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

FIRST NAME: .....

LAST NAME: .....

ID NUMBER: .....

Question 1 [6 points]

1. Introduce the main ingrediends of a generic (supervised) learning problem and clearly define its goals.

2. Exploting the concepts and notation introduced in the previous item, and in particular the relation between training cost, generalization cost and model "complexity", define the notions of overfitting and underfitting.

[Solution: Question 1]

[Solution: Question 1]

Question 2 [6 points]

Consider a non linear regression problem in which data (y_i, x_i), $i = 1, .., m$ are collected according to the measurements model

$$y_i = f(x_i) + n_i$$

where $f(x)$ is the "true" unknown function to be estimated and n_i are "errors". The objective is to estimate the function $f(x)$ and you consider a "model class" for $f(x)$ of the form

$$\mathcal{H} := \{h(x) : h(x) = \sum_{j=1}^{M} \theta_j \phi_j(x), \quad \theta_j \in \mathbb{R}, \;\; j = 1, .., M\}$$

where $\phi_j(x)$ are known and fixed basis functions.

1. Discuss the advantages and disadvantages of using a very large number of basis functions, i.e. of having M very large. What if $M >> m$?

2. How would you estimate, using data (y_i, x_i), $i = 1, .., m$, a function $\hat{h}(x)$ from the model class above, in the case in which M is large?

[Solution: Question 2]

[Solution: Question 2]

Question 3 [6 points]

1. Introduce, in the context of unsupervised learning, the (linear) dimensionality reduction problem, clearly defining the cost function that is optimized.

2. Given (training) data $\{x_i \in \mathbb{R}^d\}_{i=1,..,m}$, explaining how dimensionality reduction can be numerically performed.

[Solution: Question 3]

[Solution: Question 3]

Question 4 [6 points]

Consider the model class $\mathcal{H}$ in the non-linear regression problem in Question 2, and further assume that the coefficient α_j's are modeled as i.i.d. zero mean Gaussian random variables with common unknown variance $Var\{\alpha_j\} = \lambda$.

Assume further that the "noises" n_j are i.i.d. zero mean Gaussian, with (unknown) variance σ^2, independent of the α_j's.

1. Frame the problem of estimating $h(x)$ from data (y_i, x_i), $i = 1, .., m$ in the context of Bayesian estimation using Gaussian Processes. Which particular class of GP's does this correspond to?

2. Describe how λ and σ can be estimated.

[Solution: Question 4]

[Solution: Question 4]

EXTRA (Replacing Oral Exam/HW) [8 points]

Each of the following 4 questions, in order of appearence, refer to the previus Questions 1,2,3,4.

(a) State and explain a theorem that under the assumption of *finite* model class, guarantees that overfitting does not occur

(b) In order to solve the given learning problem (when M is large) you may need to introduce an extra "hyperparameter". Describe its role and how it can be estimated.

(c) Provide a "pictorial" representation of the solution to the dimensionality reduction problem in the case $x_i \in \mathbb{R}^2$.

(d) Provide a "deterministic" interpretation of the (Bayesian) solution of the problem in Question 4. To which type of regularization does this correspond to? How are λ and σ related to the regularization parameter?

[Solution: EXTRA (Replacing Oral Exam/HW)]

[Solution: EXTRA (Replacing Oral Exam/HW)]

[Solution: EXTRA (Replacing Oral Exam/HW)]