

## Lista de Exercícios 1

### ① Gradiente Descendente

✓(a) calcule o gradiende das seguintes funções:

- ✓ i.  $f(x) = x^2 + 2$
- ✓ ii.  $f(x) = (x - 2)^2$
- ✓ iii.  $f(x, y, z) = x^3 + y^2 + z$
- ✓ iv.  $g(z) = \frac{1}{1+e^{-z}}$

✓(b) faça 3 iterações do método de gradiente descendente para cada função utilizando  $x, y$  e  $z$  iguais a zero e  $\alpha = 0.5$ .

✓(c) implemente o código em Python que realize o gradiente ascendente e descentende de cada função apresentada.

✓(d) Explique a diferença do gradiente estocástico, mini-batch, batch? Qual é a vantagem e desvantagem de cada um? Qual deles é equivalente ao aprendizado on-line?

### ✓ 2. Regressão Linear:

✓(a) dado que

$$h_{\theta}(x^{(i)}) = \theta_k x_k^{(i)} \quad (1)$$

e

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad (2)$$

calcule  $\frac{\partial}{\partial \theta_j} J(\theta)$  e mostre a regra de atualização de  $\theta_j$ .

✓(b) A equação  $J(x^{(i)}, y^{(i)}) = \frac{1}{2}(h_{\theta}(x^{(i)}) - y^{(i)})^2$  representa o erro quadratico da regressão linear. Explique como chegar a esta fórmula.

✓(c) Considere o a Tabela 1 onde são apresentados área em metros quadrados e o preço casas.

1/2 → Adicionado afim de remover o quadrado que foi adicionado para remover o sinal

área	preço
50	120
60	150
100	250

Table 1: área e preço de casas

Utilizando regressão linear faça três iterações do gradiente descendente e apresente os pesos calculados.

✓(d) implemente o código que realize a regressão linear para o problema dos preços das casas.

### ✓ 3. Regressão Logística:

precisa mostrar pra qual lado foi passado os valores.

✓(a) A função logística é dado pela fórmula  $g(z) = \frac{1}{1+e^{-z}}$ . Desenhe a curva da função logística.  
 ✓(b) Dado um ponto e uma reta do regressor logístico responda:

- i. Qual é o valor do regressor logístico quando o ponto está em cima da reta?
- ii. Qual é o valor do regressor logístico quando o ponto está abaixo a reta?
- iii. Qual é o valor do regressor logístico quando o ponto está acima a reta?

### ④ Regra da Cadeia

✓(a) Aplique a regra da cadeia mostrando todas as derivadas parciais e grafo que representa as operações aritméticas das funções abaixo.

- ✓ i.  $f(x) = x^2 + 2$

- ii.  $f(x) = (x - 2)^2$   
 iii.  $f(x, y) = (2x + 3y)^2$   
 iv.  $f(x, y, z) = x^3 + y^2 + z$   
 v.  $g(z) = \frac{1}{1+e^{-z}}$
- (b) Considere  $x, y$  e  $z$  iguais a 1. Calcule o valor da saída de cada nó do grafo (forward pass) e depois calcule a derivada de  $\frac{\partial f(x)}{\partial x}$  utilizando o grafo (backward pass) construídos do exercício anterior.
5. Perceptron (1 camada)
- (a) Qual é a interpretação geométrica de um perceptron? Uma reta / hiperplano  
 (b) Mostre a equivalência do perceptron e o algoritmo de regressão logística.  
 (c) Dado o problema do AND fazer duas iterações do algoritmo de perceptron.  
 (d) Explique porquê XOR não pode ser aprendido por um perceptron.  
 (e) Explique a função do bias na equação. Dê exemplos onde não é possível realizar o aprendizado sem o bias (Exemplo 18 anos)

- desloca a ativação →
6. Multi Layer Perceptron (MLP)
- (a) Explique porquê uma MLP pode ser considerado um aproximador universal de funções.  
 (b) Realize o aprendizado do XOR utilizando uma MLP com uma topologia 2x1 (dois neurônios na camada intermediária e um neurônio na camada de saída). Usem a função degrau, assuma que os pesos iniciais são iguais a zero e faça 1 iteração de treino.

## 7 Convolução

- (a) Considere o filtro de convolução abaixo. Qual é o efeito deste filtro sobre uma imagem.

```
kernel =
[[1, 0, -1],
 [1, 0, -1],
 [1, 0, -1]]
```

- (b) Considere o Kernel 3x3x3 com stride 1 e padding 0. Faça a convolução da imagem 4x4x3 utilizando este kernel.

```
Kernel =
array([[[ 0,  1,  2],
       [ 3,  4,  5],
       [ 6,  7,  8]],

      [[ 9, 10, 11],
       [12, 13, 14],
       [15, 16, 17]],

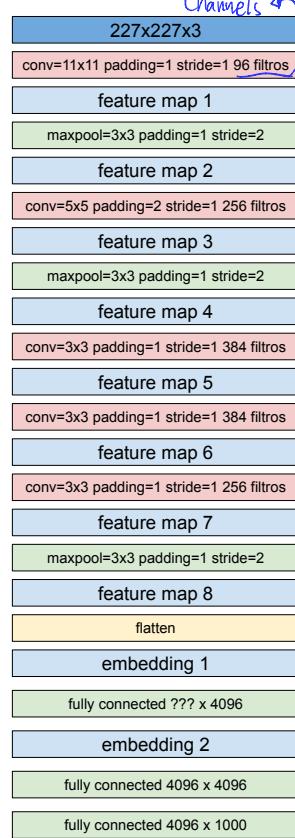
      [[18, 19, 20],
       [21, 22, 23],
       [24, 25, 26]]])

Imagem =
array([[[0, 1, 0, 0],
       [1, 1, 1, 1],
       [0, 1, 1, 1],
       [1, 0, 0, 1]],

      [[0, 0, 0, 1],
       [0, 1, 0, 1],
       [1, 0, 0, 0],
       [1, 1, 1, 1]],

      [[1, 1, 0, 0],
       [0, 1, 1, 0],
       [0, 1, 1, 1],
       [1, 0, 1, 0]]])
```

Figure 1: Rede Neural 1



✓ (c) Considere a topologia da rede neural ilustrada na Figura 1.

