Redes Neurais Artificiais

Pedro H A Konzen

27 de junho de 2023

Licença

Prefácio

Nestas notas de aula são abordados tópicos introdutórios sobre redes neurais artificiais Como ferramenta computacional de apoio, vários exemplos de aplicação de códigos Python+PyTorch são apresentados.

Agradeço a todas e todos que de modo assíduo ou esporádico contribuem com correções, sugestões e críticas. :)

Pedro H A Konzen

Conteúdo

Capa **i**

Licença **ii**

Prefácio **iii**

Sumário **v**

1 Introdução **1**

2 Perceptron **3**
- 2.1 Unidade de Processamento 3
  - 2.1.1 Um problema de classificação 4
  - 2.1.2 Problema de regressão 10
  - 2.1.3 Exercícios 13
- 2.2 Algoritmo de Treinamento 14
  - 2.2.1 Método do Gradiente Descendente 15
  - 2.2.2 Método do Gradiente Estocástico 18
  - 2.2.3 Exercícios 20

3 Perceptron Multicamadas **21**
- 3.1 Modelo MLP 21
  - 3.1.1 Treinamento 22
  - 3.1.2 Aplicação: Problema de Classificação `XOR` 24
  - 3.1.3 Exercícios 26
- 3.2 Aplicação: Aproximação de Funções 26
  - 3.2.1 Função unidimensional 26
  - 3.2.2 Função bidimensional 29

3.2.3 Exercícios 31
3.3 Aplicação: Equação de Laplace 31
3.3.1 Exercícios 36

Respostas dos Exercícios **37**

Referências Bibliográficas **38**

Capítulo 1

Introdução

Uma rede neural artificial é um modelo de aprendizagem profunda (**deep learning**), uma área da aprendizagem de máquina (**machine learning**). O termo tem origem no início dos desenvolvimentos de inteligência artificial, em que modelos matemáticos e computacionais foram inspirados no cérebro biológico (tanto de humanos como de outros animais). Muitas vezes desenvolvidos com o objetivo de compreender o funcionamento do cérebro, também tinham a intensão de emular a inteligência.

Nestas notas de aula, estudamos um dos modelos de redes neurais usualmente aplicados. A unidade básica de processamento é data do modelo de neurônio de McCulloch-Pitts (McCulloch and Pitts, 1943), conhecido como **perceptron** (Rosenblatt, 1958, 1962), o primeiro com um algoritmo de treinamento para problemas de classificação linearmente separável. Um modelo similiar é o ADALINE (do inglês, *adaptive linear element*, Widrow and Hoff, 1960), desenvolvido para a predição de números reais. Pela questão histórica, vamos usar o termo **perceptron** para designar a unidade básica (o neurônio), mesmo que o modelo de neurônio a ser estudado não seja restrito ao original.

Métodos de aprendizagem profunda são técnicas de treinamento (calibração) de composições em múltiplos níveis, aplicáveis a problemas de aprendizagem de máquina que, muitas vezes, não têm relação com o cérebro ou neurônios biológicos. Um exemplo, é a rede neural que mais vamos explorar nas notas, o **perceptron multicamada** (PMC, em inglês *multilayer per-*

ceptron, MLP), um modelo de progressão (em inglês, *feedfoward*) de rede profunda em que a informação é processada pela composição de camadas de perceptrons. Embora a ideia de fazer com que a informação seja processada através da conexão de múltiplos neurônios tenha inspiração biológica, usualmente a escolha da disposição dos neurônios em um PMC é feita por questões algorítmicas e computacionais. I.e., baseada na eficiente utilização da arquitetura dos computadores atuais.

Capítulo 2

Perceptron

2.1 Unidade de Processamento

A unidade básica de processamento (neurônio artificial) que exploramos nestas notas é baseado no **perceptron** (consultemos a Fig. 2.1). Consiste na composição de uma **função de ativação** $f : \mathbb{R} \to \mathbb{R}$ com a **pré-ativação**

$$\begin{aligned} z &= \boldsymbol{w} \cdot \boldsymbol{x} + b && (2.1)\\ &= w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b && (2.2) \end{aligned}$$

onde, $\boldsymbol{x} \in \mathbb{R}^n$ é o vetor de entrada, $\boldsymbol{w} \in \mathbb{R}^n$ é o vetor de pesos e $b \in \mathbb{R}$ é o *bias*. Escolhida uma função de ativação, a saída do neurônio é dada por

$$\begin{aligned} y &:= \mathcal{N}\left(\boldsymbol{x}; (\boldsymbol{w}, b)\right) && (2.3)\\ &= f(z) = f(\boldsymbol{w} \cdot \boldsymbol{x} + b) && (2.4) \end{aligned}$$

O treinamento (calibração) consiste em determinar os parâmetros $(\boldsymbol{w}, b)$ de forma que o neurônio forneça as saídas y esperadas com base em algum critério predeterminado.

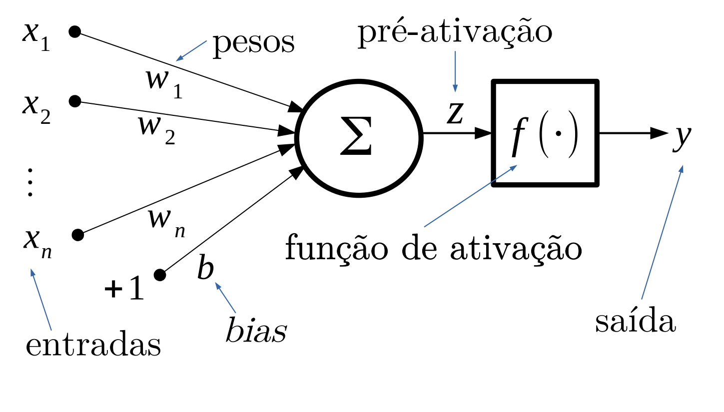

Figura 2.1: Esquema de um perceptron: unidade de processamento.

Uma das vantagens deste modelo de neurônio é sua generalidade, i.e. pode ser aplicado a diferentes problemas. Na sequência, vamos aplicá-lo na resolução de um problema de classificação e noutro de regressão.

