Aprendizagem 2024/25

Homework III – Group 041

(ist106322, ist106157)

I. Pen-and-paper

Homework Nº 3:

I. Pen-and-Paper

$\phi(y_1, y_2) = y_1 \times y_2$

1. OLS $\quad f(x) = z = w_0 + w_1 y_1 + w_2 y_2$

$W_0 = 1$ nesta cadeira de aprendizagem

$$W = (X^T X)^{-1} X^T z$$

$$X = \begin{array}{c} \text{bias} \\ \downarrow \end{array} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 3 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \\ 1 & 2 & 4 \end{pmatrix} \qquad z = \begin{pmatrix} 1,25 \\ 7,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{pmatrix}$$

norma dos pesos: $\|\vec{w}\| = \sqrt{w_1^2 + w_2^2}$

$$\phi = \begin{pmatrix} 1 & 1\times1 \\ 1 & 1\times3 \\ 1 & 3\times2 \\ 1 & 3\times3 \\ 1 & 2\times4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{pmatrix}$$

Então ao assumir uma nova parametrização: $W = (\phi^T \phi)^{-1} \phi^T z$

$$W = \left(\underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix}}_{2\times5} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{pmatrix}}_{5\times2} \right)^{-1} \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix}}_{2\times5} \underbrace{\begin{pmatrix} 1,25 \\ 7,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{pmatrix}}_{5\times1}$$

$$\overset{\text{calc.}}{=} \underbrace{\begin{pmatrix} 5 & 27 \\ 27 & 191 \end{pmatrix}^{-1}}_{2\times2} \underbrace{\begin{pmatrix} 19,65 \\ 117,25 \end{pmatrix}}_{2\times1} \overset{\text{calc.}}{\simeq} \underbrace{\begin{pmatrix} 3,31593 \rightarrow w_0 \\ 0,11372 \rightarrow w_1 \end{pmatrix}}_{2\times1}$$

$$f(x_1, x_2) = w_0 + w_1\, \phi(x_1, x_2) = w_0 + w_1 x_1 x_2 = 3,31593 + 0,11372\, x_1 x_2$$

norma dos pesos: $\|\vec{w}\| = \sqrt{3,31593^2 + 0,11372^2} \simeq 3,31788$

2. Ridge regression

penalty factor = $\lambda = 1$

$$W = (\phi^T \phi + \lambda I)^{-1} \phi^T z \qquad \lambda = 1$$

norme dos pesos: $\|\vec{w}\| = \sqrt{w_1^2 + w_2^2}$

$$\lambda I = \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$W = \left(\underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix}}_{2\times5} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{pmatrix}}_{5\times2} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix} \begin{pmatrix} 1,25 \\ 7,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{pmatrix} \overset{calc.}{=}$$

$$= \left(\begin{pmatrix} 5 & 27 \\ 27 & 191 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix} \begin{pmatrix} 1,25 \\ 7,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{pmatrix} \overset{calc.}{=}$$

$$\overset{calc.}{=} \underbrace{\begin{pmatrix} 6 & 27 \\ 27 & 192 \end{pmatrix}^{-1}}_{2\times2} \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix}}_{2\times5} \begin{pmatrix} 1,25 \\ 7,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{pmatrix} \overset{calc.}{\simeq} \begin{pmatrix} 1,81809 \\ 0,32376 \end{pmatrix} \begin{array}{l} \to w_0 \\ \to w_1 \end{array}$$

$$f(x_1, x_2) = w_0 + w_1 \phi(x_1, x_2) = w_0 + w_1 x_1 x_2 = 1,81809 + 0,32376 x_1 x_2$$

norma dos pesos: $\|\vec{w}\| = \sqrt{1,81809^2 + 0,32376^2} = 1,84669$

Comparando a norma dos coeficientes obtidos com a regularização e sem, podemos observar como:

$$\|\vec{w}\|_{OLS} > \|\vec{w}\|_{Ridge}$$

A regressão *Rigde* teve como resultados menores coeficientes, quando comparado com o modelo *OLS*. Isto era expectável, tendo em conta que o modelo Ridge adiciona um termo de penalização (neste caso de 1), sendo este responsável pela diminuição dos maiores coeficientes. Adicionalmente, para variáveis que têm pouco ou nenhum efeito no resultado, o fator de penalização também reduz os seus coeficientes. Portanto a regularização mantém todas as variáveis e enfatiza as com maior importância e as menos significativas são desvalorizadas. Como consequência, haverá uma melhoria na capacidade de generalização e menor propensão a grandes variações nos outputs provocadas por pequenas mudanças nos inputs. Do modelo