- ✓ i. Calcule as dimensões das feature maps.
- ✓ ii. Calcule as dimensões dos embeddings
- ✓ iii. Onde são os locais que deveriam ser adicionados as funções de ativação?
- ✓ iv. Qual é a dimensão de entrada da fully connected entre o embedding1 e embedding 2? → Após as camadas de convolução e as fully connected
- ✓ v. Calcule o número de parâmetros para todas as operações que possuam parâmetros.
- ✓ vi. Qual é o número total de parâmetros desta rede neural?
- ✓ vii. Escreva o código em pytorch desta rede neural.

①

a)

$$i. f(x) = x^2 + 2$$

$$f'(x) = 2x \Big|_1$$

$$ii. f(x) = (x-2)^2$$

$$\begin{aligned} f(x) &= (x-2)(x-2) & f'(x) &= 2x - 4 \Big|_1 \\ &= x^2 - 2x - 2x + 4 & & \\ &= x^2 - 4x + 4 & & \end{aligned}$$

$$iii. f(x, y, z) = x^3 + y^2 + z$$

$$\frac{\partial f(x, y, z)}{\partial x} = 3x^2 \quad \frac{\partial f(x, y, z)}{\partial y} = 2y \quad \frac{\partial f(x, y, z)}{\partial z} = 1$$

$$iv. g(z) = \frac{1}{1+e^{-z}}$$

$$g(z) = (1+e^{-z})^{-1}$$

$$\begin{aligned} f(x) &= e^x \\ \frac{d f(x)}{dx} &= e^x \end{aligned}$$

$$\begin{aligned} f(x) &= e^{-x} \\ \frac{d f(x)}{dx} &= -e^{-x} \end{aligned}$$

Regra do quociente

$$\left[ \frac{f(x)}{p(x)} \right]' = \frac{f'(x)p(x) - f(x)p'(x)}{p(x)^2}$$

logo temos ...

$$g(z) = \frac{1}{1+e^{-z}} \quad \begin{array}{l} \nearrow f(z) \\ \searrow p(z) \end{array}$$

$$f(z) = 1 \quad f'(z) = 0$$

$$p(z) = 1+e^{-z} \quad p'(z) = -e^{-z}$$

$$g'(z) = \frac{0 \cdot p'(z) - (1 \cdot (-e^{-z}))}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} \Big|_1$$

Se olharmos melhor temos:

$$\begin{aligned} g'(z) &= \frac{e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})(1+e^z)} = \boxed{\frac{1}{1+e^{-z}}} \cdot \frac{e^{-z}}{(1+e^{-z})} = g(z) \cdot \frac{e^{-z}}{1+e^{-z}} \\ &\quad \begin{array}{l} \nearrow g(z) \\ \searrow g(z) \end{array} \end{aligned}$$

$$= g(z) \cdot \frac{1+e^{-z}-1}{1+e^{-z}} = g(z) \cdot \left( \frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} \right) = g(z) \cdot \left( 1 - \boxed{\frac{1}{1+e^{-z}}} \right)$$

$$\text{logo... } g'(z) = g(z) \cdot (1 - g(z)) \Big|_1$$

$$\textcircled{1} \quad b) \quad i. \quad k = k - \alpha \cdot f'(k) \quad f'(k) = 2k \quad \alpha = 0,5$$

$$\textcircled{1} \quad k=0$$

$$k = 0 - 0,5 \cdot 2 \cdot 0$$

$$= 0 - 0,5 \cdot 0$$

$$= \underline{\underline{0}}_n$$

$$\textcircled{2} \quad k=0$$

$$k = 0 - 0,5 \cdot 2 \cdot 0$$

$$= 0 - 0,5 \cdot 0$$

$$= \underline{\underline{0}}_n$$

$$\textcircled{3} \quad k=0$$

$$k = 0 - 0,5 \cdot 2 \cdot 0$$

$$= 0 - 0,5 \cdot 0$$

$$= \underline{\underline{0}}_n$$

$$ii. \quad k = k - \alpha \cdot f'(k) \quad f'(k) = 2k - 4 \quad \alpha = 0,5$$

$$\textcircled{1} \quad k=0$$

$$k = 0 - 0,5(2 \cdot 0 - 4)$$

$$= 0 - 0,5(-4)$$

$$= \underline{\underline{2}}_n$$

$$\textcircled{2} \quad k=2$$

$$k = 2 - 0,5(2 \cdot 2 - 4)$$

$$= 2 - 0,5(0)$$

$$= 2 - 0 = \underline{\underline{2}}_n$$

$$\textcircled{3} \quad k=2$$

$$k = 2 - 0,5(2 \cdot 2 - 4)$$

$$= 2 - 0,5(0)$$

$$= 2 - 0 = \underline{\underline{2}}_n$$

$$iii. \quad k = k - \alpha \frac{\partial f(k,y,z)}{\partial k} \quad y = y - \alpha \frac{\partial f(k,y,z)}{\partial y} \quad z = z - \alpha \frac{\partial f(k,y,z)}{\partial z}$$

