -0.100000 - 0.5 \cdot (-0.000265) = -0.0998675 \approx -0.099867$$

$$w_{22}^{(1)} = w_{22}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{22}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot (-0.000135) = 0.1000675 \approx 0.100068$$

$$w_{23}^{(1)} = w_{23}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{23}}\right|_{\theta^{(0)},\#1} = 0.200000 - 0.5 \cdot (-7.801090 \times 10^{-5}) = 0.2000390 \approx 0.200039$$

$$b_2^{(1)} = b_2^{(0)} - \eta \left.\frac{\partial L}{\partial b_2}\right|_{\theta^{(0)},\#1} = 0 - 0.5 \cdot (-0.000312) = 0.000156$$

$$w_{31}^{(1)} = w_{31}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{31}}\right|_{\theta^{(0)},\#1} = 0.050000 - 0.5 \cdot (-6.612096 \times 10^{-5}) = 0.0500331 \approx 0.050033$$

$$w_{32}^{(1)} = w_{32}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{32}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot (-3.370613 \times 10^{-5}) = 0.1000169 \approx 0.100017$$

$$w_{33}^{(1)} = w_{33}^{(0)} - \eta \left.\frac{\partial L}{\partial w_{33}}\right|_{\theta^{(0)},\#1} = -0.100000 - 0.5 \cdot (-1.944734 \times 10^{-5}) = -0.0999903 \approx -0.099990$$

$$b_3^{(1)} = b_3^{(0)} - \eta \left.\frac{\partial L}{\partial b_3}\right|_{\theta^{(0)},\#1} = 0 - 0.5 \cdot (-7.778937 \times 10^{-5}) = 3.889468 \times 10^{-5}$$

Strato 2.

$$v_{11}^{(1)} = v_{11}^{(0)} - \eta \left.\frac{\partial L}{\partial v_{11}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot 0.003159 = 0.0984205 \approx 0.098421$$

$$v_{12}^{(1)} = v_{12}^{(0)} - \eta \left.\frac{\partial L}{\partial v_{12}}\right|_{\theta^{(0)},\#1} = -0.100000 - 0.5 \cdot 0.003136 = -0.101568$$

$$v_{13}^{(1)} = v_{13}^{(0)} - \eta \left.\frac{\partial L}{\partial v_{13}}\right|_{\theta^{(0)},\#1} = 0.050000 - 0.5 \cdot 0.003217 = 0.0483915 \approx 0.048391$$

$$c_1^{(1)} = c_1^{(0)} - \eta \left.\frac{\partial L}{\partial c_1}\right|_{\theta^{(0)},\#1} = 0 - 0.5 \cdot 0.006245 = -0.003123$$

$$v_{21}^{(1)} = v_{21}^{(0)} - \eta \left.\frac{\partial L}{\partial v_{21}}\right|_{\theta^{(0)},\#1} = -0.050000 - 0.5 \cdot (-0.003155) = -0.0484225 \approx -0.048423$$

$$v_{22}^{(1)} = v_{22}^{(0)} - \eta \left.\frac{\partial L}{\partial v_{22}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot (-0.003131) = 0.1015655 \approx 0.101566$$

$$v_{23}^{(1)} = v_{23}^{(0)} - \eta \left.\frac{\partial L}{\partial v_{23}}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot (-0.003213) = 0.1016065 \approx 0.101607$$

$$c_2^{(1)} = c_2^{(0)} - \eta \left.\frac{\partial L}{\partial c_2}\right|_{\theta^{(0)},\#1} = 0 - 0.5 \cdot (-0.006237) = 0.0031185 \approx 0.003118$$

Output.

$$u_1^{(1)} = u_1^{(0)} - \eta \left.\frac{\partial L}{\partial u_1}\right|_{\theta^{(0)},\#1} = 0.100000 - 0.5 \cdot 0.126554 = 0.036723$$

$$u_2^{(1)} = u_2^{(0)} - \eta \left.\frac{\partial L}{\partial u_2}\right|_{\theta^{(0)},\#1} = -0.100000 - 0.5 \cdot 0.129693 = -0.1648465 \approx -0.164847$$

$$d^{(1)} = d^{(0)} - \eta \left.\frac{\partial L}{\partial d}\right|_{\theta^{(0)},\#1} = 0 - 0.5 \cdot 0.249843 = -0.1249215 \approx -0.124921$$