2.1.1 Um problema de classificação

Vamos desenvolver um perceptron que faça a operação $\wedge$ (e-lógico). I.e, receba como entrada dois valores lógicos A_1 e A_2 (V, verdadeiro ou F, falso) e forneça como saída o valor lógico $R = A_1 \wedge A_2$. Consultemos a seguinte tabela verdade:

A_1	A_2	R
V	V	V
V	F	F
F	V	F
F	F	F

Modelo

Nosso modelo de neurônio será um perceptron com duas entradas $\boldsymbol{x} \in \{-1,1\}^2$ e a função sinal

$$f(z) = \text{sign}(z) = \begin{cases} 1 & , z > 0 \\ 0 & , z = 0 \\ -1 & , z < 0 \end{cases} \tag{2.5}$$

como função de ativação, i.e.

$$\mathcal{N}\left(\boldsymbol{x};(\boldsymbol{w},b)\right) = \operatorname{sign}(\boldsymbol{w}\cdot\boldsymbol{x}+b), \tag{2.6}$$

onde $\boldsymbol{w} \in \mathbb{R}^2$ e $b \in \mathbb{R}$ são parâmetros a determinar.

Pré-processamento

Uma vez que nosso modelo recebe valores $\boldsymbol{x} \in \{-1,1\}^2$ e retorna $y \in \{-1,1\}$, precisamos (pre)processar os dados do problema de forma a utilizá-lo. Uma forma, é assumir que todo valor negativo está associado ao valor lógico F (falso) e positivo ao valor lógico V (verdadeiro). Desta forma, os dados podem ser interpretados como na seguinte tabela

x_1	x_2	y
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

Treinamento

Agora, nos falta treinar nosso neurônio para fornecer o valor de y esperado para cada dada uma entrada $\boldsymbol{x}$. Isso consiste em um método para escolhermos os parâmetros $(\boldsymbol{w},b)$ que sejam adequados para esta tarefa. Vamos explorar mais sobre isso na sequência do texto e, aqui, apenas escolhemos

$$\boldsymbol{w} = [1, 1] \tag{2.7}$$

$$b = -1 \tag{2.8}$$

Com isso, nosso perceptron é

$$\mathcal{N}(\boldsymbol{x}) = \operatorname{sign}(x_1 + x_2 - 1) \tag{2.9}$$

Verifique que ele satisfaz a tabela verdade acima!

Implementação

Código 2.1: perceptron.py

```python
import torch
```

```python

# modelo
class Perceptron(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.linear = torch.nn.Linear(2,1)

    def forward(self, x):
        z =  self.linear(x)
        y = torch.sign(z)
        return y

model = Perceptron()
W = torch.Tensor([[1., 1.]])
b = torch.Tensor([-1.])
with torch.no_grad():
    model.linear.weight = torch.nn.Parameter(W)
    model.linear.bias = torch.nn.Parameter(b)

# dados de entrada
X = torch.tensor([[1., 1.],
                  [1., -1.],
                  [-1., 1.],
                  [-1., -1.]])

print(f"\nDados de entrada\n{X}")


# forward (aplicação do modelo)
y = model(X)

print(f"Valores estimados\n{y}")
```

Interpretação geométrica

Empregamos o seguinte modelo de neurônio

$$\mathcal{N}(\boldsymbol{x}; (\boldsymbol{w}, b)) = \operatorname{sign}(w_1 x_1 + w_2 x_2 + b) \tag{2.10}$$

Observamos que

$$w_1 x_1 + w_2 x_2 + b = 0 \tag{2.11}$$

corresponde à equação geral de uma reta no plano $\tau : x_1 \times x_2$. Esta reta divide o plano em dois semiplanos

$$\tau^+ = \{\boldsymbol{x} \in \mathbb{R}^2 : w_1x_1 + w_2x_2 + b > 0\} \tag{2.12}$$

$$\tau^- = \{\boldsymbol{x} \in \mathbb{R}^2 : w_1x_1 + w_2x_2 + b < 0\} \tag{2.13}$$

O primeiro está na direção do vetor normal a reta $\boldsymbol{n} = (w_1, w_2)$ e o segundo na sua direção oposta. Com isso, o problema de treinar nosso neurônio para nosso problema de classificação consiste em encontrar a reta

$$w_1x_1 + w_2x_2 + b = 0 \tag{2.14}$$

de forma que o ponto (1,1) esteja no semiplano positivo τ^+ e os demais pontos no semiplano negativo τ^-. Consulte a Figura 2.2.

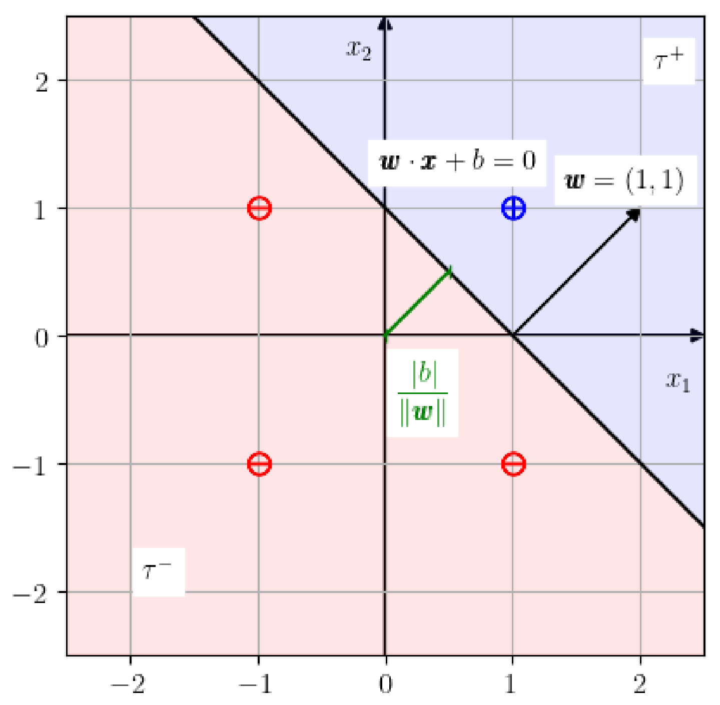

Figura 2.2: Interpretação geométrica do perceptron aplicado ao problema de classificação ralacionado à operação lógica $\wedge$ (e-lógico).

Algoritmo de treinamento: perceptron

O algoritmo de treinamento perceptron permite calibrar os pesos de um neurônio para fazer a classificação de dados linearmente separáveis. Trata-se de um algoritmo para o **treinamento supervisionado** de um neurônio, i.e. a calibração dos pesos é feita com base em um dado **conjunto de amostras de treinamento**.