$$k = k - \alpha(3k)$$

$$y = y - \alpha(2y)$$

$$z = z - \alpha \cdot 1$$

$$\alpha = 0,5$$

$$x - \textcircled{1} \quad k=0$$

$$k = 0 - 0,5 \cdot 3 \cdot 0^2$$

$$= \underline{\underline{0}}_n$$

$$\textcircled{2} \quad k=0$$

$$k = 0 - 0,5 \cdot 3 \cdot 0^2$$

$$= \underline{\underline{0}}_n$$

$$\textcircled{3} \quad k=0$$

$$k = 0 - 0,5 \cdot 3 \cdot 0^2$$

$$= \underline{\underline{0}}_n$$

$$y - \textcircled{1} \quad y=0$$

$$y = 0 - 0,5 \cdot 20$$

$$= \underline{\underline{0}}_n$$

$$\textcircled{2} \quad y=0$$

$$y = 0 - 0,5 \cdot 20$$

$$= \underline{\underline{0}}_n$$

$$\textcircled{3} \quad y=0$$

$$y = 0 - 0,5 \cdot 20$$

$$= \underline{\underline{0}}_n$$

$$z - \textcircled{1} \quad z=0$$

$$z = 0 - 0,5 \cdot 1$$

$$= -0,5\underline{\underline{1}}_n$$

$$\textcircled{2} \quad z=-0,5$$

$$z = -0,5 - 0,5 \cdot 1$$

$$= -0,5 - 0,5$$

$$= -1\underline{\underline{1}}_n$$

$$\textcircled{3} \quad z=-1$$

$$z = -1 - 0,5 \cdot 1$$

$$= -1 - 0,5$$

$$= -1,5\underline{\underline{1}}_n$$

$$\textcircled{1} \quad b) \text{ iv. } z = z - \alpha g'(z) \quad g'(z) = g(z) \cdot (1 - g(z)) \quad \alpha = 0,5$$

$$\textcircled{1} \quad z=0$$

$$\begin{aligned} z &= 0 - 0,5 \left( \frac{1}{1+e^{-0}} \cdot \left( 1 - \frac{1}{1+e^{-0}} \right) \right) \\ &= 0 - 0,5 \left( \frac{1}{1+1} \cdot \left( 1 - \frac{1}{1+1} \right) \right) \\ &= -0,5 \left( \frac{1}{2} \cdot \left( 1 - \frac{1}{2} \right) \right) \\ &= -0,5 \left( \frac{1}{2} \cdot \frac{1}{2} \right) \\ &= -\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= -\frac{1}{8} = -0,125 \end{aligned}$$

$$\textcircled{2} \quad z = -0,125$$

$$\begin{aligned} z &= -0,125 - 0,5 \cdot g'(-0,125) \\ &= -0,125 - 0,5 \cdot 0,249 \\ &= -0,125 - 0,1245 \\ &= 0,2495 \end{aligned}$$

$$\textcircled{3} \quad z = -0,2495$$

$$\begin{aligned} z &= -0,2495 - 0,5 \cdot g'(-0,2495) \\ &= -0,2495 - 0,5 \cdot 0,246 \\ &= -0,2495 - 0,123 \\ &= -0,3725 \end{aligned}$$

c) Add link

- d) Estocástico: atualiza apenas um ponto do dataset por vez
- ↳ Melhor uso da memória
  - ↳ mais fácil de evitá-lo mínimos locais
  - ↳ Computacionalmente mais rápido
  - ↳ Fácil implementação
  - ↳ Convergência não é direta, o que pode resultar nas iterações terminarem antes de se atingir o valor desejado.
  - ↳ Sensível ao aumento de variáveis

Batches: atualiza todos os valores do dataset de uma só vez. Através do conceito de multiplicação de matrizes. Basicamente estamos calculando o vetor resultante do gradiente.

- ↳ Convergência direta
- ↳ consome muita memória
- ↳ Permite paralelizações
- ↳ Pode ser demorado
- ↳ Pode não ser viável p/ datasets muito grandes

Mini Batch: Meio termo entre estocástico e batches. É possível definir o tamanho do lote do dataset a ser carregado e processado.

- ↳ maior controle do consumo de memória
- ↳ diminui a frequência de atualizações do estocástico
- ↳ **Novo parâmetro que deve-se adotar (batch\_size)**

O aprendizado online diz respeito à capacidade de atualizações de um modelo. Isto é, a cada nova informação esse dado é inserido e considerado. Dessa forma o modelo estará sempre de acordo com as informações mais atuais. O gradiente estocástico possui uma natureza semelhante uma vez que cada dado do dataset é individualmente apresentado e atualizado.

②

$$a) h_{\theta}(x^{(i)}) = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)}$$

*equação da reta*

$$h_{\theta}(x) = \theta_0 x_0 + \theta_1 x_1$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{2} \sum_{i=1}^m (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} - y^{(i)})^2$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{2} (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} - y^{(i)}) x_j$$

$$\frac{\partial J(\theta)}{\partial \theta_k} = \frac{1}{2} (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} - y^{(i)}) x_k$$

$$\theta_j = \theta_j - \alpha \cdot \frac{1}{2} (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} - y^{(i)}) x_j$$

$$\theta_k = \theta_k - \alpha \cdot \frac{1}{2} (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} - y^{(i)}) x_k$$