Iterazione 2: SGD su esempio #2

Input: $x_1 = 0.800000,\ x_2 = 0.533300,\ x_3 = 0.450000,\ y = 1$

Parametri all'inizio dell'iterazione

Strato 1 (w, b):
$w_{11} = 0.099901,\ w_{12} = -0.200051,\ w_{13} = 0.099971,\ b_1 = -0.000117$
$w_{21} = -0.099867,\ w_{22} = 0.100068,\ w_{23} = 0.200039,\ b_2 = 0.000156$
$w_{31} = 0.050033,\ w_{32} = 0.100017,\ w_{33} = -0.099990,\ b_3 = 3.889468e-05$

Strato 2 (v, c):
$v_{11} = 0.098421,\ v_{12} = -0.101568,\ v_{13} = 0.048391,\ c_1 = -0.003123$
$v_{21} = -0.048423,\ v_{22} = 0.101566,\ v_{23} = 0.101607,\ c_2 = 0.003118$

Output (u, d):
$u_1 = 0.036723,\ u_2 = -0.164847,\ d = -0.124921$

Forward (tutti i valori intermedi)

Strato 1:

$z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1 = (0.099901)(0.800000) + (-0.200051)(0.533300) + (0.099971)(0.450000) + (-0$
$a_1 = \sigma(z_1) = \sigma(0.018103) = 0.504526$

$z_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2 = (-0.099867)(0.800000) + (0.100068)(0.533300) + (0.200039)(0.450000) + (0.0$
$a_2 = \sigma(z_2) = \sigma(0.063646) = 0.515906$

$z_3 = w_{31}x_1 + w_{32}x_2 + w_{33}x_3 + b_3 = (0.050033)(0.800000) + (0.100017)(0.533300) + (-0.099990)(0.450000) + (3.8$
$a_3 = \sigma(z_3) = \sigma(0.048409) = 0.512100$

Strato 2:

$s_1 = v_{11}a_1 + v_{12}a_2 + v_{13}a_3 + c_1 = (0.098421)(0.504526) + (-0.101568)(0.515906) + (0.048391)(0.512100) + (-0.0$
$g_1 = \sigma(s_1) = \sigma(0.018915) = 0.504729$

$s_2 = v_{21}a_1 + v_{22}a_2 + v_{23}a_3 + c_2 = (-0.048423)(0.504526) + (0.101566)(0.515906) + (0.101607)(0.512100) + (0.00$
$g_2 = \sigma(s_2) = \sigma(0.083119) = 0.520768$

Output:

$t = u_1g_1 + u_2g_2 + d = (0.036723)(0.504729) + (-0.164847)(0.520768) + (-0.124921) = -0.192233$
$\hat{y} = \sigma(t) = \sigma(-0.192233) = 0.452089$

Loss (MSE su 2 esempio)

$$L = (\hat{y} - y)^2 = (0.452089 - 1)^2 = 0.300206$$

Backprop: deltas (error signals)

$$\frac{dL}{d\hat{y}} = 2(\hat{y} - y) = 2(0.452089 - 1) = -1.095822$$

$$\frac{d\hat{y}}{dt} = \hat{y}(1 - \hat{y}) = (0.452089)(1 - 0.452089) = 0.247705$$

$$\delta_{out} = \frac{dL}{dt} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dt} = -1.095822 \cdot 0.247705 = -0.271440$$

$$\delta_{g1} = \frac{dL}{ds_1} = \delta_{out} u_1 g_1(1 - g_1) = -0.271440 \cdot 0.036723 \cdot (0.504729)(1 - 0.504729) = -0.002492$$

$$\delta_{g2} = \frac{dL}{ds_2} = \delta_{out} u_2 g_2(1 - g_2) = -0.271440 \cdot -0.164847 \cdot (0.520768)(1 - 0.520768) = 0.011167$$