Seja dado um **conjunto de treinamento** $\{\boldsymbol{x}^{(s)},y^{(s)}\}_{s=1}^{n_s}$, onde n_s é o número de amostras. O algoritmo consiste no seguinte:

1. $\boldsymbol{w} \leftarrow \boldsymbol{0}$, $b \leftarrow 0$.
2. Para $e \leftarrow 1, \ldots, n_e$:
   (a) Para $s \leftarrow 1, \ldots, n_s$:
      i. Se $y^{(s)}\mathcal{N}\left(\boldsymbol{x}^{(s)}\right) \leq 0$:
         A. $\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(s)}\boldsymbol{x}^{(s)}$
         B. $b \leftarrow b + y^{(s)}$

onde, n_e é um dado número de épocas[1].

Código 2.2: perceptron_train.py

```python
import torch

# modelo

class Perceptron(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.linear = torch.nn.Linear(2,1)

    def forward(self, x):
        z =  self.linear(x)
        y = torch.sign(z)
        return y

model = Perceptron()
```

[1]Número de vezes que as amostrar serão percorridas para realizar a correção dos pesos.

```
with torch.no_grad():
    W = model.linear.weight
    b = model.linear.bias

# dados de treinamento
X_train = torch.tensor([[1., 1.],
                        [1., -1.],
                        [-1., 1.],
                        [-1., -1.]])
y_train = torch.tensor([1., -1., -1., -1.]).reshape(-1,1)

## número de amostras
ns = y_train.size(0)

print("\nDados de treinamento")
print("X_train =")
print(X_train)
print("y_train = ")
print(y_train)

# treinamento

## num max épocas
nepochs = 100

for epoch in range(nepochs):

    # forward
    y_est = model(X_train)

    # update
    not_updated = True
    for s in range(ns):
        if (y_est[s]*y_train[s] <= 0.):
            with torch.no_grad():
                W += y_train[s]*X_train[s,:]
                b += y_train[s]
                not_updated = False

    if (not_updated):
```

```
            print('Training ended.')
            break


# verificação
print(f'W =\n{W}')
print(f'b =\n{b}')
y = model(X_train)
print(f'y =\n{y}')
```

2.1.2 Problema de regressão

Vamos treinar um perceptron para resolver o problema de regressão linear para os seguintes dados

s	$x^{(s)}$	$y^{(s)}$
1	0.5	1.2
2	1.0	2.1
3	1.5	2.6
4	2.0	3.6

Modelo

Vamos determinar o perceptron[2]

$$\tilde{y} = \mathcal{N}(x; (w, b)) = wx + b \tag{2.15}$$

que melhor se ajusta a este conjunto de dados $\left\{(x^{(s)}, y^{(s)})\right\}_{s=1}^{n_s}$, $n_s = 4$.

Treinamento

A ideia é que o perceptron seja tal que minimize o erro quadrático médio (EQM)[3], i.e.

$$\min_{w,b} \frac{1}{n_s} \sum_{s=1}^{n_s} \left(\tilde{y}^{(s)} - y^{(s)}\right)^2 \tag{2.16}$$

[2]Escolhendo $f(z) = z$ como função de ativação.
[3]Em inglês, *mean squared error (MSE)*.

Vamos denotar a **função erro**[4] por

$$\begin{aligned}\varepsilon(w,b) &:= \frac{1}{n_s}\sum_{s=1}^{n_s}\left(\tilde{y}^{(s)} - y^{(s)}\right)^2 && (2.17)\\ &= \frac{1}{n_s}\sum_{s=1}^{n_s}\left(wx^{(s)} + b - y^{(s)}\right)^2 && (2.18)\end{aligned}$$

Observamos que (2.16) é equivalente a um problema linear de mínimos quadrados. A solução é obtida resolvendo-se a equação normal[5]

$$M^TM\boldsymbol{c} = M^T\boldsymbol{y}, \tag{2.19}$$

onde $\boldsymbol{c} = (w,p)$ é o vetor dos parâmetros a determinar e M é a matriz $n_s \times 2$ dada por

$$M = \begin{bmatrix}\boldsymbol{x} & \mathbf{1}\end{bmatrix} \tag{2.20}$$

Implementação

Código 2.3: perceptron_mq.py

```python
import torch

# modelo

class Perceptron(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.linear = torch.nn.Linear(1,1)

    def forward(self, x):
        z =  self.linear(x)
        return z

model = Perceptron()
with torch.no_grad():
    W = model.linear.weight
```

[4] Em inglês, *loss function*.
[5] Consulte o Exercício 2.1.2.

```
    b = model.linear.bias

# dados de treinamento
X_train = torch.tensor([0.5,
                        1.0,
                        1.5,
                        2.0]).reshape(-1,1)
y_train = torch.tensor([1.2,
                        2.1,
                        2.6,
                        3.6]).reshape(-1,1)

## número de amostras
ns = y_train.size(0)

print("\nDados de treinamento")
print("X_train =")
print(X_train)
print("y_train = ")
print(y_train)

# treinamento

## matriz
M = torch.cat((X_train,
               torch.ones((ns,1))), dim=1)
## solucão M.Q.
c = torch.linalg.lstsq(M, y_train)[0]
with torch.no_grad():
    W = c[0]
    b = c[1]

# verificação
print(f'W =\n{W}')
print(f'b =\n{b}')
y = model(X_train)
print(f'y =\n{y}')
```

Resultado

Nosso perceptron corresponde ao modelo

$$\mathcal{N}(x;(w,b)) = wx + b \tag{2.21}$$

com os pesos treinados $w = 1.54$ e $b = 0.45$. Ele corresponde à reta que melhor se ajusta ao conjunto de dados de $\left\{x^{(s)}, y^{(s)}\right\}$. Consulte a Figura 2.3.

