WRITE FIRST NAME, LAST NAME, AND ID NUMBER ("MATRICOLA") BELOW AND READ ALL INSTRUCTIONS BEFORE STARTING WITH THE EXAM! TIME: 1.5 hours.

FIRST NAME: ....

LAST NAME: ....

ID NUMBER: ....

INSTRUCTIONS

- solutions to exercises must be in the appropriate spaces, that is:
  - Exercise 1: pag. 1, 2
  - Exercise 2: pag. 3, 4
  - Exercise 3: pag. 5, 6,
  - Exercise 4: pag. 7, 8, 9

  Solutions written outside the appropriate spaces (including other papersheets) will not be considered.
- the use of notes, books, or any other material is forbidden and will make your exam invalid;
- electronic devices (smartphones, calculators, etc.) must be turned off; their use will make your exam invalid;
- this booklet must be returned in its entirety;
- if your solution describes material not related to the question, you will get negative points for your answer.

Exercise 1 [8 points]

Consider the *regression* problem with *squared loss*.

1. Provide a formal definition of the problem, describing: the data, the learner's input, the learner's output, the loss function, the (most general) assumed generative model for the data, the learner's goal, and what choices the learner has to make.

2. Assume the hypothesis class is $\mathcal{H}$, the data generative distribution is $\mathcal{D}$, and the training data is S. Provide the definition of the training error $L_S(h)$ and of generalization error $L_{\mathcal{D}}(h)$, where $h \in \mathcal{H}$, and prove that for each $h \in \mathcal{H}$, the expectation (over the distribution of the training set) of $L_S(h)$ is equal to $L_{\mathcal{D}}(h)$, justifying all steps of the proof.

[Solution: Exercise 1]

[Solution: Exercise 1]

Exercise 2 [8 points]

Consider a binary classification problem where the training data is $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ for $i = 1, \dots, m$.

Assume the hypothesis class $\mathcal{H}$ is given by linear models with parameters $\mathbf{w} \in \mathbb{R}^d$, that the loss function is $\ell(h_{\mathbf{w}}, (\mathbf{x}, y)) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}$, and that SGD is used to learn a model from the training data S.

Write the SGD algorithm, and derive its update, when the hypothesis class $\mathcal{H}$ is as described above.

[Solution: Exercise 2]

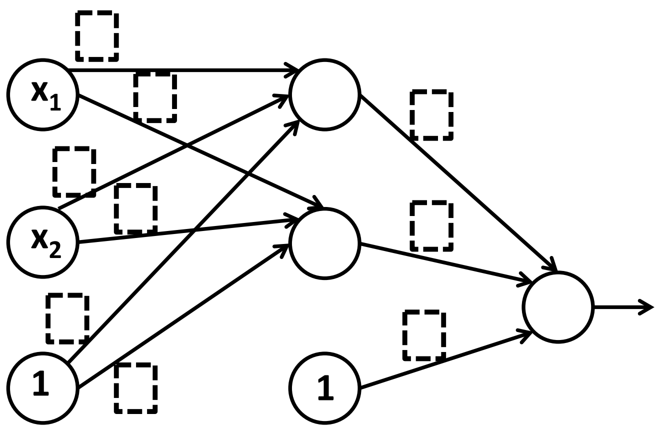
[Solution: Exercise 2]

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).

   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.*



[Solution: Exercise 3]

