TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

- Homework limited to 6 pages according to the provided template
- Include your programming code as an Appendix (no page limits)
- Submission Gxxx.PDF in Fenix where xxx is your group number. Please note that it is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is considered valid
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic/manual clearance, homework is nullified and IST guidelines apply for content sharers and consumers, irrespectively of the underlying intent
- Please consult the FAQ before posting questions to your faculty hosts

I. Pen-and-paper [12v]

Consider the following bivariate observations in a Euclidean space:

	y_1	y_2
$\mathbf{x}_1$	2	4
$\mathbf{x}_2$	-1	-4
$\mathbf{x}_3$	-1	2
$\mathbf{x}_4$	4	0

1) [6v] Compute and sketch the clustering solution given by EM assuming considering $\mathbf{x}_1$ and $\mathbf{x}_2$ to be the centroid means and:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \pi_1 = p(c_1 = 1) = 0.7, \pi_2 = p(c_2 = 1) = 0.3$$

2) [3v] Compare the quality of the produced clustering solutions using silhouette.

3) [3v] Identify the VC dimension of the following two-class/binary classifiers: **i)** MLP with three hidden layers with as much nodes as the number of input variables; **ii)** decision tree assuming input variables are discretized using three bins; and **iii)** Bayesian classifier with a multivariate Gaussian likelihood.

   (a) Assume the data dimensionality is five.

   (b) Plot in a single chart how the VC dimension varies with data dimensionality for $m \in \{2,5,10,12,13\}$. What can you conclude (one sentence, English or Portuguese)?

   (c) Plot in a single chart how the VC dimension of **i)** and **iii)** with data dimensionality for $m \in \{2,5,10,30, 100,300,1000\}$. What can you conclude (one sentence, English or Portuguese)?

II. Programming and critical analysis [8v]

Recall the `breast.w.arff` dataset from previous homeworks.

4) [4v] Apply k-means clustering unsupervised on the original data with $k = 2$ and $k = 3$.
   a. Compare the produced solutions against the ECR (external measure)
   b. Compare the produced solutions against the Silhouette coefficient (internal measure).

5) [2v] Visually plot the $k = 3$ clustering solution using the top-2 features with higher mutual information.

6) [2v] Using empirical results from (5), comment on the quality of the produced clustering solution.

END