OLS, por outro lado, ao não utilizar a regularização, já resultam coeficientes com um valor absoluto mais elevado, e pode ajustar-se excessivamente aos dados de treino, especialmente quando há multicolinearidade, isto é, quando duas ou mais variáveis estão fortemente correlacionadas. Em tais situações, o modelo *OLS* tende a atribuir coeficientes muito grandes para tentar ajustar pequenas variações entre estas variáveis, o que pode levar a *overfitting*. Este modelo corre o risco de capturar as "*noisy features*" durante o treino, e pode ficar demasiado complexo para fazer uma generalização equilibrada. Como tal, podemos concluir como a regularização procura simplificar o modelo e melhorar a sua generalização como também prevenir o *overfitting*.

3.

D	y_1	y_2	y_{num}
x_6	2	2	0,7
x_7	1	2	1,1
x_8	5	1	2,2

compare the train and test RMSE of the 2 models in 1) and 2)

Em teste,

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{m} \sum (y_{train} - y'_{train})^2}$$

$x_{new}' = f(x_{new}) =$
$= w_0 + w_1 x_{new1} + \dots +$
$+ w_m x_{newm}$

Modelo 1): $\boxed{OLS}$

$x_6' = f(x_6) = f(2,2) = 3{,}31593 + 0{,}11372 \times 2 \times 2 = 3{,}77081$

$x_7' = f(x_7) = f(1,2) = 3{,}31593 + 0{,}11372 \times 1 \times 2 = 3{,}54337$

$x_8' = f(x_8) = f(5,1) = 3{,}31593 + 0{,}11372 \times 5 \times 1 = 3{,}88453$

$x_1' = f(x_1) = f(1,1) = 3{,}31593 + 0{,}11372 \times 1 \times 1 = 3{,}42965$

$x_2' = f(x_2) = f(1,3) = 3{,}31593 + 0{,}11372 \times 1 \times 3 = 3{,}65709$

$x_3' = f(x_3) = f(3,2) = 3{,}31593 + 0{,}11372 \times 3 \times 2 = 3{,}99825$

$x_4' = f(x_4) = f(3,3) = 3{,}31593 + 0{,}11372 \times 3 \times 3 = 4{,}33941$

$x_5' = f(x_5) = f(2,4) = 3{,}31593 + 0{,}11372 \times 2 \times 4 = 4{,}22569$

	y_1	y_2	y_{num}	y'_{num}	$(y_{num} - y'_{num})$
x_1	1	1	1,25	3,42965	−2,17965
x_2	1	3	7,0	3,65709	3,34291
x_3	3	2	2,7	3,99825	−1,29825
x_4	3	3	3,2	4,33941	−1,13941
x_5	2	4	5,5	4,22569	1,27431
x_6	2	2	0,7	3,77081	−3,07081
x_7	1	2	1,1	3,54337	−2,44337
x_8	5	1	2,2	3,88453	−1,68453

$$RMSE_{train} = \sqrt{MSE_{train}} = \sqrt{\frac{1}{5}\left((-2{,}17965)^2 + (3{,}34291)^2 + (-1{,}29825)^2 + (-1{,}13941)^2 + (1{,}27431)^2\right)} = \sqrt{4{,}106699115} \simeq 2{,}02650$$

$$RMSE_{test} = \sqrt{MSE_{test}} = \sqrt{\frac{1}{3}\left((-3{,}07081)^2 + (-2{,}44337)^2 + (-1{,}68453)^2\right)}$$

