TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV

Deadline 30/10/2023 (Monday) 23:59 via Fenix as PDF

- Submit Gxxx.ZIP in Fenix where xxx is your group number. The ZIP should contain two files: Gxxx_report.pdf with your report and Gxxx_notebook.ipynb with your notebook demo according to the suggested templates
- It is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is kept
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic or 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 [11v]

Given the following observations, $\left\{\begin{pmatrix}1\\0.6\\0.1\end{pmatrix},\begin{pmatrix}0\\-0.4\\0.8\end{pmatrix},\begin{pmatrix}0\\0.2\\0.5\end{pmatrix},\begin{pmatrix}1\\0.4\\-0.1\end{pmatrix}\right\}$.

Consider a Bayesian clustering that assumes $\{y_1\} \perp\!\!\!\perp \{y_2, y_3\}$, two clusters following a Bernoulli distribution on y_1 (p_1 and p_2), a multivariate Gaussian on $\{y_2, y_3\}$ (N_1 and N_2), and the following initial mixture:

$$\pi_1 = 0.5, \pi_2 = 0.5$$

$$p_1 = P(y_1 = 1) = 0.3, \quad p_2 = P(y_1 = 1) = 0.7$$

$$N_1\left(\mathbf{u}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix}1 & 1\\0 & 1\end{pmatrix}\right), \quad N_2\left(\mathbf{u}_2 = \begin{pmatrix}0\\0\end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}\right).$$

Note: you can solve this exercise by neglecting y_1 and still scoring up to 70% of its grade.

1) [6v] Perform one epoch of the EM clustering algorithm and determine the new parameters.
   Hint: we suggest you to use numpy and scipy, however disclose the intermediary results step by step.

2) [2v] Given the new observation, $\mathbf{x}_{new} = \begin{pmatrix}1\\0.3\\0.7\end{pmatrix}$, determine the cluster memberships (posteriors).

3) [2.5v] Performing a hard assignment of observations to clusters under a ML assumption, identify the silhouette of the larger cluster under a Manhattan distance.

4) [0.5v] Knowing the purity of the clustering solution is 0.75, identify the number of possible classes (ground truth).

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV

Deadline 30/10/2023 (Monday) 23:59 via Fenix as PDF

II. Programming and critical analysis [9v]

Recall the `column_diagnosis.arff` dataset from previous homeworks. For the following exercises, normalize the data using sklearn's MinMaxScaler.

1) [4v] Using `sklearn`, apply k-means clustering fully unsupervisedly on the normalized data with $k \in \{2,3,4,5\}$ (`random=0` and remaining parameters as default). Assess the silhouette and purity of the produced solutions.

2) [2v] Consider the application of PCA after the data normalization:
   i. Identify the variability explained by the top two principal components.
   ii. For each one of these two components, sort the input variables by relevance by inspecting the absolute weights of the linear projection.

3) [2v] Visualize side-by-side the data using: i) the ground diagnoses, and ii) the *previously* learned $k = 3$ clustering solution. To this end, projected the normalized data onto a 2-dimensional data space using PCA and then color observations using the reference and cluster annotations.

4) [1v] Considering the results from questions (1) and (3), identify two ways on how clustering can be used to characterize the population of ill and healthy individuals.

END