$$\delta_{a1} = \frac{dL}{dz_1} = (\delta_{g1} v_{11} + \delta_{g2} v_{21})\, a_1(1 - a_1) = (-0.002492 \cdot 0.098421 + 0.011167 \cdot -0.048423) \cdot (0.504526)(1 - 0.5045$$

$$\delta_{a2} = \frac{dL}{dz_2} = (\delta_{g1} v_{12} + \delta_{g2} v_{22})\, a_2(1 - a_2) = (-0.002492 \cdot -0.101568 + 0.011167 \cdot 0.101566) \cdot (0.515906)(1 - 0.5159$$

$$\delta_{a3} = \frac{dL}{dz_3} = (\delta_{g1} v_{13} + \delta_{g2} v_{23})\, a_3(1 - a_3) = (-0.002492 \cdot 0.048391 + 0.011167 \cdot 0.101607) \cdot (0.512100)(1 - 0.51210$$

Tutti i gradienti (23 valori)

$$\frac{dL}{du_1} = \delta_{out} g_1 = -0.271440 \cdot 0.504729 = -0.137004$$
$$\frac{dL}{du_2} = \delta_{out} g_2 = -0.271440 \cdot 0.520768 = -0.141357$$
$$\frac{dL}{dd} = \delta_{out} = -0.271440$$

$$\frac{dL}{dv_{11}} = \delta_{g1} a_1 = -0.002492 \cdot 0.504526 = -0.001257$$
$$\frac{dL}{dv_{12}} = \delta_{g1} a_2 = -0.002492 \cdot 0.515906 = -0.001286$$
$$\frac{dL}{dv_{13}} = \delta_{g1} a_3 = -0.002492 \cdot 0.512100 = -0.001276$$
$$\frac{dL}{dc_1} = \delta_{g1} = -0.002492$$

$$\frac{dL}{dv_{21}} = \delta_{g2} a_1 = 0.011167 \cdot 0.504526 = 0.005634$$
$$\frac{dL}{dv_{22}} = \delta_{g2} a_2 = 0.011167 \cdot 0.515906 = 0.005761$$
$$\frac{dL}{dv_{23}} = \delta_{g2} a_3 = 0.011167 \cdot 0.512100 = 0.005719$$
$$\frac{dL}{dc_2} = \delta_{g2} = 0.011167$$

$$\frac{dL}{dw_{11}} = \delta_{a1} x_1 = -0.000196 \cdot 0.800000 = -0.000157$$
$$\frac{dL}{dw_{12}} = \delta_{a1} x_2 = -0.000196 \cdot 0.533300 = -0.000105$$
$$\frac{dL}{dw_{13}} = \delta_{a1} x_3 = -0.000196 \cdot 0.450000 = -8.841657e-05$$
$$\frac{dL}{db_1} = \delta_{a1} = -0.000196$$

$$\frac{dL}{dw_{21}} = \delta_{a2} x_1 = 0.000346 \cdot 0.800000 = 0.000277$$
$$\frac{dL}{dw_{22}} = \delta_{a2} x_2 = 0.000346 \cdot 0.533300 = 0.000185$$
$$\frac{dL}{dw_{23}} = \delta_{a2} x_3 = 0.000346 \cdot 0.450000 = 0.000156$$
$$\frac{dL}{db_2} = \delta_{a2} = 0.000346$$

$$\frac{dL}{dw_{31}} = \delta_{a3} x_1 = 0.000253 \cdot 0.800000 = 0.000203$$
$$\frac{dL}{dw_{32}} = \delta_{a3} x_2 = 0.000253 \cdot 0.533300 = 0.000135$$
$$\frac{dL}{dw_{33}} = \delta_{a3} x_3 = 0.000253 \cdot 0.450000 = 0.000114$$
$$\frac{dL}{db_3} = \delta_{a3} = 0.000253$$

Update (SGD)