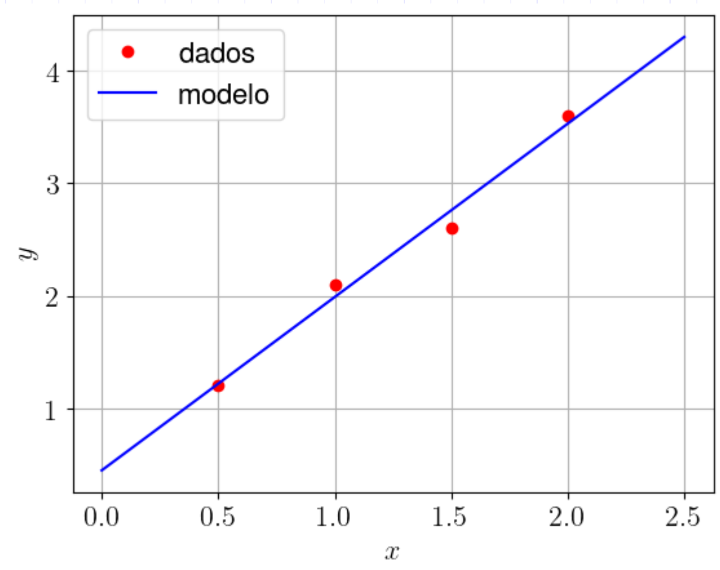

Figura 2.3: Interpretação geométrica do perceptron aplicado ao problema de regressão linear.

2.1.3 Exercícios

[[tag:construcao]]

Exercício 2.1.1. Assumindo o modelo de neurônio (2.15), mostre que (2.17) é função convexa.

Exercício 2.1.2. Mostre que a solução do problema (2.16) é dada por (2.19).

2.2 Algoritmo de Treinamento

Na seção anterior, desenvolvemos dois modelos de neurônios para problemas diferentes, um de classificação e outro de regressão. Em cada caso, utilizamos algoritmos de treinamento diferentes. Agora, vamos estudar algoritmos de treinamentos mais gerais[6], que podem ser aplicados a ambos os problemas.

Ao longo da seção, vamos considerar o modelo de neurônio

$$\mathcal{N}\left(\boldsymbol{x};(\boldsymbol{w},b)\right) = f(\boldsymbol{w}\cdot\boldsymbol{x}+b), \tag{2.22}$$

com dada função de ativação $f:\mathbb{R}\to\mathbb{R}$, sendo os vetores de entrada $\boldsymbol{x}$ e dos pesos $\boldsymbol{w}$ de tamanho n_{in}. A pré-ativação do neurônio é denotada por

$$z := \boldsymbol{w}\cdot\boldsymbol{x}+b \tag{2.23}$$

Fornecido um conjunto de treinamento $\left\{\left(\boldsymbol{x}^{(s)},y^{(s)}\right)\right\}_1^{n_s}$, com n_s amostras, o objetivo é calcular os parâmetros $(\boldsymbol{w},b)$ que minimizam a função erro quadrático médio

$$\varepsilon(\boldsymbol{w},b) := \frac{1}{n_s}\sum_{s=1}^{n_s}\left(\tilde{y}^{(s)}-y^{(s)}\right)^2 \tag{2.24}$$

$$= \frac{1}{n_s}\sum_{s=1}^{n_s}\varepsilon^{(s)} \tag{2.25}$$

onde $\tilde{y}^{(s)} = \mathcal{N}\left(\boldsymbol{x}^{(s)};(\boldsymbol{w},b)\right)$ é o valor estimado pelo modelo para a s-ésima amostra e $\varepsilon^{(s)} := \left(\tilde{y}^{(s)}-y^{(s)}\right)^2$. I.e., queremos resolver o seguinte problema de otimização

$$\min_{(\boldsymbol{w},b)} \varepsilon(\boldsymbol{w},b) \tag{2.26}$$

Para resolver este problema de otimização, vamos empregar o Método do Gradiente Descendente.

[6] Aqui, vamos explorar apenas algoritmos de treinamento supervisionado.

2.2.1 Método do Gradiente Descendente

O Método do Gradiente Descendente[7] (GD) é um método de declive. Aplicado ao nosso modelo de perceptron consiste no seguinte algoritmo:

1. $\boldsymbol{w}, b$ aproximações iniciais.
2. Para $e \leftarrow 1, \dots, n_e$:
   (a) $(\boldsymbol{w}, b) \leftarrow (\boldsymbol{w}, b) - l_r \dfrac{\partial \varepsilon}{\partial (\boldsymbol{w}, b)}$

onde, n_e é o **número de épocas**, l_r é uma dada **taxa de aprendizagem**[8] e o gradiente é

$$\frac{\partial \varepsilon}{\partial (\boldsymbol{w}, b)} := \left(\frac{\partial \varepsilon}{\partial w_1}, \dots, \frac{\partial \varepsilon}{\partial w_{n_{in}}}, \frac{\partial \varepsilon}{\partial b}\right) \tag{2.27}$$

O cálculo do gradiente para os pesos $\boldsymbol{w}$ pode ser feito como segue

$$\frac{\partial \varepsilon}{\partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}} \left[\frac{1}{n_s} \sum_{s=1}^{n_s} \varepsilon^{(s)}\right] \tag{2.28}$$

$$= \frac{1}{ns} \sum_{s=1}^{ns} \frac{\partial \varepsilon^{(s)}}{\partial \tilde{y}^{(s)}} \frac{\partial \tilde{y}^{(s)}}{\partial \boldsymbol{w}} \tag{2.29}$$

$$\frac{\partial \varepsilon}{\partial \boldsymbol{w}} = \frac{1}{ns} \sum_{s=1}^{ns} \frac{\partial \varepsilon^{(s)}}{\partial \tilde{y}^{(s)}} \frac{\partial \tilde{y}^{(s)}}{\partial z^{(s)}} \frac{\partial z^{(s)}}{\partial \boldsymbol{w}} \tag{2.30}$$

Observando que

$$\frac{\partial \varepsilon^{(s)}}{\partial \tilde{y}^{(s)}} = 2\left(\tilde{y}^{(s)} - y^{(s)}\right) \tag{2.31}$$

$$\frac{\partial \tilde{y}^{(s)}}{\partial z^{(s)}} = f'\left(z^{(s)}\right) \tag{2.32}$$

$$\frac{\partial z^{(s)}}{\partial \boldsymbol{w}} = \boldsymbol{x}^{(s)} \tag{2.33}$$

obtemos

$$\frac{\partial \varepsilon}{\partial \boldsymbol{w}} = \frac{1}{n_s} \sum_{s=1}^{n_s} 2\left(\tilde{y}^{(s)} - y^{(s)}\right) f'\left(z^{(s)}\right) \boldsymbol{x}^{(s)} \tag{2.34}$$

