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, 6, 7
  - Exercise 3: pag. 8, 9, 10
  - Exercise 4: pag. 11, 12, 13, 14

  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.

Exercise 1 [8 points]

1. Formulate the supervised learning problem, highlighting its main objective in terms of Generalization Error (= Expected Risk) and Training Error (= Empirical Risk).

2. Discuss one approach to pursue this objective for a given finite data set (x_i, y_i), $i = 1, .., n$.

3. Denoting with $\hat{\theta}_n$ the estimated model with n training samples, and with θ^* the best model in the given class, draw in Fig. 1 the typical behaviour of the Generalization Error of θ^*, the Generalization Error of $\hat{\theta}_n$ and the Training Error of $\hat{\theta}_n$ as a function of the number of samples n.

[Solution: Exercise 1]

Figure 1

[Solution: Exercise 1]

[Solution: Exercise 1]

Exercise 2 [8 points]

You want to cluster the points in the figure below using k-means, with $k = 3$.

1. Is there a way to cluster the points using k-means so that in the solution the three clusters corresponds to the three sets with different marks (triangles, squares, circles)? Given a short explanation for your answer.

2. Before clustering, you can apply a transformation to the dataset. Describe a transformation such that the application of k-means with $k = 3$ to the transformed dataset results in three clusters corresponding to the three sets with different marks, and plot the transformed dataset.

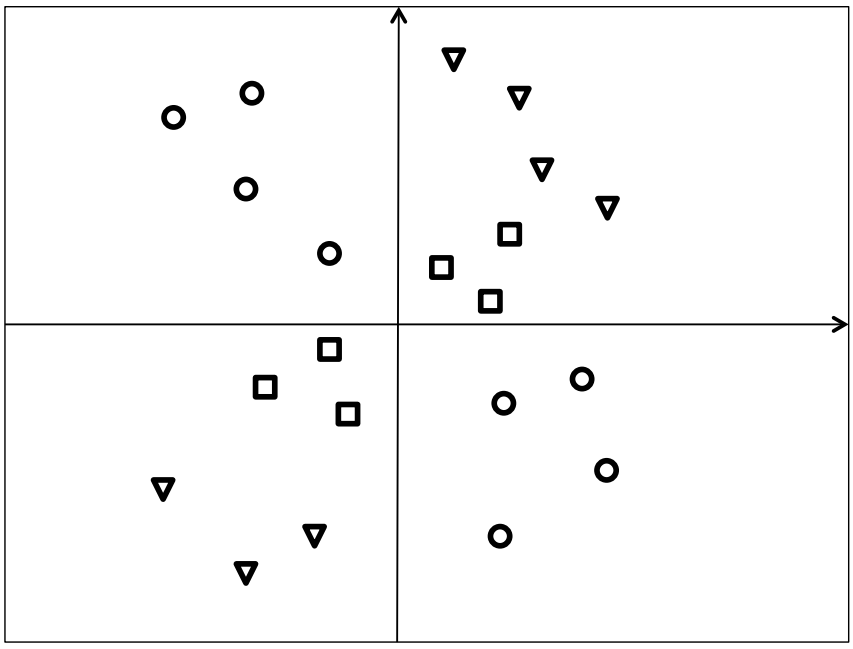
3. Assume that instead of a clustering that minimizes the k-means objective (cost), you are interested in a clustering minimizing the following objective:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \boldsymbol{\mu}_i)^3$$

where the distance $d(\mathbf{x}, \boldsymbol{\mu}), \mathbf{x} \in \mathbb{R}^d, \boldsymbol{\mu} \in \mathbb{R}^d$ is defined as:

$$d(\mathbf{x}, \boldsymbol{\mu}) = \left(\sum_{j=1}^{d} (x_j - \mu_j)^2 \right)^{1/3}.$$

What is a good rule to choose (or update) the cluster centers for a given assignment of points to clusters (i.e., once the assignment of points to clusters is fixed)? Briefly motivate your answer.



[Solution: Exercise 2]

[Solution: Exercise 2]

[Solution: Exercise 2]

Exercise 3 [8 points]

1. Consider the neural network in the figure and assume that a Rectified Linear Unit (ReLU) activation function
$$\sigma(x) = \max(0, x)$$
is used for all neurons. Compute the value of the output y when the input $\mathbf{z}$ is the vector $[z_1 \;\; z_2] = [1 \;\; 3]$.