$$= \sqrt{6{,}079100778} \simeq 2{,}46560$$

Modelo 2: $\boxed{\text{Ridge regression}}$

$$x_1' = f(x_1) = f(1,1) = 1{,}81809 + 0{,}32376 \times 1 \times 1 = 2{,}14185$$
$$x_2' = f(x_2) = f(1,3) = 1{,}81809 + 0{,}32376 \times 1 \times 3 = 2{,}78937$$
$$x_3' = f(x_3) = f(3,2) = 1{,}81809 + 0{,}32376 \times 3 \times 2 = 3{,}76065$$
$$x_4' = f(x_4) = f(3,3) = 1{,}81809 + 0{,}32376 \times 3 \times 3 = 4{,}73193$$
$$x_5' = f(x_5) = f(2,4) = 1{,}81809 + 0{,}32376 \times 2 \times 4 = 4{,}40817$$
$$x_6' = f(x_6) = f(2,2) = 1{,}81809 + 0{,}32376 \times 2 \times 2 = 3{,}11313$$
$$x_7' = f(x_7) = f(1,2) = 1{,}81809 + 0{,}32376 \times 1 \times 2 = 2{,}46561$$
$$x_8' = f(x_8) = f(5,1) = 1{,}81809 + 0{,}32376 \times 5 \times 1 = 3{,}43689$$

	y_1	y_2	y_{num}	y'_{num}	$(y_{num} - y'_{num})$
x_1	1	1	1,25	2,14185	−0,89185
x_2	1	3	7,0	2,78937	4,21063
x_3	3	2	2,7	3,76065	−1,06065
x_4	3	3	3,2	4,73193	−1,53193
x_5	2	4	5,5	4,40817	1,09183
x_6	2	2	0,7	3,11313	−2,41313
x_7	1	2	1,1	2,46561	−1,36561
x_8	5	1	2,2	3,43689	−1,23689

$$RMSE_{train} = \sqrt{MSE_{train}} = \sqrt{\frac{1}{5}\left((-0,89185)^2 + (4,21063)^2 + (-1,06065)^2 + (-1,53193)^2 + (1,09183)^2\right)} = \sqrt{4,637736423} \simeq 2,15354$$

$$RMSE_{test} = \sqrt{MSE_{test}} = \sqrt{\frac{1}{3}\left((-2,41313)^2 + (-1,36561)^2 + (-1,23689)^2\right)}$$

$$= \sqrt{3,072661314} \simeq 1,75290$$

Comparando os resultados:

Após determinar os valores de *RMSE* para cada modelo, podemos observar que o erro de treino para o modelo que utiliza *OLS* é inferior ao erro de teste respetivo, enquanto para o modelo de regressão *Ridge* se verifica o contrário.

O facto de que o erro para o conjunto de treino do modelo *OLS* é inferior ao erro para o conjunto de teste, indica que o modelo faz overfitting aos dados de treino, isto é, adequa-se aos dados conhecidos, mas não é tão capaz de estimar resultados para observações novas. Efetivamente, esta conclusão está de acordo com o esperado para o modelo *OLS*, já que é sabido que este modelo tem tendência a apresentar *overfitting*, sendo bastante flexível e não restringido pela regularização.

Por outro lado, para o modelo *Ridge*, constatamos que o erro para o conjunto de treino é superior ao erro para o conjunto de teste. Neste caso, a regularização reduz o problema de overfitting do *OLS*, sendo que a penalização introduzida torna o modelo menos complexo, permitindo uma melhor generalização e capacidade de avaliar novas observações. Note-se ainda que apesar de o erro de treino da regressão *Ridge* ser ligeiramente superior ao de treino do *OLS*, o erro de teste é significativamente inferior, o que corrobora a expectativa de que o *Ridge* apresente uma melhor capacidade de generalização. Mais uma vez, estes resultados concordam com o expectável, visto que a regularização é uma forma de reduzir o risco de *overfitting*, forçando a diminuição dos coeficientes com valores maiores e tornando o modelo mais simples e com menor variância.