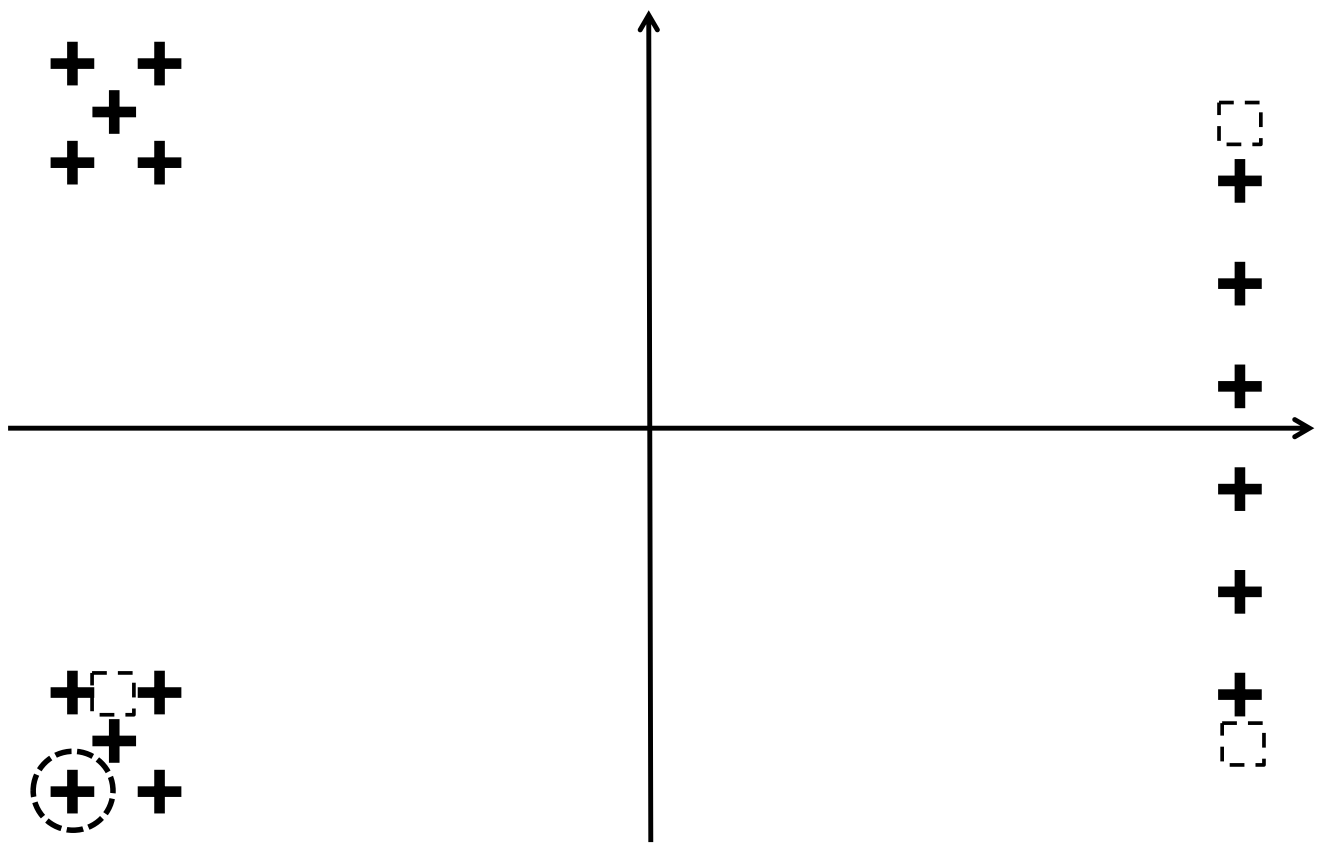
[Solution: Exercise 3]

Exercise 4 [8 points]

Consider the *k-means* clustering problem.

1. Describe Lloyd's algorithm.
2. Describe the `k-means++` algorithm for the initialization of the centers.
3. Consider the points in $\mathbb{R}^2$ shown as crosses below. Plot *in the figure below*:
   (a) the output of Lloyd's algorithm with $k = 3$ when the initial centers are the dashed squares;
   (b) the *most likely other centers* chosen by `k-means++`, and the order in which they are chosen, when the first center chosen is the point circled in the figure;
   (c) the output of Lloyd's algorithm with $k = 3$ when the initial centers are chosen by `k-means++` according to your answer to point (b).

   Briefly motivate your plots.



[Solution: Exercise 4]

[Solution: Exercise 4]