2. Describe how a neural network can be trained using the backpropagation algorithm (only the main structure of the algorithm; details of the derivation are not required).

3. The Rectified Linear Unit (ReLU) is typically preferred to hyperbolic tangent or sigmoid activation functions. With reference to the gradient descent algorithm used for training the neural network, which are the main advantages of the ReLU with respect to the other two activation functions ?

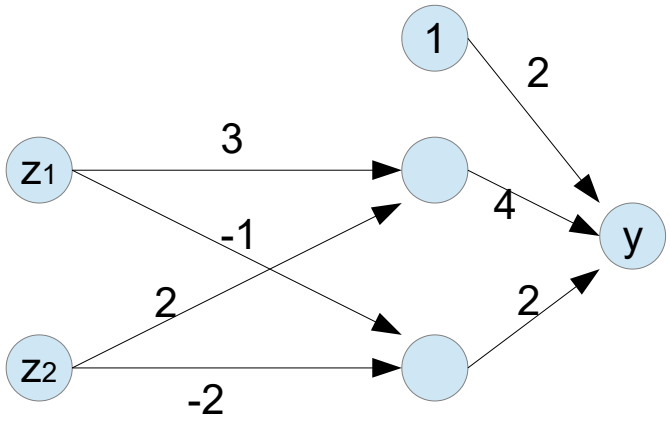


Figure 3

[Solution: Exercise 3]

[Solution: Exercise 3]

Exercise 4 [8 points]

Three classifiers C_1, C_2 and C_3 are trained iteratively using gradient descent minimizing the training error. Let us denote with k the iteration index.

1. Fig. 4a shows the final models for the different binary classifiers on the training set (the circles and crosses correspond to the 2 different classes). In Fig. 4b the behaviour of the training error as a function of the iteration step k is shown. Associate each curve to the corresponding classifier and justify your answer.

2. Assume that you are given a validation dataset (not used to train C_1, C_2, C_3). Fig. 4c shows the behaviour of the validation error as a function of the iteration step k. Associate each curve in Fig. 4c to the corresponding classifier and justify your answer, pointing out which classifier is overfitting or underfitting?

3. Which is a possible technique to reduce the overfitting issue?

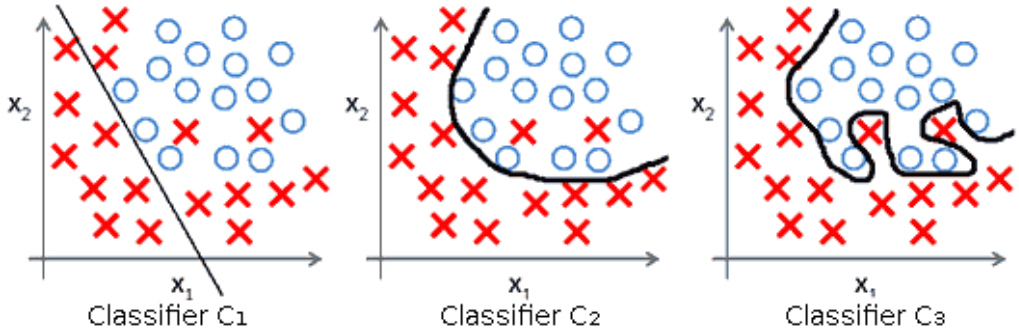


Figure 4a

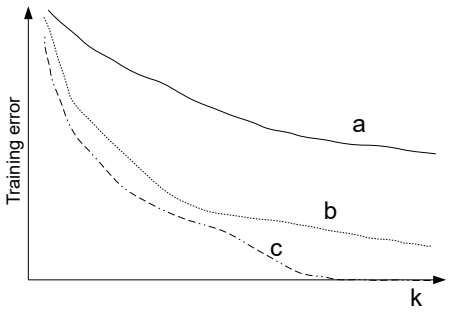


Figure 4b

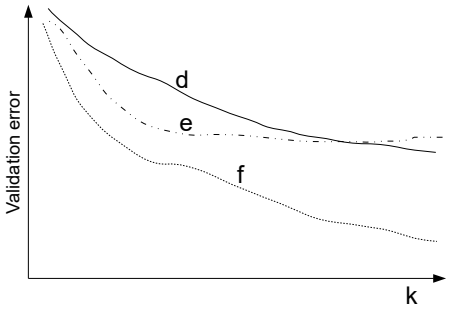


Figure 4c

[Solution: Exercise 4]

[Solution: Exercise 4]

[Solution: Exercise 4]