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.

Solution:

The objective of this question is to define the function modeled by the neural network.

Let $a^{(1)} \in \mathbb{R}^1$, $a^{(2)} \in \mathbb{R}^1$ the parameters. The neural network is modeled by the following function:

$$\begin{aligned} f &: \mathbb{R} \to \mathbb{R} \\ & x \mapsto (a^{(2)}(\sigma(a^{(1)}x))) \end{aligned}$$

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.

Solution: In this solution we retrieve the formula of the course from the general formula of NLL. We consider a binary classification problem. We recall that loglikelihood is defined by:

$$NLL(x, y) = -\sum_i y_i log(p_\theta(y_i|x))$$

Thus in our case as $y \in \{0, 1\}$ then :

$$\begin{aligned} NLL(x, y) &= -y_i log(p_\theta(y = 1|x)) - (1 - y_i) log(p_\theta(y = 0|x)) \\ &= -y_i log(p_\theta(y = 1|x)) - (1 - y_i) log(1 - p_\theta(y = 1|x)) \end{aligned}$$

Where $p(y|x)$ is modeled with :

$$p(y = 1|x) = \frac{1}{1 + e^{-f(x)}}$$

$$
\begin{aligned}
NLL(x,y) &= -\left[y_i log(\frac{1}{1+e^{-f(x)}}) + (1-y_i) log(1-\frac{1}{1+e^{-f(x)}})\right] \\
&= -\left[\ y_i log(1) - y_i log(1+e^{-f(x)}) + (1-y_i) log\left(\frac{1+e^{-f(x)}-1}{1+e^{-f(x)}}\right)\right] \\
&= -\left[-y_i log(1+e^{-f(x)}) + (1-y_i) log\left(\frac{e^{-f(x)}}{1+e^{-f(x)}}\right)\right] \\
&= -\left[y_i log(1+e^{-f(x)}) + log\left(\frac{e^{-f(x)}}{1+e^{-f(x)}}\right) - y_i log\left(\frac{e^{-f(x)}}{1+e^{-f(x)}}\right)\right] \\
&= -\left[-y_i log(1+e^{-f(x)}) + log\left(\frac{e^{-f(x)}}{1+e^{-f(x)}}\right) + y_i log(1+e^{-f(x)}) + y_i f(x)\right] \\
&= -\left[log\left(\frac{e^{-f(x)}e^{f(x)}}{(1+e^{-f(x)})e^{f(x)}}\right) + y_i f(x)\right] \\
&= -\left[log\left(\frac{e^{f(x)-f(x)}}{(e^{f(x)}+e^{f(x)-f(x)})}\right) + y_i f(x)\right] \\
&= -\left[log\left(\frac{1}{e^{f(x)}+1}\right) + y_i f(x)\right] \\
&= -\left[-log\left(e^{f(x)}+1\right) + y_i f(x)\right] \\
&= -y_i f(x) + log\left(e^{f(x)}+1\right)
\end{aligned}
$$

Question 3 :

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

Solution: We first want the gradient $\nabla_{f(x)}\mathcal{L}(f(x), y)$ which is

$$\frac{\partial - y_i o + log\left(e^o + 1\right)}{\partial o} = -y_i + \frac{e^o}{1+e^0}$$

The gradient according to $a^{(2)}$ is thus $\frac{\partial - y_i o + log(e^o+1)}{\partial o}\frac{o}{a^2} = \frac{L}{\partial o}\frac{o}{a^2} = \frac{L}{\partial o}z^{(2)}$ with $z^{(2)} = \sigma(a^{(1)}x)$

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 ($\sigma(x) = \frac{1}{1+e^{-x}}$) and α a constant.

Question 1 :

What operations will you need to build the graph of this function? Define the functions and it's partial derivative.

