WRITE FIRST NAME, LAST NAME, AND ID NUMBER (“MATRICOLA”) BELOW AND READ ALL INSTRUCTIONS BEFORE STARTING WITH THE EXAM! TIME: 2.5 hours.

FIRST NAME: ....

LAST NAME: ....

ID NUMBER: ....

INSTRUCTIONS

- solutions to exercises must be in the appropriate spaces, that is:
  - Exercise 1: pag. 1, 2, 3
  - Exercise 2: pag. 4, 5
  - Exercise 3: pag. 6, 7, 8
  - Exercise 4: pag. 9, 10, 11, 12

  Solutions written outside the appropriate spaces (including other paper-sheets) 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.

Exercise 1 [8 points]

In the context of supervised learning:

1. provide the definition of the regression task
2. consider the following model class that is linear in the parameter:

$$h(x) := \mathbf{w}^\top \Psi(x) \quad \Psi(x) = [\psi_1(x), .., \psi_L(x)]^\top \quad x \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^L$$

   where $\Psi(x) = [\psi_1(x), .., \psi_L(x)]^\top$ can be a generic function, e.g., recall the polynomial regression case where $\Psi(x) = [1, x, x^2, ..., x^{L-1}]^\top$. Write the explicit expression of the least squares estimator of $\mathbf{w}$ given data (x_k, y_k), $k = 1, ..m$.

3. Recalling the answer to the previous question, consider the one-hidden-layer neural network

$$h(x) := \sum_{i=1}^{L} w_i \sigma(\alpha_i (x - \beta_i)) \quad x \in \mathbb{R}$$

   where $\alpha_i, w_i, \beta_i,\ i = 1, .., L$, are the network parameters. Show that for α_i and β_i fixed, the optimal w_i can be found in closed form under the square loss.

[Solution: Exercise 1]

[Solution: Exercise 1]

[Solution: Exercise 1]

Exercise 2 [8 points]

Consider a generic machine learning problem and assume that a regularized loss function has been used by the selected algorithm A. In the loss function the relevance of the regularization term is controlled by a parameter λ. Let us denote with h_A the solution found by algorithm A and with $L_S(h_A)$ its empirical risk while the true risk (generalization error) of h_A is $L_D(h_A)$.

1. Which is the impact of the λ parameter on the empirical risk $L_S(h_A)$ of the solution found by A?

2. Which is the expected behavior of the true risk $L_D(h_A)$ of the found solution as a function of the λ parameter?

3. Describe how the behavior of the empirical risk and of the true risk in the answers to the previous questions are related to the bias-complexity trade-off.

[Solution: Exercise 2]

[Solution: Exercise 2]

Exercise 3 [8 points]

Consider a classification problem with 0-1 loss.

1. Provide the definition of VC dimension $VCdim(\mathcal{H})$ of a hypothesis set $\mathcal{H}$, and of empirical error and true risk (generalization error) for an arbitrary hypothesis $h \in \mathcal{H}$. What is the relation between the empirical error and the true risk in terms of the VC dimension of $\mathcal{H}$?

2. Consider the hypothesis set $\mathcal{H}$ defined as: $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b} : \mathbb{R} \mapsto \{0, 1\}$ is

$$h_{a,b}(x) = \begin{cases} 1 & \text{if } x \leq a \; OR \; x \geq b \\ 0 & \text{otherwise} \end{cases}$$

   What's the value of $VCdim(\mathcal{H})$? Provide a proof of your claim.

3. Assume that you have many hypothesis sets, denoted by $\mathcal{H}_i, i = 1, 2, \dots, n$. Describe one strategy to choose a good hypothesis set $\mathcal{H}_i$ and a good model $\hat{h}_i \in \mathcal{H}_i$.