Aprendizagem 2024/25

Homework III – Group 041

(ist106322, ist106157)

4. MLP → Multilayer Perceptron

$\eta = 0,1$

one stochastic gradient descent update → the observation is x_1

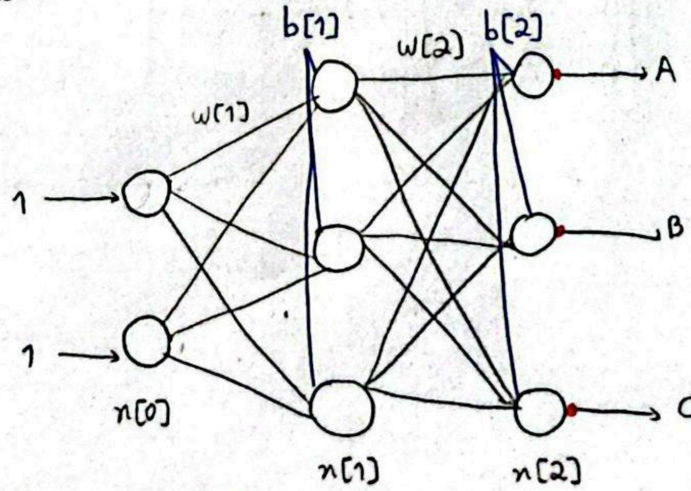
⊕1 de output ↓ classificação

$$W[1] = \underbrace{\begin{bmatrix} 0,1 & 0,1 \\ 0,1 & 0,2 \\ 0,2 & 0,1 \end{bmatrix}}_{3\times 2} \qquad b[1] = \underbrace{\begin{bmatrix} 0,1 \\ 0 \\ 0,1 \end{bmatrix}}_{3\times 1} \qquad W[2] = \underbrace{\begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}}_{3\times 3} \qquad b[2] = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{3\times 1}$$

3×2: 2ª camada, 1ª camada

3×3: 3ª camada, 2ª camada

Rede:



Função de ativação

$$\text{softmax}(z_c^{out}) = \frac{e^{z_c^{out}}}{\sum_{l=1}^{|C|} e^{z_l^{out}}}$$

Esquema:

$$x[0] \rightarrow z[1] \rightarrow x[1] \rightarrow z[2] \rightarrow x[2] \rightarrow E$$

$w[1] \rightarrow z[1]$, $b[1] \rightarrow z[1]$, $w[2] \rightarrow z[2]$, $b[2] \rightarrow z[2]$

$\delta[1]$

$\delta[2]$

1) Propagação:

$$\boxed{z[i] = w[i]\,x[i-1] + b[i] \quad , \quad x[i] = f(z[i])}$$

$$z[1] = w[1]\,x[0] + b[1] = \underbrace{\begin{bmatrix} 0,1 & 0,1 \\ 0,1 & 0,2 \\ 0,2 & 0,1 \end{bmatrix}}_{3\times 2} \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{2\times 1} + \underbrace{\begin{bmatrix} 0,1 \\ 0 \\ 0,1 \end{bmatrix}}_{3\times 1} = \begin{bmatrix} 0,2 \\ 0,3 \\ 0,3 \end{bmatrix} + \begin{bmatrix} 0,1 \\ 0 \\ 0,1 \end{bmatrix}$$

$$= \begin{bmatrix} 0,3 \\ 0,3 \\ 0,4 \end{bmatrix}$$

$x[1] = z[1]$ (pois não há função de ativação nos "hidden layers".

$$z[2] = w[2]\,x[1] + b[2] = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0,3 \\ 0,3 \\ 0,4 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\text{cálc.} \approx \begin{bmatrix} 2,7 \\ 2,3 \\ 2 \end{bmatrix}$$

$$x[2] = softmax(z[2]) = \frac{e^{z[2]}}{\sum_{l=1}^{|C|} e^{z[2]_l}} = \begin{bmatrix} softmax(2,7) \\ softmax(2,3) \\ softmax(2) \end{bmatrix} =$$

$$= \begin{bmatrix} e^{2,7} / (e^{2,7} + e^{2,3} + e^{2}) \\ e^{2,3} / (e^{2,7} + e^{2,3} + e^{2}) \\ e^{2} / (e^{2,7} + e^{2,3} + e^{2}) \end{bmatrix}$$

$$\overset{calc}{\simeq} \begin{bmatrix} 0,4614876234 \\ 0,3093444405 \\ 0,2291679717 \end{bmatrix}$$