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta} \qquad (\eta = 0.5)$$

Parametri dopo l'update (fine iterazione):
Strato 1 (w, b):
$w_{11} = 0.099979,\ w_{12} = -0.199998,\ w_{13} = 0.100015,\ b_1 = -1.878695e-05$
$w_{21} = -0.100006,\ w_{22} = 0.099975,\ w_{23} = 0.199961,\ b_2 = -1.721421e-05$
$w_{31} = 0.049932,\ w_{32} = 0.099949,\ w_{33} = -0.100047,\ b_3 = -8.779105e-05$

Strato 2 (v, c):
$v_{11} = 0.099049,\ v_{12} = -0.100925,\ v_{13} = 0.049029,\ c_1 = -0.001877$
$v_{21} = -0.051240,\ v_{22} = 0.098685,\ v_{23} = 0.098747,\ c_2 = -0.002465$

Output (u, d):
$u_1 = 0.105225,\ u_2 = -0.094168,\ d = 0.010799$

Calcolo esplicito degli aggiornamenti (Iterazione 2)

Regola teorica di update (SGD su 1 esempio).

$$\theta^{(2)} = \theta^{(1)} - \eta \left.\frac{\partial L}{\partial \theta}\right|_{\theta^{(1)},\,\text{esempio } \#2}$$

dove $\eta = 0.5$ e il gradiente è quello calcolato nella backprop dell'iterazione 2.

Strato 1.

$$w_{11}^{(2)} = w_{11}^{(1)} - \eta \left.\frac{\partial L}{\partial w_{11}}\right|_{\theta^{(1)},\,\#2} = 0.099901 - 0.5 \cdot (-0.000157) = 0.0999795 \approx 0.099979$$

$$w_{12}^{(2)} = w_{12}^{(1)} - \eta \left.\frac{\partial L}{\partial w_{12}}\right|_{\theta^{(1)},\,\#2} = -0.200051 - 0.5 \cdot (-0.000105) = -0.1999985 \approx -0.199998$$

$$w_{13}^{(2)} = w_{13}^{(1)} - \eta \left.\frac{\partial L}{\partial w_{13}}\right|_{\theta^{(1)},\,\#2} = 0.099971 - 0.5 \cdot (-8.841657 \times 10^{-5}) = 0.1000152 \approx 0.100015$$

$$b_1^{(2)} = b_1^{(1)} - \eta \left.\frac{\partial L}{\partial b_1}\right|_{\theta^{(1)},\,\#2} = -0.000117 - 0.5 \cdot (-0.000196) = -1.9 \times 10^{-5} \approx -1.878695 \times 10^{-5}$$

(...e così via identico per tutti gli altri parametri, usando sempre $\theta^{(1)}$ come "valore prima" e i gradienti dell'iterazione 2.)

Iterazione 3: SGD su esempio #3

Input: $x_1 = 0.900000,\ x_2 = 0.600000,\ x_3 = 0.350000,\ y = 1$

Parametri all'inizio dell'iterazione

Strato 1 (w, b):
$w_{11} = 0.099979,\ w_{12} = -0.199998,\ w_{13} = 0.100015,\ b_1 = -1.878695e-05$
$w_{21} = -0.100006,\ w_{22} = 0.099975,\ w_{23} = 0.199961,\ b_2 = -1.721421e-05$
$w_{31} = 0.049932,\ w_{32} = 0.099949,\ w_{33} = -0.100047,\ b_3 = -8.779105e-05$

Strato 2 (v, c):
$v_{11} = 0.099049,\ v_{12} = -0.100925,\ v_{13} = 0.049029,\ c_1 = -0.001877$

$v_{21} = -0.051240,\ v_{22} = 0.098685,\ v_{23} = 0.098747,\ c_2 = -0.002465$

Output (u, d):

$u_1 = 0.105225,\ u_2 = -0.094168,\ d = 0.010799$

Forward (tutti i valori intermedi)

Strato 1:

$z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1 = (0.099979)(0.900000) + (-0.199998)(0.600000) + (0.100015)(0.350000) + (-$

$a_1 = \sigma(z_1) = \sigma(0.004969) = 0.501242$

$z_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2 = (-0.100006)(0.900000) + (0.099975)(0.600000) + (0.199961)(0.350000) + (-$

