Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne Winter term 23/24

Mathematics of Learning – Worksheet 11

- The exercise sheets will be uploaded every Monday. Solution sketches will be uploaded one week later.
- You can hand in your own solutions via StudOn and we correct them - this is not mandatory. Please hand in small groups of 2-3 students.
- For questions, please use the forum on StudOn since other students may have similar questions. If you have a more personal question about the exercises please send an email to ehsan.waiezi@fau.de or lars.weidner@fau.de respectively.

Exercise 1 [More about neural networks - theoretical and difficult].
Consider F_I^L the set of Lipschitz continuous functions mapping from a compact real interval I to the real numbers. Lipschitz continuous means,

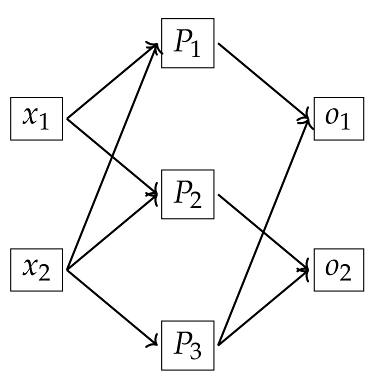
$$\exists L > 0 : |f(x) - f(y)| \leq L|x - y| \text{ for all } x, y \in I.$$

Consider the set F_I of functions which can be linearly combined of right shifted Heavyside step functions, i.e.

$$F_I = \{f : I \to \mathbb{R} : \ \exists n \in \mathbb{N}, \lambda, b \in \mathbb{R}^n : f(x) = \sum_{i=1}^{n} \lambda_i H(x + b_i) \text{ for all } x \in I\}.$$

H denotes the Heavyside step function, i.e., $H(x) = 1$ if x is nonnegative and 0 otherwise.
Prove or disprove: The closure, corresponding to the $||\cdot||_\infty$ norm on functions, of F_I, is a superset of F_I^L.

Exercise 2 [Neural networks - Calculations].
Given the following network:



The initial weights are all 1, the biases of the output layers are 0, the biases of P_i are 0. Activation functions for all layers are $\psi(t) := \frac{1}{1+e^{-t}}$. The input points are
$X = \left\{ \begin{pmatrix} 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 7 \\ -2 \end{pmatrix}, \begin{pmatrix} 8 \\ 4 \end{pmatrix}, \begin{pmatrix} -4 \\ -2 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \end{pmatrix} \right\}, Y = \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}.$

1. Look at the data and formulate a guess what the targets Y express.
2. Do feedforward passes for all given data points.
3. Do 2 or 3 rounds of backpropagation passes (training) with the neural network, using either "pure" stochastic gradient descent (with batch size 1) or stochastic gradient descent with a larger batch size. (If you do not know about stochastic gradient descent methods, wait until the lecture on July 05)
4. Generate a few test data points on your own and check the classification accuracy.
5. Propose an alternative, smaller network architecture, which does the same.

Exercise 3 [Proof Exercise: Discriminatory Activation Functions].
Let $\Omega \subset \mathbb{R}^d$ be a compact set. A continuous function $\psi : \mathbb{R} \to \mathbb{R}$ is called discriminatory if for any signed measure $\mu \in \mathfrak{M}(\Omega)$ it holds

$$\int_\Omega \psi(w \cdot x + b) \,\mathrm{d}\mu(x) = 0, \quad \forall w \in \mathbb{R}^d,\, b \in \mathbb{R} \implies \mu = 0.$$

- Prove that the polynomials $\psi(t) = 1$ and $\psi(t) = t$ are *not discriminatory* by finding non-zero signed measures $\mu \in \mathfrak{M}([-1, 1])$ with

  $$\int_{-1}^{1} \psi(wx + b) \,\mathrm{d}\mu(x) = 0, \quad \forall w, b \in \mathbb{R}.$$

- Using the general convergence theorem from the lecture, prove that polynomials are *not discriminatory*.

  Hint: Characterize the approximation space

  $$\Sigma_d(\psi) = \left\{ \sum_{i=1}^{N} \alpha_i \psi(w_i \cdot x + b_i) \;:\; N \in \mathbb{N},\, w_i \in \mathbb{R}^d,\, b_i \in \mathbb{R} \right\}$$

  in the case that ψ is a polynomial of degree $p \in \mathbb{N}$.

  It suffices to argue in the one-dimensional case $d = 1$.

- Prove that $\mathrm{ReLU}(t) := \max(t, 0)$ is a discriminatory activation.

  Hint: You can cleverly combine two ReLU functions into a sigmoidal function, as defined in the lecture.