[7]Em inglês, *Gradiente Descent, GD Method.*
[8]Em inglês, *learning rate.*

$$\frac{\partial \varepsilon}{\partial b} = \frac{1}{ns} \sum_{s=1}^{ns} \frac{\partial \varepsilon^{(s)}}{\partial \tilde{y}^{(s)}} \frac{\partial \tilde{y}^{(s)}}{\partial z^{(s)}} \frac{\partial z^{(s)}}{\partial b} \tag{2.35}$$

$$\frac{\partial \varepsilon}{\partial b} = \frac{1}{n_s} \sum_{s=1}^{n_s} 2 \left(\tilde{y}^{(s)} - y^{(s)} \right) f' \left(z^{(s)} \right) \cdot 1 \tag{2.36}$$

Aplicação: Problema de Classificação

Na Subseção 2.1.1, treinamos um Perceptron para o problema de classificação do e-lógico. Usamos $f(x) = \text{sign}(x)$ como função de ativação. Ocorre, que para o Método do Gradiente, esta função de ativação não é apropriada, pois $f'(x) \equiv 0$ para $x \neq 0$. Aqui, vamos usar

$$f(x) = \tanh(x). \tag{2.37}$$

Código 2.4: perceptron_gd.py

```python
import torch

# modelo

class Perceptron(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.linear = torch.nn.Linear(2,1)

    def forward(self, x):
        z =  self.linear(x)
        y = torch.tanh(z)
        return y

model = Perceptron()

# treinamento

## optimizador
optim = torch.optim.SGD(model.parameters(), lr=1e-1)

## função erro
loss_fun = torch.nn.MSELoss()

```

```python
## dados de treinamento
X_train = torch.tensor([[1., 1.],
                        [1., -1.],
                        [-1., 1.],
                        [-1., -1.]])
y_train = torch.tensor([1., -1., -1., -1.]).reshape(-1,1)

print("\nDados de treinamento")
print("X_train =")
print(X_train)
print("y_train = ")
print(y_train)

## num max épocas
nepochs = 5000
tol = 1e-3

for epoch in range(nepochs):

    # forward
    y_est = model(X_train)

    # erro
    loss = loss_fun(y_est, y_train)

    print(f'{epoch}: {loss.item():.4e}')

    # critério de parada
    if (loss.item() < tol):
        break

    # backward
    optim.zero_grad()
    loss.backward()
    optim.step()


# verificação
y = model(X_train)
print(f'y_est = {y}')
```

2.2.2 Método do Gradiente Estocástico

O Método do Gradiente Estocástico é um variação do método anterior. A ideia é atualizar os parâmetros do modelo com base no gradiente do erro de cada amostra. A estocasticidade é obtida da randomização com que as amostras são escolhidas a cada época. O algoritmos consiste no seguinte:

1. **w**, b aproximações inicial.

2. Para $e \leftarrow 1, \dots, n_e$:

    1.1. Para $s \leftarrow \texttt{random}(1, \dots, n_s)$:

$$(\boldsymbol{w}, b) \leftarrow (\boldsymbol{w}, b) - l_r \frac{\partial \varepsilon^{(s)}}{\partial (\boldsymbol{w}, b)} \tag{2.38}$$

Aplicação: Problema de Classificação

Código 2.5: perceptron_sgd.py

```python
import torch
import numpy as np

# modelo

class Perceptron(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.linear = torch.nn.Linear(2,1)

    def forward(self, x):
        z =  self.linear(x)
        y = torch.tanh(z)
        return y

model = Perceptron()

# treinamento

## optimizador
optim = torch.optim.SGD(model.parameters(), lr=1e-1)

```

```
## função erro
loss_fun = torch.nn.MSELoss()

## dados de treinamento
X_train = torch.tensor([[1., 1.],
                        [1., -1.],
                        [-1., 1.],
                        [-1., -1.]])
y_train = torch.tensor([1., -1., -1., -1.]).reshape(-1,1)

## num de amostras
ns = y_train.size(0)

print("\nDados de treinamento")
print("X_train =")
print(X_train)
print("y_train = ")
print(y_train)

## num max épocas
nepochs = 5000
tol = 1e-3

for epoch in range(nepochs):

    # forward
    y_est = model(X_train)

    # erro
    loss = loss_fun(y_est, y_train)

    print(f'{epoch}: {loss.item():.4e}')

    # critério de parada
    if (loss.item() < tol):
        break

    # backward
    for s in torch.randperm(ns):
        loss_s = (y_est[s,:] - y_train[s,:])**2
```

```
        optim.zero_grad()
        loss_s.backward()
        optim.step()
        y_est = model(X_train)


# verificação
y = model(X_train)
print(f'y_est = {y}')
```

2.2.3 Exercícios

[[tag:construcao]]

Exercício 2.2.1. Calcule a derivada da função de ativação

$$f(x) = \tanh(x). \tag{2.39}$$

Exercício 2.2.2. Use o Método do Gradiente para treinar um Perceptron para o problema de classificação estudado na Subseção 2.1.2.

Exercício 2.2.3. Refaça o Exercício 2.2.2 usando o Método do Gradiente Estocástico como otimizador.

Capítulo 3

Perceptron Multicamadas

3.1 Modelo MLP

Uma Perceptron Multicamadas (MLP, do inglês, *Multilayer Perceptron*) é um tipo de Rede Neural Artificial formada por composições de camadas de perceptrons. Consulte a Figura 3.1.