[Solution: Exercise 3]

[Solution: Exercise 3]

[Solution: Exercise 3]

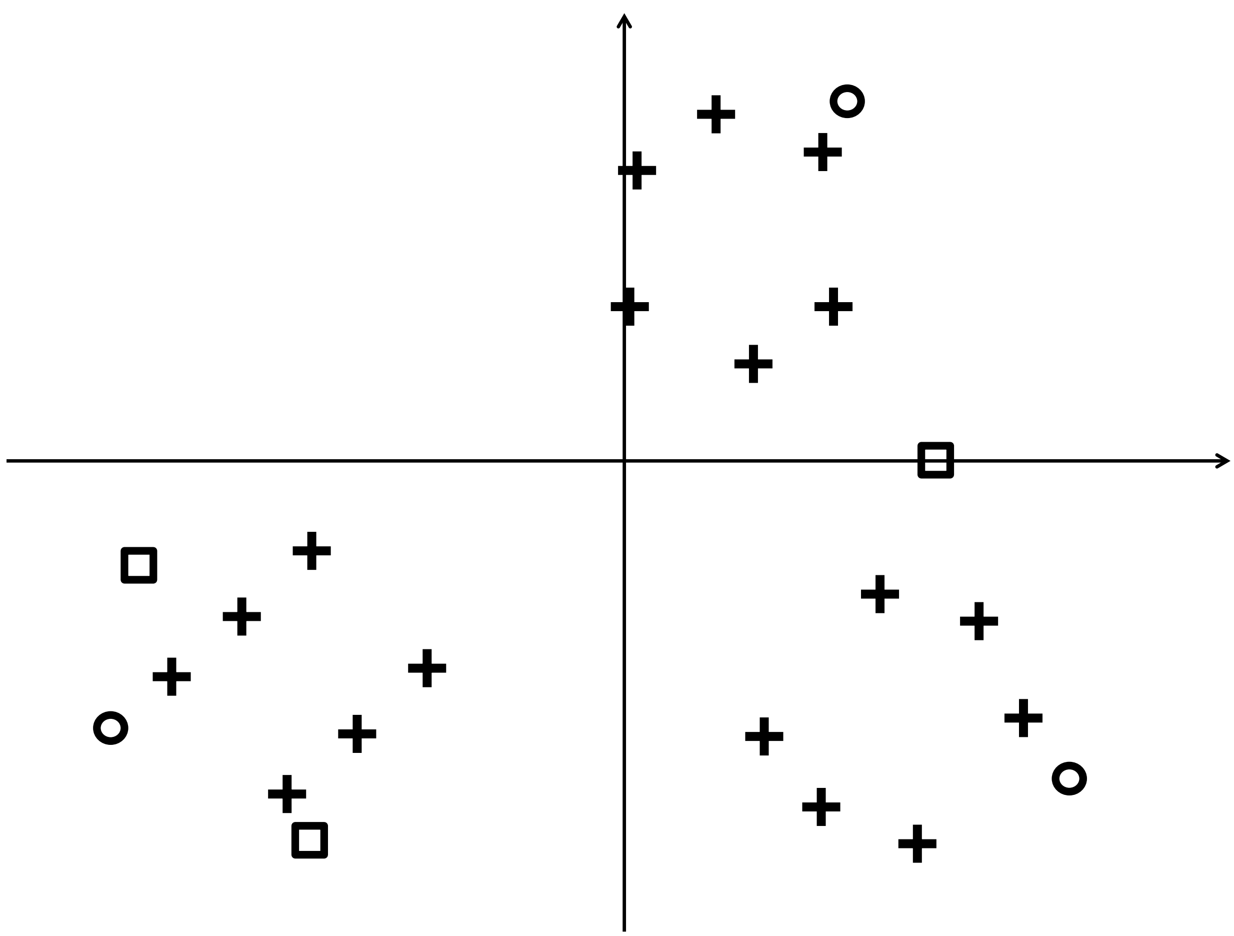
[Solution: Exercise 3]

Exercise 4 [8 points]

Consider the problem of clustering.

1. Introduce the k-means clustering problem, rigorously defining its cost function.
2. Consider Lloyd's algorithm. What is the rule that is used to update the cluster centers after the points are assigned to clusters? Prove that such rule minimizes the k-means cost for the given assignment of points to clusters (i.e., once the assignment of points to clusters is fixed).
3. Consider the data in the figure below where each point $\mathbf{x} \in \mathbb{R}^2$ is represented by a cross. Draw (approximately) the output of Lloyd's algorithm for $k = 3$ when
   (a) the initial centers for the algorithm are the circles;
   (b) the initial centers for the algorithm are the squares.

   Which one of the two resulting clusterings has a lower cost?



[Solution: Exercise 4]

[Solution: Exercise 4]

[Solution: Exercise 4]