Machine Learning – March 23, 2018

Time limit: **1.5 hours**.

Last Name	First Name	Matricola
.....	.....	.....

For AIML students, choose at least 3 questions from the 6 below.

EXERCISE A1

1. Provide a formal definition of overfitting.
2. Discuss the problem of overfitting in learning with Decision Trees and illustrate possible solutions to it.

EXERCISE A2

1. Describe the *Naive Bayes Classifier* and highlight the approximation made with respect to the Bayes Optimal Classifier.
2. Provide design and implementation choices for solving the following problem through *Naive Bayes Classifier*:

   Classification of scientific papers in categories according to their main subject. The categories to be considered are: ML (Machine Learning), KR (Knowledge Representation), PL (Planning). Data available for each scientific paper are: title, authors, abstract and publication site (name of the journal and/or of the conference).

EXERCISE A3

1. Define with a precise formal definition the unsupervised learning problem.
2. Provide a full example of unsupervised learning problem (i.e., a specific invented data set), possibly in a graphical form.
3. Describe a solution to the defined problem based on K-Means, providing examples of execution of some steps of the algorithm and a reasonable solution.

EXERCISE B1

1. Provide the main steps of classification based on K-nearest neighbors (K-NN).
2. Draw an example in 2D demonstrating the application of the 3-NN algorithm for the classification of 3 points given a dataset consisting of points from 4 different classes.

Notes: You can choose how the points of the 4 classes are distributed. Use a different symbol for each class (e.g. use (*,x,+,-) for the classes and (o) for the points to be classified).

EXERCISE B2

1. Describe the role of the following notions related to parameter estimation of an artificial neural network:
   - backpropagation
   - forward and backward pass
   - Stochastic Gradient Descent
2. Provide the main steps of the backpropagation algorithm.

EXERCISE B3

1. Briefly describe the goal of linear regression and define the corresponding model.
2. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1^T, t_1)^T, \ldots, (\mathbf{x}_N^T, t_N)^T\}$ with $\mathbf{x}_n$ the input values and t_n the corresponding target values, explain how the parameters of the model can be estimated either in a batch or in a sequential mode.