► Outra interpretação (correta)

$$h_{\theta}(x^{(i)}) = \theta_0 x_0^{(i)}$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{2} \sum_{i=1}^m (\theta_0 x_0^{(i)} - y^{(i)})^2$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = (\theta_0 x_0^{(i)} - y^{(i)}) \cdot x_j$$

$$\theta_j = \theta_j - \alpha (\theta_0 x_0^{(i)} - y^{(i)}) x_j$$

② b) Uma maneira de calcularmos o erro entre o valor real e o valor predito pelo modelo é tirar a diferença entre os dois. Portanto, temos:

$$\text{erro} = \hat{y}_i - y_i$$

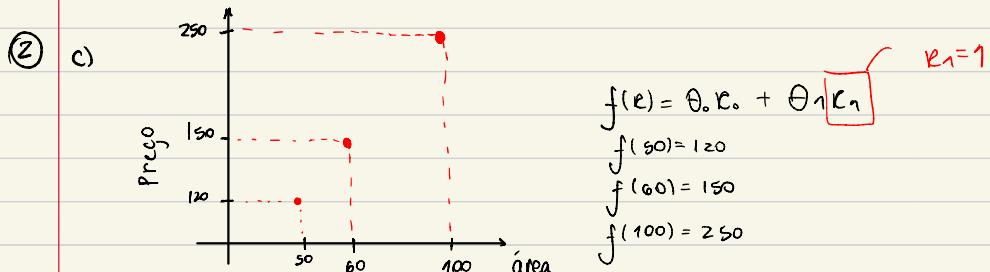
valor predito  $\rightarrow$  valor real

Entretanto, é necessário apenas o valor absoluto dessa subtração logo vamos elevar essa expressão ao quadrado.

$$\text{erro} = (\hat{y}_i - y_i)^2$$

Dessa forma já temos o erro quadrático, porém multiplicamos o termo por 1/2 com o fim de removermos a constante 2 que surgirá da derivada do erro logo temos:

$$\text{erro} = \frac{1}{2} (\hat{y}_i - y_i)^2$$



$$\textcircled{1} \quad \theta_0 = 0,1 \quad \alpha = 0,01 \quad x_0^{(1)} = 50 \quad y^{(1)} = 120$$

$$\textcircled{2} \quad \theta_0 = 57,6 \quad \alpha = 0,01 \quad x_0^{(1)} = 60 \quad y^{(1)} = 150$$

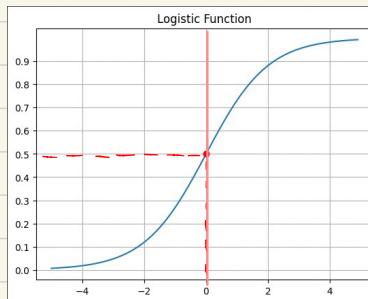
$$\begin{aligned}\theta_0 &= 0,1 - 0,01 (0,1 \cdot 50 - 120) \cdot 50 \\ &= 0,1 - 0,3 \cdot (-115) \\ &= 0,1 + 37,5 \\ &= 37,6\end{aligned}$$

$$\begin{aligned}\theta_0 &= 57,6 - 0,01 (57,6 \cdot 60 - 150) \cdot 60 \\ &= 57,6 - 0,6 \cdot 3306 \\ &= 57,6 - 1983,6 \\ &= -1926\end{aligned}$$

$$\textcircled{3} \quad \theta_0 = -1926 \quad \alpha = 0,01 \quad R_0^{(z)} = 100 \quad y^{(z)} = 250$$

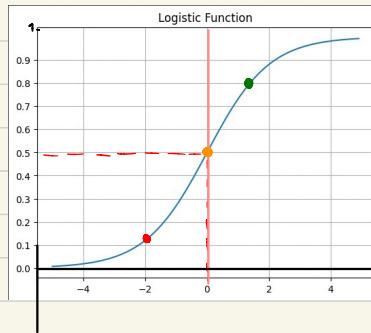
$$\begin{aligned}\theta_0 &= -1926 - 0,01(-1926 \cdot 100 - 250) \cdot 100 \\ &= -1926 - (-192850) \\ &= 190924\end{aligned}$$

\textcircled{3} a)



Se considerarmos a equação geral da reta com os termos isolados à esquerda:

b)



i) 0,5

ii) Tende a 0 ( ponto < 0,5 )

iii) Tende a 1 ( ponto > 0,5 )

## ④ Regra da cadeia

a)

$$i - f(x) = x^2 + 2$$

$$f_1 = x^2 \quad \frac{df_1}{dx} = 2x$$

$$f_2 = f_1 + 2 \quad \frac{df_2}{df_1} = 1$$

$$\frac{df(x)}{dx} = \frac{df_2}{df_1} \cdot \frac{df_1}{dx} = 2x \cdot 1,$$



$$ii - f(x) = (x-2)^2$$

$$f_1 = x-2 \quad \frac{df_1}{dx} = 1$$

$$\frac{df(x)}{dx} = \frac{df_2}{df_1} \cdot \frac{df_1}{dx} = 1 \cdot 2f_1$$

$$f_2 = f_1^2 \quad \frac{df_2}{df_1} = 2f_1 \quad x \rightarrow \textcircled{-2} \xrightarrow{\textcircled{f1}} \textcircled{12} \xrightarrow{\textcircled{f2}}$$

$$iii - f(x,y) = (2x+3y)^2$$

$$f_1 = 2x \quad \frac{df_1}{dx} = 2$$

$$f_2 = f_1 + 3y \quad \frac{\partial f_1}{\partial f_1} = 1 \quad \frac{\partial f_2}{\partial y} = 3$$

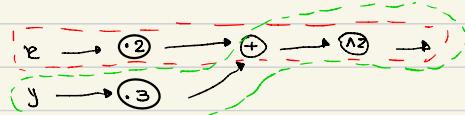
$$\frac{\partial f(x,y)}{\partial x} = \frac{\partial f_1}{\partial x} \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_2}{\partial x}$$

