Machine Learning: **II Session 16/02/2023**

Exercise 1 [8 points]

Consider the *regression* problem with *squared loss*.

1. Provide a formal definition of the problem, describing: the data, the learner's input, the learner's output, the loss function, the (most general) assumed generative model for the data, the learner's goal, and what choices the learner has to make.
2. Assume the hypothesis class is $\mathcal{H}$, the data generative distribution is $\mathcal{D}$, and the training data is S. Provide the definition of the training error $L_S(h)$ and of generalization error $L_{\mathcal{D}}(h)$, where $h \in \mathcal{H}$, and prove that for each $h \in \mathcal{H}$, the expectation (over the distribution of the training set) of $L_S(h)$ is equal to $L_{\mathcal{D}}(h)$, justifying all steps of the proof.
3. Use the result above to argue that the ERM procedure can be appropriate when the training data is large.

[Solution: Exercise 1]

Machine Learning: **II Session 16/02/2023**

Exercise 2 [8 points]

Consider the regression problem with *linear models* and *squared loss* where an instance $\mathbf{x} = [x_1, x_2, x_3]$ contains 3 real features (i.e., $\mathcal{X} = \mathbb{R}^3$).

1. Derive the ERM hypothesis.
2. Assume that, given the application domain, you know that $(x_1)^\alpha$, $(x_2)^\alpha$, and $(x_3)^\alpha$ would be the best features for a polynomial model, for some $\alpha \in \mathbb{N}^+$. Describe:
   (a) how you can learn such a polynomial model for a fixed value α;
   (b) a strategy to choose a good value fo α.

[Solution: Exercise 2]

Exercise 3 [8 points]

Consider the binary classification problem with domain set $\mathcal{X} = \mathbb{R}^2$ and label set $\mathcal{Y} = \{-1; 1\}$. You decide to consider the following hypothesis class:

$$\mathcal{H} = \left\{ \mathbf{x} \to \text{sign}\left(-b + \sum_{i=1}^{5} SVM_{\mathbf{w}_i}(\mathbf{x}) \right) : \mathbf{w}_i \in \mathbb{R}^3 \; \forall i \in \{1, \ldots, 5\}, b \in \mathbb{R} \right\}$$

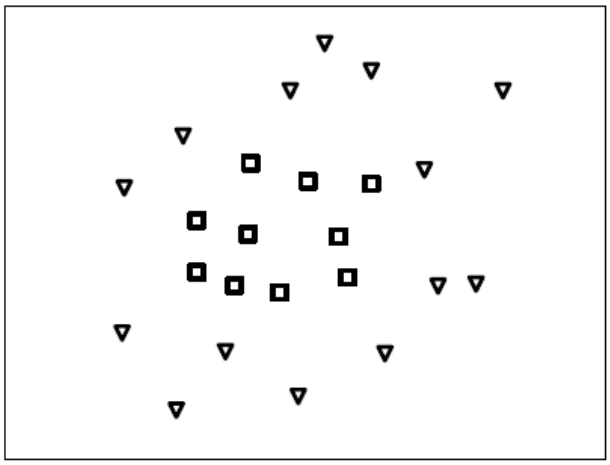
where $SVM_{\mathbf{w}_i}$ is a hard-SVM with parameters $\mathbf{w}_i \in \mathbb{R}^3$ ($\mathbf{w}_i$ includes the bias for the model $SVM_{\mathbf{w}_i}$). (Remember that the output of $SVM_{\mathbf{w}_i} \in \{-1; 1\}$.)

Assume the training data is represented by the figure below: the input $\mathbf{x} \in \mathbb{R}^2$ is given by the coordinates of the point while the label y is 1 if there is a square, and -1 if there is a triangle.

1. Is there an hypothesis $\bar{h}$ in $\mathcal{H}$ that perfectly classifies the training data? What could be a value for b in such hypothesis $\bar{h}$?
2. Plot the decision region of such a model $\bar{h}$ (i.e., where does the model predicts label 1 and where the model predicts label -1) *in the figure below.*

Motivate your answers.

(***Note***: you do **not** need to find the weights $\mathbf{w}_i$ of the SVMs!)



Exercise 4 [8 points]

Consider the *k-means* clustering problem.

1. Describe Lloyd's algorithm.
2. Describe the `k-means++` algorithm for the initialization of the centers.
3. Consider the points in $\mathbb{R}^2$ shown as crosses below. Plot *in the figure below*:
   (a) the output of Lloyd's algorithm with $k = 3$ when the initial centers are the dashed squares;
   (b) the *most likely other centers* chosen by `k-means++`, and the order in which they are chosen, when the first center chosen is the point circled in the figure;
   (c) the output of Lloyd's algorithm with $k = 3$ when the initial centers are chosen by `k-means++` according to your answer to point (b).

   Briefly motivate your plots.