$a_2 = \sigma(z_2) = \sigma(0.039949) = 0.509986$

$z_3 = w_{31}x_1 + w_{32}x_2 + w_{33}x_3 + b_3 = (0.049932)(0.900000) + (0.099949)(0.600000) + (-0.100047)(0.350000) + (-$

$a_3 = \sigma(z_3) = \sigma(0.069804) = 0.517444$

Strato 2:

$s_1 = v_{11}a_1 + v_{12}a_2 + v_{13}a_3 + c_1 = (0.099049)(0.501242) + (-0.100925)(0.509986) + (0.049029)(0.517444) + (-0.0$

$g_1 = \sigma(s_1) = \sigma(0.021671) = 0.505417$

$s_2 = v_{21}a_1 + v_{22}a_2 + v_{23}a_3 + c_2 = (-0.051240)(0.501242) + (0.098685)(0.509986) + (0.098747)(0.517444) + (-0.0$

$g_2 = \sigma(s_2) = \sigma(0.073276) = 0.518311$

Output:

$t = u_1g_1 + u_2g_2 + d = (0.105225)(0.505417) + (-0.094168)(0.518311) + (0.010799) = 0.015173$

$\hat{y} = \sigma(t) = \sigma(0.015173) = 0.503793$

Loss (MSE su 3 esempio)

$$L = (\hat{y} - y)^2 = (0.503793 - 1)^2 = 0.246221$$

Backprop: deltas (error signals)

$$\frac{dL}{d\hat{y}} = 2(\hat{y} - y) = 2(0.503793 - 1) = -0.992414$$

$$\frac{d\hat{y}}{dt} = \hat{y}(1 - \hat{y}) = (0.503793)(1 - 0.503793) = 0.249986$$

$$\delta_{out} = \frac{dL}{dt} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dt} = -0.992414 \cdot 0.249986 = -0.248089$$

$$\delta_{g1} = \frac{dL}{ds_1} = \delta_{out}u_1g_1(1 - g_1) = -0.248089 \cdot 0.105225 \cdot (0.505417)(1 - 0.505417) = -0.006526$$

$$\delta_{g2} = \frac{dL}{ds_2} = \delta_{out}u_2g_2(1 - g_2) = -0.248089 \cdot -0.094168 \cdot (0.518311)(1 - 0.518311) = 0.005833$$

$$\delta_{a1} = \frac{dL}{dz_1} = (\delta_{g1}v_{11} + \delta_{g2}v_{21})\, a_1(1 - a_1) = (-0.006526 \cdot 0.099049 + 0.005833 \cdot -0.051240) \cdot (0.501242)(1 - 0.5012$$

$$\delta_{a2} = \frac{dL}{dz_2} = (\delta_{g1}v_{12} + \delta_{g2}v_{22})\, a_2(1 - a_2) = (-0.006526 \cdot -0.100925 + 0.005833 \cdot 0.098685) \cdot (0.509986)(1 - 0.5099$$

$$\delta_{a3} = \frac{dL}{dz_3} = (\delta_{g1}v_{13} + \delta_{g2}v_{23})\, a_3(1 - a_3) = (-0.006526 \cdot 0.049029 + 0.005833 \cdot 0.098747) \cdot (0.517444)(1 - 0.51744$$

Tutti i gradienti (23 valori)

$$\frac{dL}{du_1} = \delta_{out} g_1 = -0.248089 \cdot 0.505417 = -0.125389$$
$$\frac{dL}{du_2} = \delta_{out} g_2 = -0.248089 \cdot 0.518311 = -0.128587$$
$$\frac{dL}{dd} = \delta_{out} = -0.248089$$

$$\frac{dL}{dv_{11}} = \delta_{g1} a_1 = -0.006526 \cdot 0.501242 = -0.003271$$
$$\frac{dL}{dv_{12}} = \delta_{g1} a_2 = -0.006526 \cdot 0.509986 = -0.003328$$
$$\frac{dL}{dv_{13}} = \delta_{g1} a_3 = -0.006526 \cdot 0.517444 = -0.003377$$
$$\frac{dL}{dc_1} = \delta_{g1} = -0.006526$$

