Tutorial Class 1: Backpropagation and Gradient Descent

Exercice 1: Derivative

Question 1 :

Considering a binary classification problem, define formally the MLP function f such that we get two layers producing a 1 dimensional representation (the output layer is the second layer). The input of the neural network is a scalar $x \in \mathbb{R}$ and output $y \in \{0, 1\}$. We consider σ as the activation function and no bias. Explicit the different variables of the function.

Question 2 :

Considering a dataset $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$ with $y_i \in \mathcal{Y}$ is 1 (if x_i belong to the class 1) or 0 (if x_i belong to the class 0). Explicitly define the error $E(\mathcal{D}, \Theta)$ function considering the log-likelihood loss for a binary classification problem. Notably, $p_\theta(y_i = 1|x_i)$ is modeled by $\sigma(f_\theta(x_i))$ with σ the sigmoid function.

Question 3 :

Explicitly provide the gradient of $\mathcal{L}(f(x), y)$ according to the different parameters.

Exercise 2: Computational graph (Optional)

Let f be the following function :

$$\begin{aligned} f &: \mathbb{R} \to \mathbb{R} \\ & x \mapsto a^{(3)}\sigma\left(\alpha a^{(1)}x + (1-\alpha)a^{(2)}(x)\right) \end{aligned}$$

With $a^{(i)} \in \mathbb{R}$, σ the sigmoid function and α a constant.

Question 1 :

What operations will you need to build the graph of this function?

Question 2 :

Draw the computational graph

Exercise 3: Backpropagation

Let consider the following variables:

- $\mathcal{D} \subset \mathbb{R}^N \times \mathbb{R}^M$ and $\mathcal{D}$ is a finite set
- We denote $(x^{(i)}, y^{(i)})$ the i^{th} element of the dataset $\mathcal{D}$
- $x^{(i)} \in \mathbb{R}^N$ and $y^{(i)} \in \{0,1\}^M$, $\sum_{j=1}^{M} y_j^{(i)} = 1$ and $c^{(i)}$ being the index of the non zero component.
- $W_1 \in \mathbb{R}^{Z \times N}$ and $b_1 \in \mathbb{R}^Z$ and $W_2 \in \mathbb{R}^{M \times Z}$ and $b_2 \in \mathbb{R}^M$

Let f_θ with $\theta = \{W_1, W_2, b_1, b_2\}$ and the neural network function defined by:

$$\begin{aligned} f_\theta \colon \mathbb{R}^N &\to \mathbb{R}^M \\ x^{(i)} &\mapsto W_2 ReLU\left(W_1 x^{(i)} + b_1\right) + b_2 \end{aligned}$$

The function $ReLU$ is defined by :

$$ReLU(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

We denote $o^{(i)} = f_\theta(x^{(i)})$, and the error function to minimize is given by:

$$E(\mathcal{D}, \theta) = \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \mathcal{L}(o^{(i)}, y^{(i)})$$

And $\mathcal{L}(o^{(i)}, y^{(i)})$ by:

$$\mathcal{L}(o^{(i)}, y^{(i)}) = -\sum_{j=1}^{M} y_j^{(i)} \ln\left(p_\theta(y_j | x^{(i)})\right) = -\sum_{j=1}^{M} y_j^{(i)} \ln\left(\frac{e^{o_j^{(i)}}}{\sum_{k=1}^{M} e^{o_k^{(i)}}}\right)$$

Question 1 :

Provide the formula of $\nabla_{o^{(i)}} \mathcal{L}(o^{(i)}, y^{(i)})$

Question 2 :

What is the expression of $\nabla_{W_2} \mathcal{L}(o^{(i)}, y^{(i)})$ and $\nabla_{b_2} \mathcal{L}(o^{(i)}, y^{(i)})$ considering $z^{(i)} = ReLU\left(W_1 x^{(i)} + b_1\right)$ and $\nabla_{o^{(i)}} \mathcal{L}(o^{(i)}, y^{(i)})$ known

Question 3 :

What is the expression of $\nabla_{W_1} \mathcal{L}(o^{(i)}, y^{(i)})$ and $\nabla_{b_1^{(i)}} \mathcal{L}(o^{(i)}, y^{(i)})$

Question 4 :

Define one step of gradient descent and how each weight update with W_i^t being previous weigths (resp b_i^t) and W_i^{t+1} being the updated weights (resp b_i^{t+1})

Question 5 :

Draw a graph of the neural network and define for each node the attributes, the forward function and the backward function inputs and outputs.