TD 3

Exercice 1 – Learning an XOR Network

The XOR (exclusive or) problem is as follows : the points $(-1,-1),(1,1)$ are considered negative, and the points $(-1,1),(1,-1)$ are positive.

Q 1.1 Draw a neural network with 2 hidden neurons for these data. List some possible activation functions. Which ones are the most appropriate in this case for the different layers ?

Q 1.2 Propose values for the weights of the network. Is the solution unique ?

Q 1.3 Same question for an 8-square checkerboard.

Exercice 2 – Characterization of the Solution Learned by a Neural Network

Consider a network with a single hidden layer parameterized by the vector $\mathbf{w}$. Let $f_{\mathbf{w}}(\mathbf{x})$ denote the output for an input $\mathbf{x}$.
We will use the following notations :

- A sample $\mathbf{x}^i = \{x^i_j\}_{j=1,\dots,d}$, its label y^i, and a training set $\mathcal{D} = \{(\mathbf{x}^i, y^i)\}_{i=1,\dots,N}$;
- The weights towards the hidden layer are $\mathbf{w}^1 = \{w^1_{jh}\}_{j=1,\dots,d,\, h=1,\dots,H}$, the weights towards the output layer are $\mathbf{w}^s = \{w^s_{hk}\}_{h=1,\dots,H,\, k=1,\dots,K}$.
- The activation functions g^1, g^s for the two layers.

Q 2.1 How many hidden neurons does the network have ? How many outputs ? Draw the network. What does it mean for the number of outputs to be greater than one ?

Q 2.2 Express the output $f_{\mathbf{w}}(\mathbf{x})$ in terms of the components of $\mathbf{x}$ and $\mathbf{w}$.

Q 2.3 Give the expression for the cost (least squares) in terms of the training set $\mathcal{D}$. What is its theoretical formulation (using the expectation of a quantity) ?

Q 2.4 Show that at each $\mathbf{x}$, the optimal solution corresponds to $f^\star(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$. What does this result mean ?

Q 2.5 For multi-class classification, the output is a vector : $\mathbf{y} = [\dots, 1, \dots]$ with a 1 in the k-th position if the class of $\mathbf{x}$ is k. What is $f^\star_k(\mathbf{x})$ an approximation of in this case ?

Q 2.6 In the case of regression, what does $f^\star(\mathbf{x})$ correspond to ? Give a graphical example of noisy 1D regression where several values of y correspond to a given $\mathbf{x}$.

Q 2.7 Decomposition and interpretation of the cost.

- Rewrite the cost criterion at a point x to involve the terms $y - f^\star(\mathbf{x})$ and $f^\star(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})$, then $E_{y|\mathbf{x}}\left[\|y - f^\star(\mathbf{x})\|^2\right]$.
- Provide an interpretation of what this term represents, as well as the other terms you have introduced. Why does learning not always result in a zero cost ?

Q 2.8 Is the solution obtained by gradient descent unique ? Why ? What does it depend on ?

Exercice 3 (8 points) – Highway to Gradient

The `Highway Network` architecture was proposed in 2015 specifically for very deep networks dedicated to image processing. A layer of this type of network is very similar to a classical network layer, but it mixes the layer inputs with the layer outputs.

For example, consider a non-linear fully connected layer of a classical network defined by the function $H(\mathbf{x}, \mathbf{W}_H) = \sigma(\mathbf{W}_H^t \mathbf{x} + b_H)$. The Highway Network uses a transformation $T(\mathbf{x}, \mathbf{W}_T) = \sigma(\mathbf{W}_T^t \mathbf{x} + b_T)$ to mix the input $\mathbf{x}$ and the usual output of the layer $H(\mathbf{x}, \mathbf{W}_H)$. The output $\mathbf{y}$ of the layer is (where $\odot$ denotes the element-wise product) :

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W}_H) \odot T(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \odot (1 - T(\mathbf{x}, \mathbf{W}_T))$$

Q 3.1 Suppose the input $\mathbf{x}$ is of dimension d : $\mathbf{x} \in \mathbb{R}^d$. According to the definition, what are the dimensions of $\mathbf{W}_H, \mathbf{W}_T, \mathbf{y}$? (Assume the biases b_H and b_T are scalars in $\mathbb{R}$).

Q 3.2 Compute the derivative of the sigmoid function. In the following, you can denote it as $\sigma'(x)$ without expanding. Recall : $\sigma(x) = \frac{1}{1+e^{-x}}$

Q 3.3 A network is assumed to be composed first of a Highway Network layer, denoted $\mathbf{z}$, followed by a linear layer M with a sigmoid activation function, and a least squares cost :

- $\mathbf{z} = H(\mathbf{x}, \mathbf{W}_H) \odot T(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \odot (1 - T(\mathbf{x}, \mathbf{W}_T))$
- $\hat{\mathbf{y}} = \sigma(\mathbf{W}_M^t \mathbf{z} + b_M)$
- $L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2$

Q 3.3.1 Suppose $\mathbf{y} \in \mathbb{R}^p$. What should be the dimensions of $\mathbf{W}_M$ and $\hat{\mathbf{y}}$? (Assume the bias b_M is a scalar in $\mathbb{R}$).

Q 3.3.2 Compute $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \hat{\mathbf{y}}_i}$, the derivative of the cost with respect to the i-th output of the network.

Q 3.3.3 Compute $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial w_{i,j}^M}$ for a weight $w_{i,j}^M$ of $\mathbf{W}_M$.

Q 3.3.4 Compute $\delta_i = \frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial z_i}$ for the i-th output $\mathbf{z}$ of the Highway layer.

Q 3.3.5 Compute for a weight $w_{i,j}^T$ the derivative $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial w_{i,j}^T}$.

Q 3.3.6 Compute for a weight $w_{i,j}^H$ the derivative $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial w_{i,j}^H}$.

Q 3.3.7 Provide the optimization algorithm for the network.

Q 3.4 (bonus) In your opinion, what problem(s) can a Highway network solve and why ?