WRITE FIRST NAME, LAST NAME, AND ID NUMBER ("MATRICOLA") BELOW AND READ ALL INSTRUCTIONS BEFORE STARTING WITH THE EXAM! TIME: 2 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
  - Exercise 3: pag. 7, 8, 9
  - Exercise 4: pag. 10, 11, 12

  Solutions written outside the appropriate spaces (including other paper-sheets) 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]

Consider the regression problem when the training data is $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i = [x_{i,1}, x_{i,2}] \in \mathbb{R}^2$ and $y_i \in \mathbb{R}$ for $i = 1, \dots, m$.

1. Formally define the problem when the hypothesis class is $\mathcal{H} = \{h(\mathbf{x}) \text{ s.t. } h(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2; w_0, w_1, w_2 \in \mathbb{R}\}$ and the squared loss is used. In particular, describe what the goal is.

2. Describe ℓ_2 regularization within the context described above. Given an hypothesis $h \in \mathcal{H}$, let $L_S(h)$ be its training error (i.e., the average loss on the training set), $J(h) = L_S(h) + \lambda R(h)$ be the regularized training error, where $R(h)$ is the ℓ_2 regularization function. Formally define the regularization function for ℓ_2 regularization and *derive* the hypothesis that minimizes the ℓ_2 regularized training error.

3. Given an hypothesis h, let $\mathcal{L}(h)$ be the true (or generalization) error of h. Let h_S be the hypothesis that, given data S, minimizes the regularized training error $J(h)$. Plot the typical behavior of $L_S(h_S)$ and $\mathcal{L}(h_S)$ as a function of $\lambda \geq 0$, and describe how this is linked to overfitting.

[Solution: Exercise 1]

[Solution: Exercise 1]

[Solution: Exercise 1]

Exercise 2 [8 points]

Consider the classification problem in machine learning.

1. Provide a formal definition, describing data, loss functions, classification rules etc.

2. Assuming inputs (or features) $x \in \mathbb{R}$, and consider the model class, which is a modified version of logistic regression, defined as the set of models obtained composing the sigmoid function
$$\frac{1}{1+e^{-z}}$$
with the function
$$z = h_{\mathbf{w}}(x) = w_1 + w_2 x^2$$
where the parameters are $\mathbf{w} = [w_1,\ w_2]^\top \in \mathbb{R}^2$. Assume the loss is $\ell(h,(x_i,y_i)) = (y_i - h(x_i))$ if $y_i = 1$, and $\ell(h,(x_i,y_i)) = h(x_i)$ if $y_i = 0$. Write the stoochastic gradient descent update for learning this model from data (x_i, y_i), $i = 1,..,m$.

[Solution: Exercise 2]

[Solution: Exercise 2]

[Solution: Exercise 2]

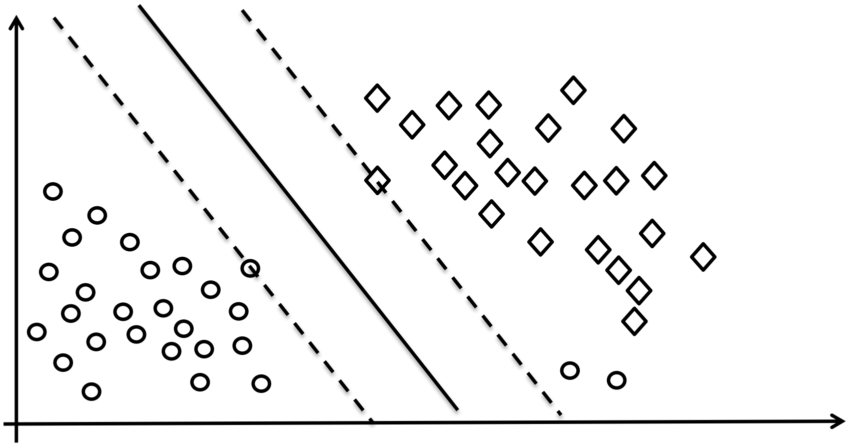
Exercise 3 [8 points]

The Soft-SVM classifier aims at minimizing the following function: $\lambda||\mathbf{w}||^2 + \frac{1}{m}\sum_i \xi_i$.

1. Briefly explain how the Soft-SVM classification method works and which are the constraints under which the function has to be minimized.

2. The figure shows the results of a binary classification performed using a Soft-SVM model with parameter $\lambda = 1$. The training samples are the circles and diamonds and the two shapes correspond to the two classes to which the samples belong. The solid line is the computed separating hyperplane, while the dotted lines represent the margins. For which points ξ_i is different from 0?

3. Does the margin increase or decrease when λ decreases? Guess how the solution changes when a very small value for the λ parameter (i.e., $\lambda \approx 0$) is used, and draw an estimate of the separating hyperplane that could be obtained in this case.



[Solution: Exercise 3]

[Solution: Exercise 3]

[Solution: Exercise 3]

Exercise 4 [8 points]

1. Briefly introduce the clustering problem.

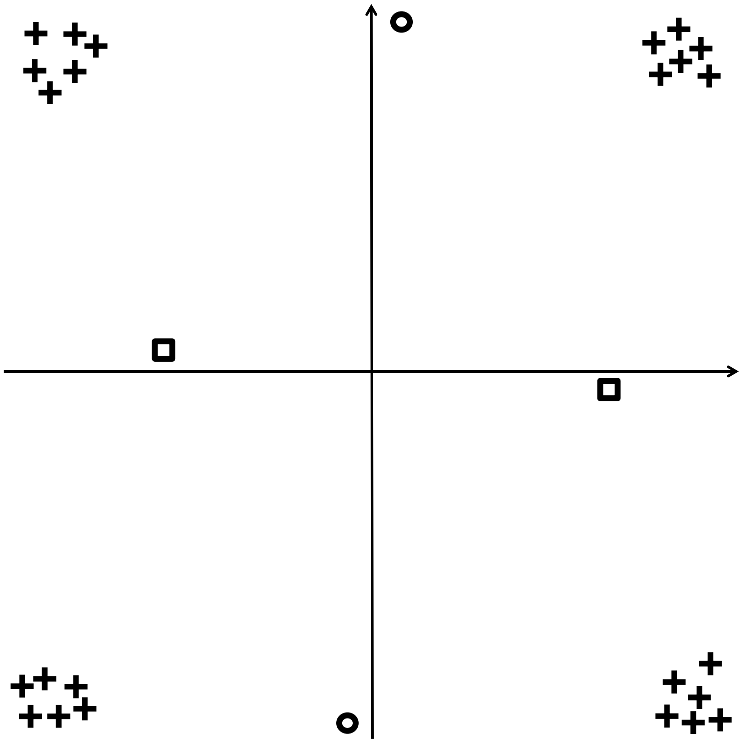
2. Define and explain the cost function used in K-means clustering.

3. Consider the data in the figure below where each point $\mathbf{x} \in \mathbb{R}^2$ is represented by a cross. Show the results (i.e., draw approximately the final centroid locations and the final assignment of the points to the clusters) of clustering into $k = 2$ clusters with K-means when

   (a) the initial centers for the algorithm are the circles;

   (b) the initial centers for the algorithm are the squares.

   Is one solution *significantly* better than the other one? Briefly motivate your answer.



[Solution: Exercise 4]

[Solution: Exercise 4]

[Solution: Exercise 4]