$$\frac{dL}{dv_{21}} = \delta_{g2} a_1 = 0.005833 \cdot 0.501242 = 0.002924$$
$$\frac{dL}{dv_{22}} = \delta_{g2} a_2 = 0.005833 \cdot 0.509986 = 0.002975$$
$$\frac{dL}{dv_{23}} = \delta_{g2} a_3 = 0.005833 \cdot 0.517444 = 0.003018$$
$$\frac{dL}{dc_2} = \delta_{g2} = 0.005833$$

$$\frac{dL}{dw_{11}} = \delta_{a1} x_1 = -0.000236 \cdot 0.900000 = -0.000213$$
$$\frac{dL}{dw_{12}} = \delta_{a1} x_2 = -0.000236 \cdot 0.600000 = -0.000142$$
$$\frac{dL}{dw_{13}} = \delta_{a1} x_3 = -0.000236 \cdot 0.350000 = -8.270554e-05$$
$$\frac{dL}{db_1} = \delta_{a1} = -0.000236$$

$$\frac{dL}{dw_{21}} = \delta_{a2} x_1 = 0.000308 \cdot 0.900000 = 0.000278$$
$$\frac{dL}{dw_{22}} = \delta_{a2} x_2 = 0.000308 \cdot 0.600000 = 0.000185$$
$$\frac{dL}{dw_{23}} = \delta_{a2} x_3 = 0.000308 \cdot 0.350000 = 0.000108$$
$$\frac{dL}{db_2} = \delta_{a2} = 0.000308$$

$$\frac{dL}{dw_{31}} = \delta_{a3} x_1 = 6.392714e-05 \cdot 0.900000 = 5.753443e-05$$
$$\frac{dL}{dw_{32}} = \delta_{a3} x_2 = 6.392714e-05 \cdot 0.600000 = 3.835629e-05$$
$$\frac{dL}{dw_{33}} = \delta_{a3} x_3 = 6.392714e-05 \cdot 0.350000 = 2.237450e-05$$
$$\frac{dL}{db_3} = \delta_{a3} = 6.392714e-05$$

Update (SGD)

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta} \qquad (\eta = 0.5)$$

Parametri dopo l'update (fine iterazione):
Strato 1 (w, b):
$w_{11} = 0.100085,\ w_{12} = -0.199927,\ w_{13} = 0.100056,\ b_1 = 9.936382e-05$
$w_{21} = -0.100145,\ w_{22} = 0.099883,\ w_{23} = 0.199907,\ b_2 = -0.000171$
$w_{31} = 0.049903,\ w_{32} = 0.099930,\ w_{33} = -0.100058,\ b_3 = -0.000120$

Strato 2 (v, c):
$v_{11} = 0.100685,\ v_{12} = -0.099261,\ v_{13} = 0.050718,\ c_1 = 0.001386$
$v_{21} = -0.052701,\ v_{22} = 0.097198,\ v_{23} = 0.097238,\ c_2 = -0.005381$

Output (u, d):
$u_1 = 0.167919,\ u_2 = -0.029874,\ d = 0.134843$

Calcolo esplicito degli aggiornamenti (Iterazione 3)

Regola teorica di update (SGD su 1 esempio).

$$\theta^{(3)} = \theta^{(2)} - \eta \left.\frac{\partial L}{\partial \theta}\right|_{\theta^{(2)},\,\text{esempio } \#3}$$

dove $\eta = 0.5$ e il gradiente è quello calcolato nella backprop dell'iterazione 3.

Strato 1.

$$w_{11}^{(3)} = w_{11}^{(2)} - \eta \left.\frac{\partial L}{\partial w_{11}}\right|_{\theta^{(2)},\,\#3} = 0.099979 - 0.5 \cdot (-0.000213) = 0.1000855 \approx 0.100085$$

(. . . e così via per tutti gli altri parametri, usando $\theta^{(2)}$ come "valore prima" e i gradienti dell'iterazione 3.)