$\downarrow = 2 \cdot 1 \cdot 2f_2$

$$f_3 = f_2^2 \quad \frac{df_3}{dx} = 2f_2$$

$$\frac{\partial f(x,y)}{\partial y} = \frac{\partial f_2}{\partial y} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_3}{\partial y}$$

$$\downarrow = 3 \cdot 2f_2$$



$$\text{IV) } f(x, y, z) = x^3 + y^2 + z$$

$$f_1 = x^3$$

$$\frac{\partial f_1}{\partial x} = 3x^2$$

$$f_2 = f_1 + y^2$$

$$\frac{\partial f_2}{\partial f_1} = 1 \quad \frac{\partial f_2}{\partial y} = 2y$$

$$f_3 = f_2 + z$$

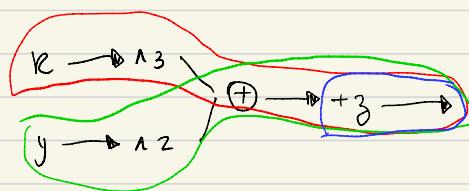
$$\frac{\partial f_3}{\partial f_2} = 1 \quad \frac{\partial f_3}{\partial z} = 1$$

$$\frac{\partial f(x, y, z)}{\partial x} = \frac{\partial f_1}{\partial x} \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_3}{\partial x}$$

$$= 3x^2 \cdot 1 \cdot 1$$

$$\frac{\partial f(x, y, z)}{\partial y} = \frac{\partial f_2}{\partial y} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_3}{\partial y}$$

$$= 2y \cdot 1$$



$$\frac{\partial f(x, y, z)}{\partial z} = \frac{\partial f_3}{\partial z}$$

$$= 1$$

$$\left( \frac{de^x}{dx} = e^x \right)$$

$$\text{V) } g(z) = \frac{1}{1 + e^{-z}}$$

$$f_1 = -1 \cdot z$$

$$\frac{\partial f_1}{\partial x} = -1$$

$$\frac{\partial g(z)}{\partial z} = -1 \cdot e^{f_1} \cdot 1 \cdot -1 f_1^{-2}$$

$$f_2 = e^{f_1}$$

$$\frac{\partial f_2}{\partial f_1} = e^{f_1}$$

$$= e^{f_1} \cdot f_1^{-2}$$

$$f_3 = 1 + f_2 \quad \frac{\partial f_3}{\partial f_2} = 1$$

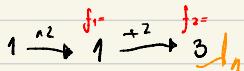
$$f_4 = \frac{1}{f_3} = f_3^{-1} \quad \frac{\partial f_4}{\partial f_3} = -1 f_3^{-2}$$

$$z \rightarrow \textcircled{-1} \rightarrow \textcircled{e^z} \rightarrow \textcircled{+1} \rightarrow \textcircled{1-1} \rightarrow$$

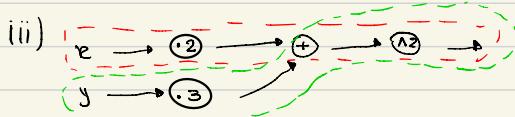
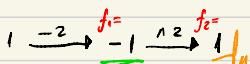
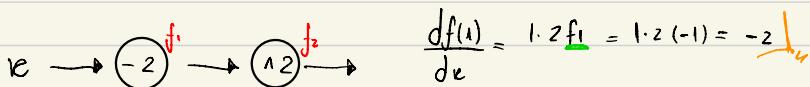
(4)

b)

$$\text{i) } \frac{df(x)}{dx} = \frac{df_2}{df_1} \cdot \frac{df_1}{dx} = 2x \cdot 1 \downarrow, \quad x=1$$



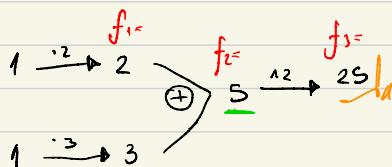
$$\text{ii) } \frac{df(x)}{dx} = \frac{df_2}{df_1} \cdot \frac{df_1}{dx} = 1 \cdot 2 f_1, \quad x=1$$



$$\frac{\partial f(x,y)}{\partial x} = \frac{\partial f_1}{\partial x} \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_3}{\partial x} \quad \frac{\partial f(x,y)}{\partial y} = \frac{\partial f_2}{\partial y} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_3}{\partial y}$$

$\hookrightarrow = 2 \cdot 1 \cdot 2 f_2 \downarrow$

$\hookrightarrow = 3 \cdot 2 f_2 \downarrow$

 $x=1 \quad y=1$ 

$$\frac{df(x,y)}{dx} = 2 \cdot 1 \cdot 2 f_2 = 2 \cdot 1 \cdot 2 \cdot 5 = 20 \downarrow$$

$$\frac{\partial f(x,y)}{\partial y} = 3 \cdot 2 f_2 = 3 \cdot 2 \cdot 5 = 30 \downarrow$$

$$\frac{\partial f(x,y,z)}{\partial x} = \frac{\partial f_1}{\partial x} \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_2}{\partial x}$$