2) Propagação para trás : 1º calcular as derivadas; 2º deltas; 3º regra

Derivadas: $\frac{\partial z[i]}{\partial W[i]} = x[i-1]$ $\quad \frac{\partial z[i]}{\partial x[i-1]} = W[i]$

$\frac{\partial z[i]}{\partial b[i]} = 1$ $\quad \frac{\partial x[i]}{\partial z[i]} = ?$

Na função softmax, todas as saídas estão interconectadas por meio do denominador que é a soma das exponenciais de todas as entradas. Alterar uma entrada afeta tanto a própria saída (caso $i = j$) quanto as saídas das outras entradas (caso $i \neq j$)

Calcula a derivada parcial de softmax(z_i) em relação a z_j, ou seja,

$$\frac{\partial \text{softmax}(z_i)}{\partial z_j}$$

Caso $i = j$:

$$\frac{\partial \text{softmax}(z_i)}{\partial z_i} = \frac{(e^{z_i})' \cdot \left(\sum_{l=1}^{m} e^{z_l}\right) - e^{z_i} \cdot \left(\sum_{l=1}^{m} e^{z_l}\right)'}{\left(\sum_{l=1}^{m} e^{z_l}\right)^2}$$

$$\overset{*}{=} \frac{e^{z_i}\left(\sum_{l=1}^{m} e^{z_l}\right) - e^{z_i} \cdot e^{z_i}}{\left(\sum_{l=1}^{m} e^{z_l}\right)^2} = \frac{e^{z_i}\left(\sum_{l=1}^{m} e^{z_l} - e^{z_i}\right)}{\left(\sum_{l=1}^{m} e^{z_l}\right)^2}$$

$$= \frac{e^{z_i}}{\left(\sum_{l=1}^{m} e^{z_l}\right)^2} \cdot \left(\frac{\sum_{l=1}^{m} e^{z_l}}{\sum_{l=1}^{m} e^{z_l}} - \frac{e^{z_i}}{\left(\sum_{l=1}^{m} e^{z_l}\right)}\right) =$$

$$= \text{softmax}(z_i) \cdot (1 - \text{softmax}(z_i))$$

* $\left(\sum_{l=1}^{m} e^{z_l}\right)'_{z_i} = e^{z_i}$ pois apenas o termo correspondente a z_i contribui para a derivada de $\sum_{l=1}^{m} e^{z_l}$, os outros termos que não dependem de z_i, têm derivada $= 0$.

Caso $i \neq j$:

$$\frac{\partial \text{softmax}(z_i)}{\partial z_j} = \frac{\cancel{(e^{z_i})'}^{\,0} \cdot \left(\sum_{l=1}^{m} e^{z_l}\right) - e^{z_i} \cdot \left(\sum_{l=1}^{m} e^{z_l}\right)'}{\left(\sum_{l=1}^{m} e^{z_l}\right)^2} =$$

$$= \frac{-e^{z_i} e^{z_j}}{\left(\sum_{l=1}^{m} e^{z_l}\right)^2} = -\text{softmax}(z_i) \cdot (\text{softmax}(z_j)$$

Homework III – Group 041

(ist106322, ist106157)

Então:

$$\frac{\partial x[i]}{\partial z[j]} = \begin{cases} \text{softmax}(z[i]).(1-\text{softmax}(z[i])) & \text{se } i=j \\ -\text{softmax}(z[i]).\text{softmax}(z[j]) & \text{se } i\neq j \end{cases}$$

então

$$\frac{\partial x[i]}{\partial z[i]} = \text{softmax}(z[i]).(1-\text{softmax}(z[i]))$$

para $i=2$

se $i \neq 2$ não há função de ativação e $x[i]=z[i]$, logo

$$\frac{\partial x[i]}{\partial z[i]} = \frac{\partial z[i]}{\partial z[i]} = 1$$

nr samples, nr classes, true label for sample i, class l

Erro: $$-\sum_{i=1}^{N}\sum_{l=1}^{|C|} t_l^{(i)} \log\left(X_l^{[out](i)}\right) = E$$

ii)

resultou de derivada softmax com entropia cruzada

$$t = \text{target} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\delta[2] = \frac{\partial E}{\partial x[2]} \circ \frac{\partial x[2]}{\partial z[2]} = x[2] - t =$$

$$= \begin{bmatrix} 0,4614876234 \\ 0,3093444049 \\ 0,2291679716 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0,4614876234 \\ -0,6906555951 \\ 0,2291679716 \end{bmatrix}$$