7. Parametri finali dopo 3 iterazioni

Strato 1 (w, b):
$w_{11} = 0.100085,\ w_{12} = -0.199927,\ w_{13} = 0.100056,\ b_1 = 9.936382e-05$
$w_{21} = -0.100145,\ w_{22} = 0.099883,\ w_{23} = 0.199907,\ b_2 = -0.000171$
$w_{31} = 0.049903,\ w_{32} = 0.099930,\ w_{33} = -0.100058,\ b_3 = -0.000120$

Strato 2 (v, c):
$v_{11} = 0.100685,\ v_{12} = -0.099261,\ v_{13} = 0.050718,\ c_1 = 0.001386$
$v_{21} = -0.052701,\ v_{22} = 0.097198,\ v_{23} = 0.097238,\ c_2 = -0.005381$

Output (u, d):
$u_1 = 0.167919,\ u_2 = -0.029874,\ d = 0.134843$

Perché la loss può aumentare in SGD

Contesto

Consideriamo la rete neurale 3→3→2→1 definita nelle sezioni precedenti, addestrata con **Stochastic Gradient Descent** (SGD) su esempi singoli.

Loss sui primi 3 esempi

	Esempio #1 ($y=0$)	Esempio #2 ($y=1$)	Esempio #3 ($y=1$)
Prima di update	0.249686	0.250332	0.250329
Dopo update #1	0.204396	0.300210	0.300210
Dopo update #2	0.253822	0.246224	0.246222

Interpretazione

- Dopo l'update sull'esempio #1 la loss su #1 diminuisce, ma quella su #2 e #3 può aumentare. Questo accade perché l'update è ottimizzato *solo* per l'esempio corrente.
- In SGD non è garantita una diminuzione monotona della loss per singolo esempio.
- È quindi corretto monitorare la **loss media** su più esempi o su un mini-batch.

Come leggere la tabella delle loss (training vs valutazione)

È importante distinguere tra:

- **loss di training (SGD)**: è la loss usata per calcolare i gradienti e aggiornare i pesi nell'iterazione corrente. Alla iterazione k (SGD "vero") si usa *solo* l'esempio corrente k:

$$L^{(k)}(\theta_k) = \big(\hat{y}^{(k)}(\theta_k) - y^{(k)}\big)^2, \qquad \theta_{k+1} = \theta_k - \eta\, \nabla_\theta L^{(k)}(\theta_k).$$

  Questa $L^{(k)}$ è l'unica loss che influenza l'update dei pesi in quell'iterazione.
- **loss di valutazione (monitoraggio)**: dopo aver ottenuto un certo set di pesi θ (ad esempio "prima di update", "dopo update #1", "dopo update #2"), si può calcolare *a posteriori* la loss su più esempi per capire come l'update ha influenzato la rete. In questo caso i pesi θ sono *fissi* e non si fa backprop su questi esempi: si fa solo forward.

Cosa rappresenta ogni cella della tabella. Ogni cella è una loss *per-esempio* calcolata con un set di pesi fissato:

$$L_i(\theta) = \big(\hat{y}_i(\theta) - y_i\big)^2,$$

dove $i \in \{1,2,3\}$ indica l'esempio (colonna della tabella) e θ indica lo stato dei pesi (riga della tabella).

Significato delle righe.

- **Prima di update**: loss calcolata sui tre esempi usando i pesi iniziali θ_0.
- **Dopo update #1**: loss calcolata sui tre esempi usando i pesi θ_1 ottenuti *dopo* aver fatto SGD sull'esempio #1.
- **Dopo update #2**: loss calcolata sui tre esempi usando i pesi θ_2 ottenuti *dopo* aver fatto SGD sull'esempio #2.

Perché la loss può aumentare tra iterazioni. In SGD non è garantito che la loss su *tutti* gli esempi diminuisca ad ogni update: l'update è ottimizzato per ridurre $L^{(k)}$ dell'esempio corrente, e può temporaneamente aumentare L_i su altri esempi. Per questo, in pratica, si monitora spesso la **loss media** su un insieme di esempi (o su un mini-batch), non la loss di un singolo esempio.