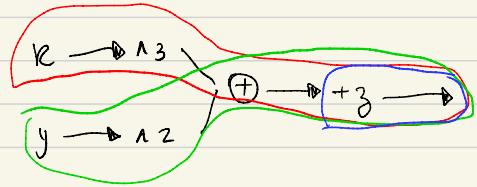
$$\hookrightarrow = 3e^2 \cdot 1 \cdot 1$$

$$\frac{\partial f(x,y,z)}{\partial y} = \frac{\partial f_2}{\partial y} \cdot \frac{\partial f_3}{\partial f_2} = \frac{\partial f_3}{\partial y}$$

$$\hookrightarrow = z_2 \cdot 1 \cdot 1$$

$$\frac{\partial f(x,y,z)}{\partial z} = \frac{\partial f_3}{\partial z}$$

$$\hookrightarrow = 1$$



$$x=1 \quad y=1 \quad z=1$$

$$\begin{array}{c} f_1= \\ 1 \xrightarrow{x_3} 1 \\ 1 \xrightarrow{x_2} 1 \end{array} \xrightarrow{+} \begin{array}{c} f_2= \\ 2 \xrightarrow{+3} 3 \end{array}$$

$$\frac{\partial f(1,1,1)}{\partial x} = 3 \cdot 1^2 = 3 \quad \frac{\partial f(1,1,1)}{\partial y} = 2 \cdot 1 \cdot 1 = 2 \quad \frac{\partial f(1,1,1)}{\partial z} = 1$$

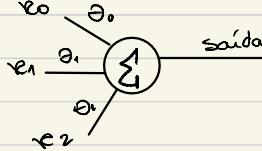
v)  $\frac{\partial g(z)}{\partial z} = -1 \cdot e^{f_1} \cdot 1 \cdot -1 f_3^{-2}$

$$\begin{aligned} & \quad z \rightarrow (-1) \rightarrow (e^1) \rightarrow (+1) \rightarrow (1-1) \rightarrow \\ & = e^{f_1} \cdot f_3^{-2} \end{aligned}$$

$$\begin{aligned} & z=1 \quad f_1=-1 \quad f_2=e^1 = 0,368 \quad f_3=1,368 \quad f_4=0,731 \\ & 1 \xrightarrow{-1} -1 \xrightarrow{e^1} 0,368 \xrightarrow{+1} 1,368 \xrightarrow{1-1} 0,731 \end{aligned}$$

$$\frac{\partial g(1)}{\partial z} = e^{f_1} \cdot f_3^{-2} = e^{-1} \cdot 1,368^{-2} = 0,368 \cdot 0,534 = 0,196$$

- 5) a) O perceptron pode ser geometricamente representado por uma reta ou hiperplano.



logo temos

$$\underbrace{\kappa_0\theta_0 + \kappa_1\theta_1 + \kappa_2\theta_2}_{\text{eq. linear}} = \theta^T X$$

$$X = \begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix}$$

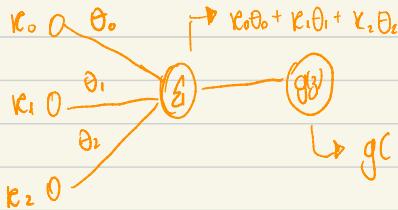
$$\theta = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 \end{bmatrix}$$

- 5) b) Como mostrado no exercício anterior a saída do perceptron (se passar por uma função de ativação) é equivalente a uma reta (hiperplano).

Assim como a regressão logística queremos encontrar os melhores pesos a fim de contemplar os dados do nosso dataset.

Se adicionarmos a arquitetura do perceptron como função de ativação a função logística temos que a atualização ocorre da mesma forma que o algoritmo da Regressão Logística.

### Perceptron



$$\theta_f = \theta_f - \alpha \cdot \frac{\partial \text{loss}}{\partial \theta_f}$$

### Regressão Logística

$$\alpha \kappa_0 + \beta \kappa_1 + \gamma \kappa_2 = 0 \rightarrow \text{Eq. geral da reta}$$

$$\theta_f = \theta_f - \alpha \cdot \frac{\partial \text{loss}}{\partial \theta_f}$$

$$\theta_0 \kappa_0 + \theta_1 \kappa_1 + \theta_2 \kappa_2$$

$$\text{loss} = g(y) = \frac{1}{1 + e^{-y}}$$

5) c) AND

	$R_0$	$R_1$	$R_2$	$y$
bias	1	0	0	0
	1	0	1	0
	1	1	0	0
	1	1	1	1

$$\Theta_0 = 1$$

$$\Theta_1 = 1$$

$$\Theta_2 = 1$$

$$\alpha = 0,5$$

Regra de atualização:

$$\Theta_j = \Theta_j - 2\alpha [ (g(\Theta x^T) - y) \cdot g(\Theta x^T) (1 - g(\Theta x^T))] \cdot c_j$$

① Primeira iteração

$$\Theta_0 = 1$$

$$\Theta = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}_{1 \times 3}$$

$$\Theta x^T = \begin{bmatrix} 1 & 2 & 2 & 3 \end{bmatrix}_{1 \times 4}$$

$$\Theta_1 = 1$$

$$\Theta_2 = 1$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}_{3 \times 4}$$

$$g(\Theta x^T) = \begin{bmatrix} 0.73 & 0.77 & 0.77 & 0.95 \end{bmatrix}$$

$$\alpha = 0,5$$

$$y = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$$

Em lotes

$$\Theta = \Theta - \left( 2 \cdot \alpha \cdot ((g(\Theta x^T) - y) \cdot g(\Theta x^T) (1 - g(\Theta x^T)) \cdot X) \right) / \text{len}(\Theta)$$

$$\Theta = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} - \left( 2 \cdot 1/2 \cdot ((g[\begin{bmatrix} 1 & 2 & 2 & 3 \end{bmatrix}] - [0 \ 0 \ 0 \ 1]) \cdot g[\begin{bmatrix} 1 & 2 & 2 & 3 \end{bmatrix}]) \cdot (1 - g[\begin{bmatrix} 1 & 2 & 2 & 3 \end{bmatrix}]) \cdot X \right) / 3$$

