WRITE FIRST NAME, LAST NAME, AND ID NUMBER ("MATRICOLA") BELOW AND READ ALL INSTRUCTIONS BEFORE STARTING WITH THE EXAM! TIME: 2.5 hours.

FIRST NAME: .....

LAST NAME: .....

ID NUMBER: .....

INSTRUCTIONS

- solutions to exercises must be in the appropriate spaces, that is:
  - Exercise 1: pag. 1, 2
  - Exercise 2: pag. 3, 4, 5
  - Exercise 3: pag. 6, 7
  - Exercise 4: pag. 8, 9, 10

  Solutions written outside the appropriate spaces (including other papersheets) 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 [10 points]

1. Provide the formulation of Agnostic PAC learning in terms of sample ($L_{\mathcal{S}}(h)$) and generalisation ($L_{\mathcal{D}}(h)$) errors.

2. Assuming that $x_i \in [a, b]$ (a and b finite but unknown), $i = 1, .., m$ are i.i.d. with mean μ and variance σ^2 and consider the problem of estimating μ solving

$$\hat{\mu} = \arg\min_\theta L_{\mathcal{S}}(\theta)$$

where

$$L_{\mathcal{S}}(\theta) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \theta)^2$$

Prove that

$$L_{\mathcal{D}}(\theta) = \sigma^2 + (\theta - \mu)^2$$

and that, $\forall \epsilon > 0$

$$\lim_{m\to\infty}\mathbb{P}[|L_{\mathcal{S}}(\theta) - (\sigma^2 + (\theta - \mu)^2)| > \epsilon] = 0$$

3. Discuss which tools can be used, in general, to prove that

$$\mathbb{P}[|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| > \epsilon]$$

is small. Under the hypothesis made in the course, how does this probability behave as the sample size m goes to infinity, and which role does it play in the agnostic PAC learning problem?

[Solution: Exercise 1]

[Solution: Exercise 1]

Exercise 2 [8 points]

1. Describe and motivate the ridge regression problem and derive its solution (showing all the steps of the derivation).

2. Let λ be the regularization parameter in ridge regression. Let $\mathcal{S}$ be a training set containing m i.i.d. examples, h_S be the hypothesis that minimises the empirical risk on $\mathcal{S}$, $h_{S,R}$ be the hypothesis that minimises the ridge regression problem defined above on $\mathcal{S}$, and
$$h^* \in \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h).$$
Assume the number of samples m of $\mathcal{S}$ is fixed and that the relation between m and the dimensionality d of the instance space (e.g., $\mathcal{X} = \mathbb{R}^d$) is $m << d$. Plot below the typical behaviour of $L_{\mathcal{S}}(h_S)$, $L_{\mathcal{S}}(h_{S,R})$, $L_{\mathcal{D}}(h_S)$, $L_{\mathcal{D}}(h_{S,R})$ as a function of the regularization parameter λ.

3. Describe the use of cross-validation to estimate the best value of the regularization parameter λ. When is cross-validation a preferable choice compared to the training-validation-test split?

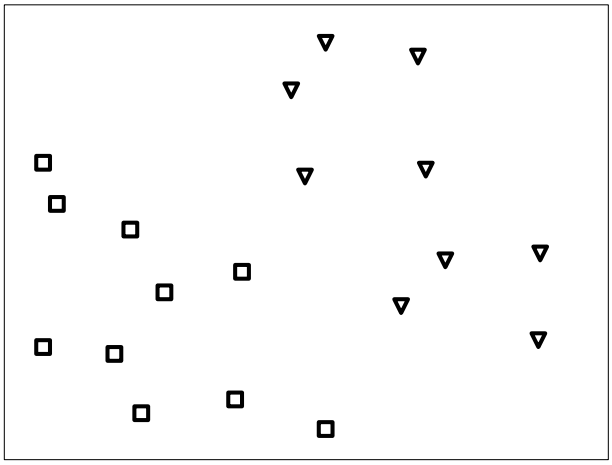
[Solution: Exercise 2]

[Solution: Exercise 2]

[Solution: Exercise 2]

Exercise 3 [7 points]

1. Describe linear SVM for binary classification in the case of linearly separable data.
2. The following figure shows an instance for binary classification with data points in $\mathbb{R}^2$, where the class of each point is represented by its shape (triangle or square). Draw (approximately) the separating hyperplane that would result from running (hard) SVM on the instance and mark the support vectors and the margin. Draw them directly in the figure.

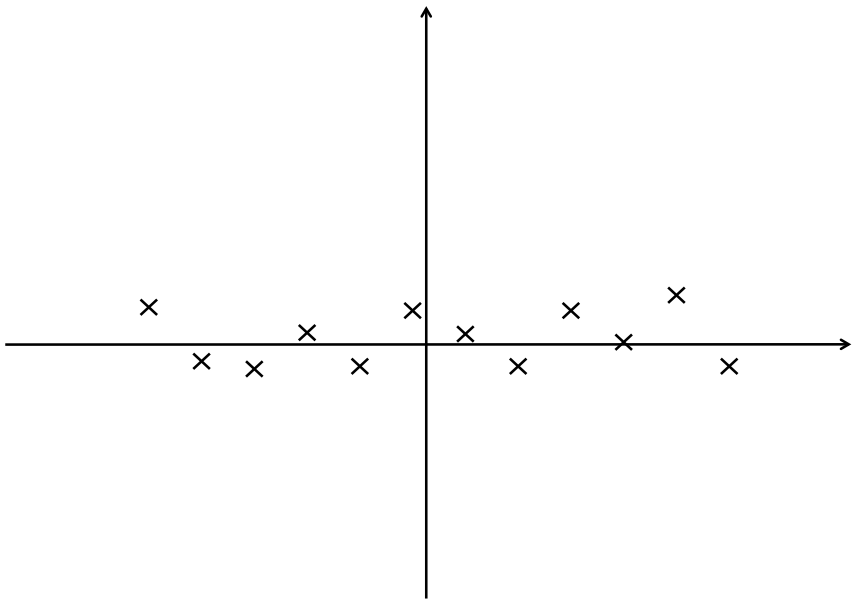


[Solution: Exercise 3]

[Solution: Exercise 3]

Exercise 4 [7 points]

1. Introduce PCA in the context of unsupervised learning.
2. Describe how to obtain the first r principal components for a data matrix $\mathbf{D}$.
3. With respect to the dataset shown in the figure below: draw approximately the first and second right principal components (directly in the figure). In a separate plot, show (approximately) the projection of the dataset on the first principal component.



[Solution: Exercise 4]

[Solution: Exercise 4]

[Solution: Exercise 4]