Hadamard

$$\delta[1] = \left(\left(\frac{\partial z[2]}{\partial x[1]}\right)^T \cdot \delta[2] \right) \circ \frac{\partial x[1]}{\partial z[1]} = \left(W[2]^T \cdot \delta[2]\right) \circ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$= \left(\begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}^T \cdot \delta[2] \right) \circ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \left(\begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0,4614876234 \\ -0,6906555951 \\ 0,2291679716 \end{pmatrix} \right)$$

$$\circ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \times 10^{-10} \\ -0,2291679718 \\ 0,4614876233 \end{pmatrix} \circ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \times 10^{-10} \\ -0,2291679718 \\ 0,4614876233 \end{pmatrix}$$

Atualizar:

$$b[2]^{new} = b[2]^{old} - \eta \frac{\partial E}{\partial b[2]} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0,1 \left(\delta[2] \cdot \cancel{\frac{\partial z[2]}{\partial b[2]}}^{\,1\,T} \right)$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0,1 \begin{pmatrix} 0,4614876234 \\ -0,6906555951 \\ 0,2291679716 \end{pmatrix} = \begin{pmatrix} 0,95385 \\ 1,06907 \\ 0,97708 \end{pmatrix}$$

Homework III – Group 041

(ist106322, ist106157)

$$W[2]^{new} = W[2]^{old} - \eta \frac{\partial E}{\partial W[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0{,}1 \left(\delta[2] \cdot \overset{x[1]}{\left(\frac{\partial z[2]}{\partial W[2]}\right)^T} \right) = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0{,}1 \left(\underset{3\times 1}{\begin{bmatrix} 0{,}4614876234 \\ -0{,}6906555951 \\ 0{,}2291679716 \end{bmatrix}} \cdot \underset{1\times 3}{\begin{bmatrix} 0{,}3 & 0{,}3 & 0{,}4 \end{bmatrix}} \right) =$$

$$= \begin{bmatrix} 0{,}98616 & 1{,}98616 & 1{,}98154 \\ 1{,}02072 & 2{,}02072 & 1{,}02762 \\ 0{,}99312 & 0{,}99312 & 0{,}99083 \end{bmatrix}$$

$$b[1]^{new} = b[1]^{old} - \eta \frac{\partial E}{\partial b[1]} = \begin{bmatrix} 0{,}1 \\ 0 \\ 0{,}1 \end{bmatrix} - 0{,}1 \left(\delta[1] \cdot \left(\cancelto{1}{\frac{\partial z[1]}{\partial b[1]}}\right)^T \right) =$$

$$= \begin{bmatrix} 0{,}1 \\ 0 \\ 0{,}1 \end{bmatrix} - 0{,}1 \begin{pmatrix} -1\times 10^{-10} \\ -0{,}2291679718 \\ 0{,}4614876233 \end{pmatrix} = \begin{pmatrix} 0{,}10000 \\ 0{,}02292 \\ 0{,}05385 \end{pmatrix}$$

$$W[1]^{new} = W[1]^{old} - \eta \frac{\partial E}{\partial W[1]} = \begin{bmatrix} 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}1 \end{bmatrix} - 0{,}1 \left(\delta[1] \cdot \overset{x[0]}{\left(\frac{\partial z[1]}{\partial W[1]}\right)^T} \right) =$$

$$= \begin{bmatrix} 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}1 \end{bmatrix} - 0{,}1 \left(\begin{bmatrix} -1\times 10^{-10} \\ -0{,}2291679718 \\ 0{,}4614876233 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \end{bmatrix} \right) =$$

$$= \begin{pmatrix} 0{,}10000 & 0{,}10000 \\ 0{,}12292 & 0{,}22292 \\ 0{,}15385 & 0{,}05385 \end{pmatrix}$$