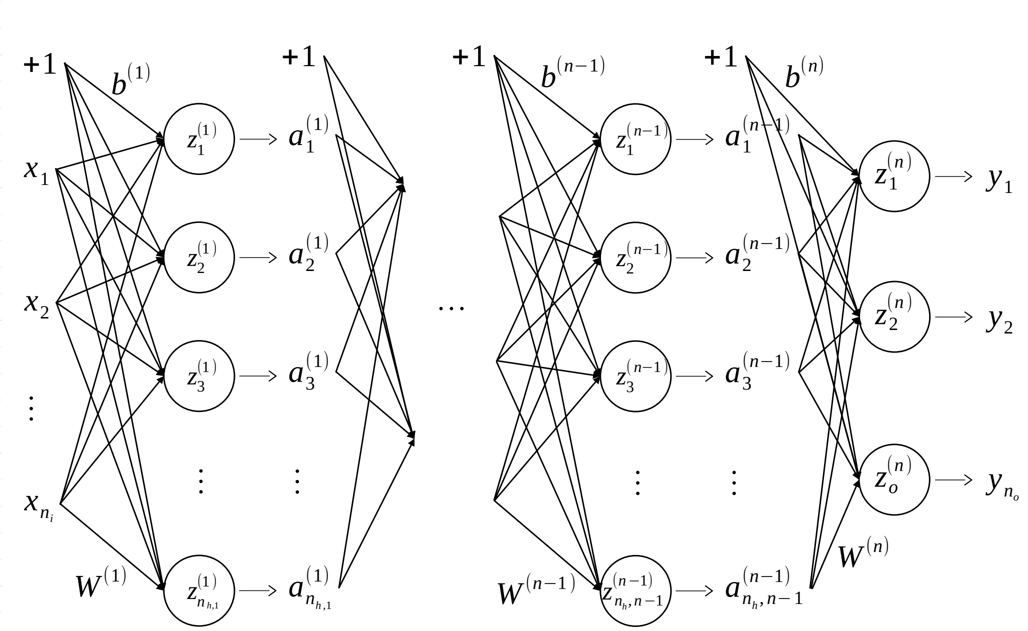

Figura 3.1: Estrutura de uma rede do tipo Perceptron Multicamadas (MLP).

Denotamos uma MLP de n camadas por

$$\boldsymbol{y} = \mathcal{N}\left(\boldsymbol{x}; \left(W^{(l)}, \boldsymbol{b}^{(l)}, f^{(l)}\right)_{l=1}^{n}\right), \tag{3.1}$$

onde $\left(W^{(l)}, \boldsymbol{b}^{(l)}, f^{(l)}\right)$ é a tripa de pesos, bias e função de ativação da l-ésima camada da rede, $l = 1, 2, \dots, n$.

A saída da rede é calculada por iteradas composições das camadas, i.e.

$$\boldsymbol{a}^{(l)} = f^{(l)} \underbrace{\left(W^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l-1)}\right)}_{\boldsymbol{z}^{(l)}}, \tag{3.2}$$

para $l = 1, 2, \dots, n$, denotando $\boldsymbol{a}^{(0)} := \boldsymbol{x}$ e $\boldsymbol{a}^{(n)} := \boldsymbol{y}$.

3.1.1 Treinamento

Fornecido um conjunto de treinamento $\{\boldsymbol{x}^{(s)}\}_{s=1}^{n_s}$, com n_s amostras, o treinamento da rede consiste em resolver o problema de minimização

$$\min_{(W,\boldsymbol{b})} \varepsilon\left(\boldsymbol{y}^{(s)}\right) \tag{3.3}$$

onde ε é uma dada função erro.

O problema de minimização pode ser resolvido por um Método de Declive e, de forma geral, consiste em:

1. $\boldsymbol{w}, b$ aproximações iniciais.
2. Para $e \leftarrow 1, \dots, n_e$:
   (a) $(W, \boldsymbol{b}) \leftarrow (W, \boldsymbol{b}) - l_r \boldsymbol{d}$

onde, n_e é o **número de épocas**, l_r é uma dada **taxa de aprendizagem**[1] e o vetor direção $\boldsymbol{d}$ depende dos gradientes

$$\nabla_{W,\boldsymbol{b}}\varepsilon := \left(\frac{\partial \varepsilon}{\partial W}, \frac{\partial \varepsilon}{\partial \boldsymbol{b}}\right). \tag{3.4}$$

O cálculo dos gradientes pode ser feito **de trás para frente**, i.e. para os pesos da última camada, temos

$$\begin{align}
\frac{\partial \varepsilon}{\partial W^{(n)}} &= \frac{\partial \varepsilon}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}^{(\boldsymbol{n})}} \frac{\partial \boldsymbol{z}^{(\boldsymbol{n})}}{\partial W^{(n)}}, \tag{3.5}\\
&= \frac{\partial \varepsilon}{\partial \boldsymbol{y}} f'\left(W^{(n)}\boldsymbol{a}^{(n-1)} + \boldsymbol{b}^{(n)}\right)\boldsymbol{a}^{(n-1)}. \tag{3.6}
\end{align}$$

Para os pesos da penúltima, temos

$$\begin{align}
\frac{\partial \varepsilon}{\partial W^{(n-1)}} &= \frac{\partial \varepsilon}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}^{(\boldsymbol{n})}} \frac{\partial \boldsymbol{z}^{(\boldsymbol{n})}}{\partial W^{(n-1)}}, \tag{3.7}\\
&= \frac{\partial \varepsilon}{\partial \boldsymbol{y}} f'\left(\boldsymbol{z}^{(n)}\right) \frac{\partial \boldsymbol{z}^{(n)}}{\partial \boldsymbol{a}^{(n-1)}} \frac{\partial \boldsymbol{a}^{(n-1)}}{\partial \boldsymbol{z}^{(n-1)}} \frac{\partial \boldsymbol{z}^{(n-1)}}{\partial W^{(n-1)}} \tag{3.8}\\
&= \frac{\partial \varepsilon}{\partial \boldsymbol{y}} f'\left(\boldsymbol{z}^{(n)}\right) W^{(n)} f'\left(\boldsymbol{z}^{(n-1)}\right)\boldsymbol{a}^{(n-2)} \tag{3.9}
\end{align}$$

e assim, sucessivamente para as demais camadas da rede. Os gradientes em relação aos *biases* podem ser analogamente calculados.

[1] Em inglês, *learning rate*.

3.1.2 Aplicação: Problema de Classificação XOR

Vamos desenvolver uma MLP que faça a operação `xor` (ou exclusivo). I.e, receba como entrada dois valores lógicos A_1 e A_2 (V, verdadeiro ou F, falso) e forneça como saída o valor lógico $R = A_1 \texttt{xor} A_2$. Consultamos a seguinte tabela verdade:

A_1	A_2	R
V	V	F
V	F	V
F	V	V
F	F	F

Assumindo $V = 1$ e $F = -1$, podemos modelar o problema tendo entradas $\boldsymbol{x} = (x_1, x_2)$ e saída y como na seguinte tabela:

x_1	x_2	y
1	1	−1
1	−1	1
−1	1	1
−1	−1	−1

Modelo

Vamos usar uma MLP de estrutura $2-2-1$ e com funções de ativação $f^{(1)}(\boldsymbol{x}) = \tanh(\boldsymbol{x})$ e $f^{(2)}(\boldsymbol{x}) = id(\boldsymbol{x})$. Ou seja, nossa rede tem duas entradas, uma camada escondida com 2 unidades (função de ativação tangente hiperbólica) e uma unidade de saída (função de ativação identidade).

