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 main learning problems (supervised/unsupervised, classification/regression etc.) pointing out the main differences.

2. Assume you are given measured inputs and noisy outputs of an unknown linear time invariant dynamical system. With reference to the taxonomy in Question 1 above, how would you position the problem of learning the unknown impulse response from measured data, which could be a reasonable cost function and how would you explicitly write training and validation error?

Exercise 2 [6 points]

With reference to the binary classification problem:

1. Define the concept of separating hyperplane
2. Highlight the main differences between the perceptron algorithm and the SVM framework.

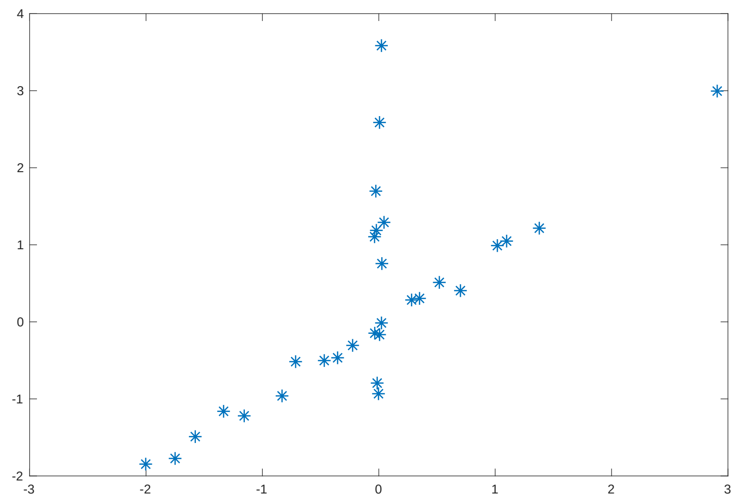
Motivate your answers.

Exercise 3 [6 points]

1. Define the clustering problem and how Gaussian Mixture Models can be used in this context.

2. Assume we would like to group in 2 clusters the data in the figure below. Highlight ONE major difference between GMM and k-means that makes the result of the clustering problem on this example significantly different using k-means or GMM.

Motivate your answers.



[Solution: Exercise 3]

[Solution: Exercise 3]

[Solution: Exercise 3]

Exercise 4 [6 points]

Assume that for a given learning problem the empirical risk is defined by $L_S(w)$ where $w \in \mathbb{R}^n$ is the parameter vector describing the model (for instance all the parameters in a Neural Network or the parameters describing the separating hyperplane in a classification problem).

Assume also a validation error $L_V(w)$ can be computed (e.g. using training-validation split or k-fold cross validation) for each value of the parameter vector w.

1. Describe the Gradient Descent (GD) method for optimizing $L_S(w)$.

2. Assume that the model class is **very rich** as compared to the number of **training** data and denote with $\hat{w}^{(k)}$ the parameter vector at the $k-th$ iteration of GD

    - Draw a qualitative plot of $L_S(\hat{w}^{(k)})$ and $L_V(\hat{w}^{(k)})$ as a function of the iteration index k.
    - How can $L_V(\hat{w}^{(k)})$ be used to decide when the iteration of GD should/could be stopped?

Motivate your answers.