Aprendizagem 2023
Homework IV – Group 28

Gonçalo Bárias (ist1103124) & Raquel Braunschweig (ist1102624)

Part I: Pen and Paper

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:

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

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

$$N_1\left(\boldsymbol{\mu}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix}2 & 0.5\\0.5 & 2\end{pmatrix}\right) \quad , \quad N_2\left(\boldsymbol{\mu}_2 = \begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix}1.5 & 1\\1 & 1.5\end{pmatrix}\right)$$

1. **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.**

   Gonçalo

2. **Given the new observation,** $x_{new} = \begin{bmatrix}1 & 0.3 & 0.7\end{bmatrix}^T$**, determine the cluster memberships (posteriors).**

   Raquel

3. **Performing a hard assignment of observations to clusters under a ML assumption, identify the silhouette of both clusters under a Manhattan distance.**

   Raquel

4. **Knowing the purity of the clustering solution is 0.75, identify the number of possible classes (ground truth).**

   Raquel

Part II: Programming and critical analysis

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

1. **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.**

   Gonçalo

2. **Consider the application of PCA after the data normalization:**

   (a) **Identify the variability explained by the top two principal components.**

   Raquel

   (b) **For each one of these two components, sort the input variables by relevance by inspecting the absolute weights of the linear projection.**

   Raquel

3. **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.**

   Raquel

4. **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.**

   Raquel