$$= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} - \left( ([0.73 \ 0.77 \ 0.77 \ 0.95] - [0 \ 0 \ 0 \ 1]) \cdot [0.73 \ 0.77 \ 0.77 \ 0.95] \cdot (1 - [0.73 \ 0.77 \ 0.77 \ 0.95]) \cdot X \right) / 3$$

$$= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} - \left( [0.73 \ 0.77 \ 0.77 \ 0.95] \cdot [0.73 \ 0.77 \ 0.77 \ 0.95] \cdot [0.27 \ 0.12 \ 0.12 \ 0.05] \cdot X \right) / 3$$

$$= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} - \left( [0.73 \ 0.77 \ 0.77 \ 0.95] \cdot [0.73 \ 0.77 \ 0.77 \ 0.95] \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) / 3$$

$$= [0.6129 \ 0.8557 \ 0.7557]$$

② Segunda iteração

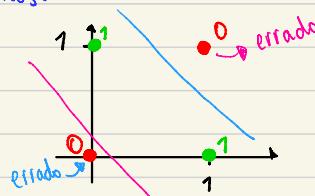
$$\Theta = [0.1476 \ 0.6614 \ 0.6614]$$

⑤ d) Tabela XOR

$\begin{matrix} \text{x} \\ \text{y} \end{matrix}$

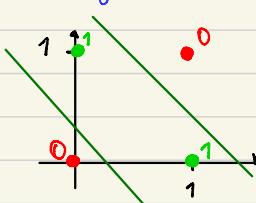
0	0	0
0	1	1
1	0	1
1	1	0

Se plotarmos essa tabela nos eixos  $(X, Y)$  temos:



Pelo gráfico plotado é notável que é impossível separar corretamente as classes 0 e 1 apenas com uma reta.

Entretanto o efeito desejado é atingido se usarmos duas retas, como o ilustrado abaixo:



Como foi demonstrado no exercício 5. a. sabemos que um perceptror geometricamente representa uma reta, logo apenas o perceptron não consegue representar a porta lógica XOR. Contudo, através da combinação de percepções (MLP) essa representação torna-se possível.

⑤

e) O bias tem como função deslocar a ativação do modelo. Se considerarmos uma reta com  $f(x) = ax + b$  o bias é representado pelo termo  $b$ , também conhecido como coeficiente linear da reta.

Se nosso objetivo for obter, por exemplo, um modelo que possa aprender dada a idade de um indivíduo, se ele pode ou não, ingerir bebidas alcoólicas legalmente, seria necessário definir nosso bias = 18. Sem o bias essa condição não seria corretamente aprendida pelo modelo.

7

a) Detector de bordas horizontais

Kernel =

b) Kernel =  
array([[[ 0, 1, 2],  
 [ 3, 4, 5], Ch1  
 [ 6, 7, 8]]],

```
[[ 9, 10, 11],  
 [12, 13, 14], Ch7  
 [15, 16, 17]],
```

```
[[18, 19, 20],  
 [21, 22, 23],  
 [24, 25, 26]]])
```

```
Imagen =  
array([[0, 1, 0, 0],  
       [1, 1, 1, 1], Ch!  
       [0, 1, 1, 1],  
       [1, 0, 0, 1]]),
```

```
[[0, 0, 0, 1],  
 [0, 1, 0, 1],  
 [1, 0, 0, 0],  
 [1, 1, 1, 1]],
```

```
[[1, 1, 0, 0],  
 [0, 1, 1, 0],  
 [0, 1, 1, 1],  
 [1, 0, 1, 0]]])
```

$$0.0 + 1 \cdot 1 + 2 \cdot 0 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 0 + 7 \cdot 1 + 8 \cdot 1 = R_1$$

$$1 + 3 + 4 + 5 + 7 + 8 = \boxed{30}$$

$$\textcolor{red}{\triangleright R_1 = z_8}$$

$$\cancel{0 \cdot 1} + \cancel{1 \cdot 0} + \cancel{2 \cdot 0} + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 1 + 7 \cdot 1 + 8 \cdot 1 = k_2$$

$$3+4+5+6+7+8 = R_2$$

$$\blacktriangleright R_2 = 33$$

~~$$0 \cdot 1 + 1 \cdot 1 + 2 \cdot 1 + 3 \cdot 0 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 1 + 7 \cdot 0 + 8 \cdot 0 = R_3$$~~

$$1+2+4+5+6 = R_3$$

$$\triangleright R_3 = 18$$

$$0 \cdot 1 + 1 \cdot 1 + 2 \cdot 1 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot Q + 7 \cdot O + 8 \cdot 1 = P_9$$

$$1+2+3+4+5+8 = R4$$

$$\rightarrow R_4 = 23$$

$$9.0 + 10.0 + 11.0 + 12.0 + 13.1 + 14.0 + 15.1 + 16.6 + 17.0 = 64.8$$

$$13 + 15 = 61$$

►  $G_1 = 28$

$$96 + 10.1 + 11.6 + 12.1 + 13.0 + 14.6 + 15.1 + 16.1 + 17.1 = 693$$

$$10 + 12 + 15 + 16 + 17 = 64$$

►  $G_3 = 70$

$$9.0 + 10.0 + 11.1 + 12.1 + 13.0 + 14.1 + 15.0 + 16.0 + 17.0 = 122$$

$$11 + 12 + 14 = 67$$

$$\Rightarrow G_2 = 37$$

$$9 \cdot 1 + 10 \cdot 6 + 11 \cdot 1 + 12 \cdot 6 + 13 \cdot 6 + 14 \cdot 6 + 15 \cdot 1 + 16 \cdot 1 + 17 \cdot 1 = 64$$

