WRITE FIRST NAME, LAST NAME, AND ID NUMBER ("MATRICOLA") ON YOUR ASSIGNMENT. TIME: 1.5 hours.

FIRST NAME: .....

LAST NAME: .....

ID NUMBER: .....

Exercise 1 [6 points]

1. Introduce the classification problem in machine learning, including discussion of a suitable loss.

2. Introduce as many approaches as you know to solve the classification problem and very briefly discuss their relative pros and cons.

[Solution: Exercise 1]

[Solution: Exercise 1]

Exercise 2 [6 points]

Consider a one-hidden-layer scalar valued ($h(x) \in \mathbb{R}$) neural network of the form

$$h(x) = \sum_{i=1}^{K} \alpha_i \sigma(w_i(x - b_i)) + c, \qquad x \in \mathbb{R}$$

and assume that all the b_i's are fixed (e.g. on a grid ranging in a suitable interval) and $w_i = w$, $\forall i \in [1, K]$.

With reference to a regression problem with data (x_i, y_i), $i = 1, .., m$ with quadratic loss:

1. Write the training error and find a *closed form expression* for the *optimal* α_i and c assuming w FIXED (and known).

2. Assume σ to be the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

which is the role of w? How would you expect the "optimal" w to vary as a function of number of data and the "true" regression function in the following two cases:

   (a) m "large" and the "true" regression function very irregular

   (b) m "small" and the "true" regression function very smooth.

Motivate your answers.

[Solution: Exercise 2]

[Solution: Exercise 2]

Exercise 3 [6 points]

Very often in machine learning problems the model class is described by some hyper-parameters that describe the model "complexity" (e.g. the number of regressors in a regression problem, the "width" of the Kernel in a kernel-SVM, the tradeoff between training loss and penalty function in a regularized problem etc.).

Assume the "hyper-parameter" governing this complexity is scalar (call it λ), non-negative, and assume the model "complexity" increases as lambda increases.

1. Draw a qualitative plot of the *Training Loss* at the optimum as a function of λ
2. Define the k-fold cross validation loss $L_{CV}(\lambda)$, draw a qualitative plot of $L_{CV}(\lambda)$ and discuss how this can be used to find an "optimal" value for λ.

Motivate your answers.

[Solution: Exercise 3]

[Solution: Exercise 3]

Exercise 4 [6 points]

1. Define the dimensionality reduction problem and introduce Principal Component Analysis (PCA). In which way is PCA used to reduce dimensionality? In which way PCA is revealing the "most suitable" dimensionality for our problem (I am not referring to a sharp numerical criterion but just to "which quantities in PCA contain information on the most suitable reduced dimension")?

2. Describe one problem where PCA can be used and how.

Motivate your answers.

[Solution: Exercise 4]

[Solution: Exercise 4]