Treinamento

Para o treinamento, vamos usar a função erro quadrático médio

$$\varepsilon := \frac{1}{n_s}\sum_{s=1}^{n_s}\left|\tilde{y}^{(s)} - y^{(s)}\right|^2, \tag{3.10}$$

onde os valores estimados $\tilde{y}^{(s)} = \mathcal{N}\left(\boldsymbol{x}^{(s)}\right)$ e $\left\{\boldsymbol{x}^{(s)}, y^{(s)}\right\}_{s=1}^{n_s}$, $n_s = 4$, conforme na tabela acima.

Implementação

O seguinte código implementa a MLP e usa o Método do Gradiente Descendente (DG) no algoritmo de treinamento.

Código 3.1: mlp_xor.py

```python
import torch

# modelo

model = torch.nn.Sequential(
    torch.nn.Linear(2,2),
    torch.nn.Tanh(),
    torch.nn.Linear(2,1)
    )

# treinamento

## optimizador
optim = torch.optim.SGD(model.parameters(), lr=1e-2)

## função erro
loss_fun = torch.nn.MSELoss()

## dados de treinamento
X_train = torch.tensor([[1., 1.],
                        [1., -1.],
                        [-1., 1.],
                        [-1., -1.]])
y_train = torch.tensor([-1., 1., 1., -1.]).reshape(-1,1)

print("\nDados de treinamento")
print("X_train =")
print(X_train)
print("y_train = ")
print(y_train)

## num max épocas
nepochs = 5000
tol = 1e-3
```

```

for epoch in range(nepochs):

    # forward
    y_est = model(X_train)

    # erro
    loss = loss_fun(y_est, y_train)

    print(f'{epoch}: {loss.item():.4e}')

    # critério de parada
    if (loss.item() < tol):
        break

    # backward
    optim.zero_grad()
    loss.backward()
    optim.step()


# verificação
y = model(X_train)
print(f'y_est = {y}')
```

3.1.3 Exercícios

[[tag::construcao]]

3.2 Aplicação: Aproximação de Funções

Redes Perceptron Multicamadas (MLP) são aproximadoras universais. Nesta seção, vamos aplicá-las na aproximação de funções uni- e bidimensionais.

3.2.1 Função unidimensional

Vamos criar uma MLP para aproximar a função gaussiana

$$y = e^{-x^2}, \tag{3.11}$$

para $x \in [-1,1]$.

Código 3.2: mlp_gaussiana_1d.py

```
import torch
import matplotlib.pyplot as plt

# modelo

model = torch.nn.Sequential(
    torch.nn.Linear(1,25),
    torch.nn.Tanh(),
    torch.nn.Linear(25,1)
    )

# treinamento

## fun obj
fobj = lambda x: torch.exp(-x**2)
a = -1.
b = 1.

## optimizador
optim = torch.optim.SGD(model.parameters(),
                        lr=1e-2, momentum=0.9)

## função erro
loss_fun = torch.nn.MSELoss()

## num de amostras por época
ns = 100
## num max épocas
nepochs = 5000
## tolerância
tol = 1e-5

for epoch in range(nepochs):

    # amostras
    X_train = (a - b) * torch.rand((ns,1)) + b
    y_train = fobj(X_train)
```

```python

    # forward
    y_est = model(X_train)

    # erro
    loss = loss_fun(y_est, y_train)

    print(f'{epoch}: {loss.item():.4e}')

    # critério de parada
    if (loss.item() < tol):
        break

    # backward
    optim.zero_grad()
    loss.backward()
    optim.step()


# verificação
fig = plt.figure()
ax = fig.add_subplot()

x = torch.linspace(a, b,
                   steps=50).reshape(-1,1)

y_esp = fobj(x)
ax.plot(x, y_esp, label='fobj')

y_est = model(x)
ax.plot(x, y_est.detach(), label='model')

ax.legend()
ax.grid()
ax.set_xlabel('x')
ax.set_ylabel('y')
plt.show()
```

3.2.2 Função bidimensional

Vamos criar uma MLP para aproximar a função gaussiana

$$y = e^{-(x_1^2+x_2^2)}, \tag{3.12}$$

para $\boldsymbol{x} = (x_1, x_2) \in [-1,1]^2$.

Código 3.3: mlp_gaussiana_2d.py

```python
import torch
import matplotlib.pyplot as plt

# modelo

model = torch.nn.Sequential(
    torch.nn.Linear(2,50),
    torch.nn.Tanh(),
    torch.nn.Linear(50,25),
    torch.nn.Tanh(),
    torch.nn.Linear(25,5),
    torch.nn.Tanh(),
    torch.nn.Linear(5,1)
    )

# treinamento

## fun obj
a = -1.
b = 1.
def fobj(x):
    y = torch.exp(-x[:,0]**2 - x[:,1]**2)
    return y.reshape(-1,1)

## optimizador
optim = torch.optim.SGD(model.parameters(),
                        lr=1e-1, momentum=0.9)

## função erro
loss_fun = torch.nn.MSELoss()

## num de amostras por eixo por época
```

```python
ns = 100
## num max épocas
nepochs = 5000
## tolerância
tol = 1e-5

for epoch in range(nepochs):

    # amostras
    x0 = (a - b) * torch.rand(ns) + b
    x1 = (a - b) * torch.rand(ns) + b
    X0, X1 = torch.meshgrid(x0, x1)
    X_train = torch.cat((X0.reshape(-1,1),
                         X1.reshape(-1,1)),
                        dim=1)
    y_train = fobj(X_train)

    # forward
    y_est = model(X_train)

    # erro
    loss = loss_fun(y_est, y_train)

    print(f'{epoch}: {loss.item():.4e}')

    # critério de parada
    if (loss.item() < tol):
        break

    # backward
    optim.zero_grad()
    loss.backward()
    optim.step()


# verificação
fig = plt.figure()
ax = fig.add_subplot()

n = 50
```

```
x0 = torch.linspace(a, b, steps=n)
x1 = torch.linspace(a, b, steps=n)
X0, X1 = torch.meshgrid(x0, x1)
X = torch.cat((X0.reshape(-1,1),
               X1.reshape(-1,1)),
              dim=1)

y_esp = fobj(X)
Y = y_esp.reshape((n,n))
levels = torch.linspace(0., 1., 10)
c = ax.contour(X0, X1, Y, levels=levels, colors='white')
ax.clabel(c)

y_est = model(X)
Y = y_est.reshape((n,n))
ax.contourf(X0, X1, Y.detach(), levels=levels)

ax.grid()
ax.set_xlabel('x_1')
ax.set_ylabel('x_2')
plt.show()
```