$$9 + 11 + 15 + 16 + 17 = 64$$

$$\triangleright G_4 = 67$$

$$18 \cdot 1 + 19 \cdot 1 + 20 \cdot 0 + 21 \cdot 0 + 22 \cdot 1 + 23 \cdot 1 + 24 \cdot 0 + 25 \cdot 1 + 26 \cdot 1 = B_1$$

$$18 + 19 + 22 + 23 + 25 + 26 = B_1$$

$$\blacktriangleright B_1 = 133$$

$$18 \cdot 1 + 19 \cdot 0 + 20 \cdot 0 + 21 \cdot 1 + 22 \cdot 1 + 23 \cdot 0 + 24 \cdot 1 + 25 \cdot 1 + 26 \cdot 1 = B_2$$

$$18 + 21 + 22 + 24 + 25 + 26 = B_2$$

$$\blacktriangleright B_2 = 136$$

$$18 \cdot 0 + 19 \cdot 1 + 20 \cdot 1 + 21 \cdot 0 + 22 \cdot 1 + 23 \cdot 1 + 24 \cdot 1 + 25 \cdot 0 + 26 \cdot 1 = B_3$$

$$19 + 20 + 22 + 23 + 24 + 26 = B_3$$

$$\blacktriangleright B_3 = 134$$

$$18 \cdot 1 + 19 \cdot 1 + 20 \cdot 0 + 21 \cdot 1 + 22 \cdot 1 + 23 \cdot 1 + 24 \cdot 0 + 25 \cdot 1 + 26 \cdot 0 = B_4$$

$$18 + 19 + 21 + 22 + 23 + 25 = B_4$$

$$\blacktriangleright B_4 = 128$$

$$R_1 + G_1 + B_1 = 28 + 27 + 133 = 188$$

$$R_2 + G_2 + B_2 = 33 + 37 + 136 = 206$$

$$R_3 + G_3 + B_3 = 18 + 70 + 134 = 222$$

$$R_4 + G_4 + B_4 = 23 + 68 + 127 = 219$$

$$\text{Feature Map final} = \begin{bmatrix} 188 & 206 \\ 222 & 219 \end{bmatrix}$$

(7)

c)

$$n_0 = \underbrace{n_1 + 2P - K}_{5} + 1$$

i.

$$f_{M1} = \frac{227 + 2 \cdot 1 - 11}{1} + 1 = 219$$

$$f_{M1} = 219 \times 219 \times 96$$

$$f_{MS} = \frac{55 + 2 \cdot 1 - 3}{1} + 1 = 55$$

$$f_{MS} = 55 \times 55 \times 384$$

$$f_{M2} = \frac{219 + 2 \cdot 1 - 3}{2} + 1 = 110$$

$$f_{M2} = 110 \times 110 \times 96$$

$$f_{M6} = \frac{55 + 2 \cdot 1 - 3}{1} + 1 = 55$$

$$f_{M6} = 55 \times 55 \times 384$$

$$f_{M3} = \frac{110 + 2 \cdot 2 - 5}{1} + 1 = 110$$

$$f_{M3} = 110 \times 110 \times 256$$

$$f_{M7} = \frac{55 + 2 \cdot 1 - 3}{1} + 1 = 55$$

$$f_{M7} = 55 \times 55 \times 256$$

$$f_{M4} = \frac{110 + 2 \cdot 1 - 3}{2} + 1 = \left\lfloor \frac{109}{2} \right\rfloor + 1 = 54 + 1 = 55$$

$$f_{M4} = 55 \times 55 \times 256$$

$$f_{M8} = \frac{55 + 2 \cdot 1 - 3}{2} + 1 = 28$$

$$f_{M8} = 28 \times 28 \times 256$$

ii.

$$emb1 = 28 \cdot 28 \cdot 256 \times 4096$$

$$emb2 = 4096 \times 4096$$

iii. Sempre após as camadas de convolução e as redes "fully conn.",  
exceto após a última camada.

iv.

$$28 \cdot 28 \cdot 256 = 200704$$

V. ①  $\frac{96(41 \times 11 \times 3 + 1)}{4} = 96 \cdot 364 = 34944 = P_1$

1 in channels      41      bias  
out channels      4      dimensions do kernel

②  $256(5 \times 5 \times 96 + 1) = 256 \cdot 2401 = 614656 = P_2$

$$\textcircled{3} \quad 384(3 \times 3 \times 256 + 1) = 384 \cdot 2305 = 885120 = P_3$$

$$\textcircled{4} \quad 384(3 \times 3 \times 384 + 1) = 384 \cdot 3457 = 1327488 = P_4$$

$$\textcircled{5} \quad 256(3 \times 3 \times 384 + 1) = 256 \cdot 3457 = 884992 = P_5$$

$$\textcircled{6} \quad (28 \cdot 28 \cdot 256 + 1) \cdot 4096 = 200705 \cdot 4096 = 822087680 = P_6$$

Bimensão da fc bias      Saída fc  
entrada

$$\textcircled{7} \quad (4096 + 1) \cdot 4096 = 16771312 = P_7$$

$$\textcircled{8} \quad (4096 + 1) \cdot 1000 = 4097 \cdot 1000 = 4097000 = P_8$$

vi.

$$\begin{aligned}\text{Total\_params} &= P_1 + P_2 + P_3 + P_4 + P_5 + P_6 + P_7 + P_8 \\ &= 846713192\end{aligned}$$

vii. add link