Solution:

Considering low granularity you only need :

Multiplication

- $f(x, y) = x \times y$
- $\frac{\partial f(x,y)}{\partial x} = y$
- $\frac{\partial f(x,y)}{\partial y} = x$

Addition

- $f(x, y) = x + y$
- $\frac{\partial f(x,y)}{\partial x} = 1$
- $\frac{\partial f(x,y)}{\partial y} = 1$

Power

- $f(x, y) = x^y$
- $\frac{\partial f(x,y)}{\partial x} = yx^{y-1}$
- $\frac{\partial f(x,y)}{\partial y} = x^y ln(x)$

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 $\nabla_{o^{(i)}} \mathcal{L}(o^{(i)}, y^{(i)})$

Solution: Lets find the derivative of $\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial o_l^{(i)}}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial o_l^{(i)}} &= -\sum_{j=1}^{M} \frac{y_j^{(i)} \ln\left(\frac{e^{o_j^{(i)}}}{\sum_{k=1}^{M} e^{o_k^{(i)}}}\right)}{\partial o_l^{(i)}} \\
&= -\sum_{j=1}^{M} \frac{y_j^{(i)} \ln\left(e^{o_j^{(i)}}\right)}{\partial o_l^{(i)}} + \sum_{j=1}^{M} \frac{y_j^{(i)} \ln\left(\sum_{k=1}^{M} e^{o_k^{(i)}}\right)}{\partial o_l^{(i)}} \\
&= -y_l^{(i)} + \sum_{j=1}^{M} \frac{y_j^{(i)} \ln\left(\sum_{k=1}^{M} e^{o_k^{(i)}}\right)}{\partial o_l^{(i)}} \\
&= -y_l^{(i)} + \sum_{j=1}^{M} \frac{y_j^{(i)} e^{o_l^{(i)}}}{\sum_{k=1}^{M} e^{o_k^{(i)}}}
\end{aligned}$$

In our case, only $y_c = 1$ then it becames

$$\begin{aligned}
\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial o_l^{(i)}} &= -\mathbb{1}_{c=l} + \frac{e^{o_l^{(i)}}}{\sum_{k=1}^{M} e^{o_k^{(i)}}} \\
&= -\mathbb{1}_{c=l} + softmax(o^{(i)})_l
\end{aligned}$$

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

Solution: Lets find the derivative of $\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial W_{n,m}^{(2)}}$:

$$\begin{aligned}\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial W^{(2)}_{n,m}} &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} \frac{\partial o^{(i)}_l}{\partial W^{(2)}_{n,m}} \\ &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} \frac{\partial \sum_{j=1}^{Z} W^{(2)}_{l,j} z^{(i)}_j + b^{(2)_n}}{\partial W^{(2)}_{n,m}} \\ &= \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_n} z^{((i))}_m\end{aligned}$$

Thus $\nabla_{W_2} \mathcal{L}(o^{(i)}, y^{(i)}) = \nabla_{o^{(i)}} \mathcal{L}(o^{(i)}, y^{(i)}).z^{(i)\intercal}$

$$\begin{aligned}\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial b^{(2)}_n} &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} \frac{\partial o^{(i)}_l}{\partial b^{(2)}_n} \\ &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} \frac{\partial \sum_{j=1}^{Z} W^{(2)}_{l,j} z^{(i)}_j + b^{(2)_n}}{\partial b^{(2)}_n} \\ &= \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_n}\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{L}(o^{(i)}, y^{(i)})}{\partial z^m} &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} \frac{\partial o^{(i)}_l}{\partial z^m} \\ &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} \frac{\partial \sum_{j=1}^{Z} W^{(2)}_{l,j} z^{(i)}_j + b^{(2)_n}}{\partial z^m} \\ &= \sum_{l=1}^{Z} \frac{\mathcal{L}(o^{(i)}, y^{(i)})}{\partial o^{(i)}_l} W_{l,n}\end{aligned}$$

Question 3 :

What is the expression of $\nabla_{W_1} \mathcal{L}(o^{(i)}, y^{(i)})$ and $\nabla_{b^{(i)}_1} \mathcal{L}(o^{(i)}, y^{(i)})$

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.