3.2.3 Exercícios

[[tag::construcao]]

3.3 Aplicação: Equação de Laplace

Vamos criar uma MLP para resolver

$$-\Delta u = f, \quad \boldsymbol{x} \in D = (-1,1)^2, \tag{3.13a}$$

$$u = 0, \quad \boldsymbol{x} \in \partial D. \tag{3.13b}$$

Para validação, vamos considerar um problema com solução manufaturada

$$u(\boldsymbol{x}) = \operatorname{sen}(\pi x_1)\operatorname{sen}(\pi x_2) \tag{3.14}$$

o que nos fornece

$$f = \pi^2 \operatorname{sen}(\pi x_1) \operatorname{sen}(\pi x_2). \tag{3.15}$$

Código 3.4: mlp_eqlaplace.py

```
import torch
import matplotlib.pyplot as plt

# modelo

model = torch.nn.Sequential(
    torch.nn.Linear(2,500),
    torch.nn.Tanh(),
    torch.nn.Linear(500,250),
    torch.nn.Tanh(),
    torch.nn.Linear(250,50),
    torch.nn.Tanh(),
    torch.nn.Linear(50,1)
    )

# treinamento

## fun obj
a = -1.
b = 1.
def exact(x):
    y = torch.sin(torch.pi*x[:,0])*torch.sin(torch.pi*x[:,1])
    return y.reshape(-1,1)

def rhs(x):
    y = torch.pi**2*torch.sin(torch.pi*x[:,0])*torch.sin(torch.pi*x[:,
    return y.reshape(-1,1)

## optimizador
optim = torch.optim.SGD(model.parameters(),
                        lr=1e-2, momentum=0.85)
## scheaduler
scheduler = torch.optim.lr_scheduler.ReduceLROnPlateau(optim,
                                                       factor = 0.6,
                                                       min_lr = 1e-6,)
```

```

## num de amostras pts internos
n_in = 200
## num de amostras pts fronteira
n_bound = 50
## num max épocas
nepochs = 5000
## tolerância
tol = 1e-5
## output freq
eout = 100

def loss_fun(X_in, X_bound, model=model):

    ## pontos internos
    n_in = X_in.size(0)
    l_in = 0.
    for s in range(n_in):
        x = X_in[s:s+1,:].detach()
        x.requires_grad = True
        u = model(x)
        grad_u = torch.autograd.grad(u, x,
                                     create_graph = True,
                                     retain_graph = True)[0]
        u_x = grad_u[0,0]
        u_y = grad_u[0,1]

        u_xx = torch.autograd.grad(u_x, x,
                                   create_graph = True,
                                   retain_graph = True)[0][0,0]
        u_yy = torch.autograd.grad(u_y, x,
                                   create_graph = True,
                                   retain_graph = True)[0][0,1]
        l_in = torch.add(l_in, (u_xx + u_yy + rhs(x))**2)
    l_in /= n_in


    ## pontos de contorno
    n_bound = X_bound.size(0)
    l_bound = 0.
```

```
    for s in range(n_bound):
        x = X_bound[s:s+1,:]
        u = model(x)
        l_bound = torch.add(l_bound, u**2)
    l_bound /= n_bound

    return l_in + l_bound


# pts de fronteira
X_bound = torch.empty((4*n_bound, 2))
# pts internos
X_in = torch.empty((n_in, 2))

# épocas
for epoch in range(nepochs):

    # amostras: pts internos
    for s in range(n_in):
        X_in[s,:] = (a-b)*torch.rand(2) + b
    # amostras: pst fronteira
    s = 0
    for i in range(n_bound):
        # a <= x0 <= b, x1 = 0
        X_bound[s,0] = (a-b)*torch.rand(1) + b
        X_bound[s,1] = a
        s += 1

        # x0 = b, a <= x1 <= b
        X_bound[s,0] = b
        X_bound[s,1] = (a-b)*torch.rand(1) + b
        s += 1

        # x0 = a, a <= x1 < b
        X_bound[s,0] = a
        X_bound[s,1] = (a-b)*torch.rand(1) + b
        s += 1

        # a <= x0 <= b, x1 = b
        X_bound[s,0] = (a-b)*torch.rand(1) + b
```

```
        X_bound[s,1] = b
        s += 1

    # erro
    loss = loss_fun(X_in, X_bound)

    print(f'{epoch}: loss = {loss.item():.4e}, lr = {optim.param_groups[0]["
    if ((epoch+1) % eout == 0):
        # verificação
        fig = plt.figure()
        ax = fig.add_subplot()

        ns = 50
        x0 = torch.linspace(a, b, steps=ns)
        x1 = torch.linspace(a, b, steps=ns)
        X0, X1 = torch.meshgrid(x0, x1)
        X = torch.cat((X0.reshape(-1,1),
                       X1.reshape(-1,1)),
                      dim=1)

        y_esp = exact(X)
        Y = y_esp.reshape((ns,ns))
        c = ax.contour(X0, X1, Y, levels=10, colors='white')
        ax.clabel(c)

        y_est = model(X)
        Y = y_est.reshape((ns,ns))
        cf = ax.contourf(X0, X1, Y.detach(), levels=10, cmap='coolwarm')
        plt.colorbar(cf)

        # amostras
        ax.plot(X_in[:,0].detach(), X_in[:,1].detach(), ls='', marker='*', c
        ax.plot(X_bound[:,0].detach(), X_bound[:,1].detach(), ls='', marker=

        ax.grid()
        ax.set_xlabel('$x_1$')
        ax.set_ylabel('$x_2$')
        plt.show()
```

```
    # critério de parada
    if (loss.item() < tol):
        break

    # backward
    optim.zero_grad()
    loss.backward()
    optim.step()
    scheduler.step(loss)
```

3.3.1 Exercícios

[[tag::construcao]]

Resposta dos Exercícios

Exercício 2.1.1. Dica: verifique que sua matriz hessiana é positiva definida.

Exercício 2.1.2. Dica: consulte a ligação Notas de Aula: Matemática Numérica: 7.1 Problemas lineares.

Exercício 2.2.1. $(\tanh x)' = 1 - \tanh^2 x$

Bibliografia

[1] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, Cambridge, MA, 2016.