TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV

Deadline 7/11/2022 (Monday) 23:59 via Fenix as PDF

I. Pen-and-paper [11v]

Given the bivariate observations $\left\{\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\}$,

and the multivariate Gaussian mixture

$$\mathbf{u}_1 = \begin{pmatrix}2\\2\end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix}0\\0\end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}, \pi_1 = 0.5, \pi_2 = 0.5.$$

1) [7v] Perform one epoch of the EM clustering algorithm and determine the new parameters. Indicate all calculus step by step (you can use a computer, however disclose intermediary steps).
2) Given the updated parameters computed in previous question:
   a. [1.5v] perform a hard assignment of observations to clusters under a MAP assumption.
   b. [2.5v] compute the silhouette of the larger cluster using the Euclidean distance.

II. Programming and critical analysis [9v]

Recall the `pd_speech.arff` dataset from earlier homeworks, centered on the Parkinson diagnosis from speech features. For the following exercises, normalize the data using `sklearn`'s MinMaxScaler.

1) [4.5v] Using `sklearn`, apply k-means clustering fully unsupervisedly (without targets) on the normalized data with $k = 3$ and three different seeds (using `random` $\in \{0,1,2\}$). Assess the silhouette and purity of the produced solutions.
2) [1.5v] What is causing the non-determinism?
3) [1.5v] Using a scatter plot, visualize side-by-side the labeled data using as labels: i) the original Parkinson diagnoses, and ii) the previously learned $k = 3$ clusters (`random`$= 0$). To this end, select the two most informative features as axes and color observations according to their label. For feature selection, select the two input variables with highest variance on the MinMax normalized data.
4) [1.5v] The fraction of variance explained by a principal component is the ratio between the variance of that component (i.e., its eigenvalue) and total variance (i.e., sum of all eigenvalues). How many principal components are necessary to explain more than 80% of variability?
   Hint: explore the `DimReduction` notebook to be familiar with PCA in `sklearn`.

END