Machine Learning: **I Session 31/01/2023**

Exercise 1 [8 points]

Consider the problem of supervised learning.

1. Describe all the components of the formal statistical learning model (domain set, ...) and the general goal of supervised learning.
2. Formally define when a training set is ε-representative.
3. *Provide* and *prove* the upper bound to the generalization error $L_{\mathcal{D}}(h_S)$ of the hypothesis h_S picked by empirical risk minimization when the training set is $\frac{\varepsilon}{2}$-representative (briefly motivating all the steps of the proof).

Machine Learning: **I Session 31/01/2023**

Exercise 2 [8 points]

1. Describe the stochastic gradient descent (SGD) algorithm (in general).
2. What is the main advantage of SGD with respect to the gradient descent (GD) algorithm?
3. Consider a regression problem where the training data is $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i = [x_{i,1}, x_{i,2}] \in \mathbb{R}^2$ and $y_i \in \mathbb{R}$ for $i = 1, \dots, m$.

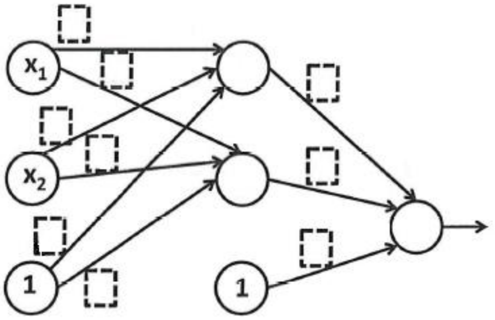
   Assume the hypothesis class $\mathcal{H}$ is given by (simplified) linear models, that is, $\mathcal{H}$ is defined as $\mathcal{H} = \{h_{\mathbf{w}} : h_{\mathbf{w}}(\mathbf{x}) = w_1 x_1 + w_2 x_2, \mathbf{w} = [w_1, w_2] \in \mathbb{R}^2\}$, with $\mathbf{x} = [x_1, x_2]$. Assume that the loss function is $\ell(h_{\mathbf{w}}, (\mathbf{x}, y)) = (h_{\mathbf{w}}(\mathbf{x}) - y)^2$, and that SGD is used to learn a model from the training data S.

   Write the SGD update when the hypothesis class $\mathcal{H}$ is as described above.

Exercise 3 [8 points]

Consider neural networks (NNs) for classification with $0-1$ loss.

1. Describe what you need to fix to define the hypothesis class of a NN, and what is instead learned from data.
2. Let $\mathbf{x} = [x_1, x_2]$, with $x_1, x_2 \in \{-1, 1\}$. Consider the NN in the figure below, where the activation function for each hidden node and the output node is the *sign* function. Assume that the weights are constrained to be in the set $\{-1, 0, 1\}$. You want your network to represent the function f that is 1 when the input is $[1, 1]$ or $[-1, -1]$ (for all other inputs the function f is -1).
   (a) Find the network's weights so that the network represents the function f described above. *Write the weights in the dashed boxes in the figure.*
   (b) Briefly describe the procedure you used to find the weights.



Exercise 4 [8 points]

Consider the clustering problem.

1. Briefly describe *linkage-based clustering*: what is the input, what is the output, the general algorithm it employs, and a termination condition.
2. Consider *single linkage* clustering. Describe how it is obtained by the general *linkage-based clustering* (no pseudocode needed).
3. Show the output of single linkage clustering when the input is given by the points in $\mathbb{R}^2$ shown as crosses below and the termination condition is given by having the points partitioned in $k = 2$ clusters. (You can draw directly in the figure below.) Briefly describe how the algorithm reaches such output.