Aprendizagem 2023
Homework I — Group 003
(ist1107028, ist1107137)

Part I: Pen and paper

Consider the bivariate observations

$$\{x_1 = \begin{bmatrix}1\\0\end{bmatrix}, x_2 = \begin{bmatrix}0\\2\end{bmatrix}, x_3 = \begin{bmatrix}3\\-1\end{bmatrix}\}$$

and the multivariate Gaussian mixture given by

$$\mathbf{u}_1 = \begin{bmatrix}2\\-1\end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix}1\\1\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}, \quad \pi_1 = 0.5, \quad \pi_2 = 0.5$$

Answer the following questions by presenting all intermediary steps, and use 3 decimal places in each.

1. **Perform two epochs of the EM clustering algorithm and determine the new parameters.**
2. **Using the final parameters computed in previous question:**
   a) **perform a hard assignment of observations to clusters under a MAP assumption.**
   b) **compute the silhouette of the larger cluster (the one that has more observations assigned to it) using the Euclidean distance.**

Part II: Programming

In the next exercise you will use the `accounts.csv` dataset. This dataset contains account details of bank clients, and the target variable y is binary ('has the client subscribed a term deposit?').

1. **Select the first 8 features and remove duplicates and null values. Normalize the data using `MinMaxScaler`. Using `sklearn`, apply `k-means clustering` (without targets) on the normalized data with $k = \{2, 3, 4, 5, 6, 7, 8\}$. Apply k-means randomly initialized, using `max_iter = 500` and `random_state = 42`. Plot the different sum of squared errors (SSE) using the `_inertia` attribute of `k-means` according to the number of clusters.**

   Hint: You can use `get_dummies()` to change the feature type from categorical to numerical (e.g. `pd.get_dummies(data, drop_first=True)`)

2. **According to the previous plot, how many underlying customer segments (clusters) should there be ? Explain based on the trade off between the clusters and inertia.**

3. **Would k-modes be a better clustering approach ? Explain why based on the dataset features.**

4. **Apply PCA to the data :**

   a) **Use `StandardScaler` to scale the data before you apply `fit_transform`. How much variability is explained by the top 2 components ?**

   b) **Provide a scatterplot according to the first 2 principal components and color the points according to $k = 3$ clusters. Can we clearly separate the clusters ? Justify.**

5. **Plot the cluster conditional features of the frequencies of 'job" and 'education" according to k-means, with `multiple='dodge'`, `stat='density'`, `shrink=0.8`, `common_norm=False`. Analyze the frequency plots using `sns.displot`, (see Data Exploration notebook). Describe the main differences between the clusters in no more than half page.**