Aprendizagem

LEIC IST-UL

RELATÓRIO - HOMEWORK 4

Grupo 10:

Gabriel Ferreira 107030
Irell Zane 107161

2024/2025 – 1st Semester, P1

Part I: Pen and paper

1. **EM Algorithm**

   Initialization Stage Initial parameters:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \; \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}, \; \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

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

   Data points:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; \mathbf{x}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \; \mathbf{x}_3 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

   (a) **Epoch 1**

   i. **E-step**

   Calculate Gaussian densities using:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

   For each data point:

$$\begin{aligned} \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) &= 0.029 & \qquad \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) &= 0.062 \\ \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) &= 0.005 & \qquad \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) &= 0.048 \\ \mathcal{N}(\mathbf{x}_3|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) &= 0.036 & \qquad \mathcal{N}(\mathbf{x}_3|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) &= 0.011 \end{aligned}$$

   Calculate responsibilities using:

$$\gamma_{ik} = \frac{\pi_k\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

   Results:

$$\begin{aligned} \gamma_{11} &= 0.322 & \qquad \gamma_{12} &= 0.678 \\ \gamma_{21} &= 0.092 & \qquad \gamma_{22} &= 0.908 \\ \gamma_{31} &= 0.770 & \qquad \gamma_{32} &= 0.230 \end{aligned}$$

ii. **M-step**

$$N_1 = \sum_{i=1}^{3} \gamma_{i1} = 1.183 \qquad \pi_1^{\text{new}} = \frac{N_1}{3} = 0.394$$
$$N_2 = \sum_{i=1}^{3} \gamma_{i2} = 1.817 \qquad \pi_2^{\text{new}} = \frac{N_2}{3} = 0.606$$

Update means:

$$\boldsymbol{\mu}_1^{\text{new}} = \frac{1}{N_1} \sum_{i=1}^{3} \gamma_{i1} \mathbf{x}_i = \begin{bmatrix} 2.223 \\ -0.496 \end{bmatrix}$$
$$\boldsymbol{\mu}_2^{\text{new}} = \frac{1}{N_2} \sum_{i=1}^{3} \gamma_{i2} \mathbf{x}_i = \begin{bmatrix} 0.754 \\ 0.873 \end{bmatrix}$$

Update covariances:

$$\begin{aligned}
\boldsymbol{\Sigma}_1^{\text{new}} &= \frac{1}{N_1} \sum_{i=1}^{3} \gamma_{i1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{\text{new}})(\mathbf{x}_i - \boldsymbol{\mu}_1^{\text{new}})^T \\
&= \begin{bmatrix} 1.182 & -0.849 \\ -0.849 & 0.714 \end{bmatrix} \\
\boldsymbol{\Sigma}_2^{\text{new}} &= \frac{1}{N_2} \sum_{i=1}^{3} \gamma_{i2} (\mathbf{x}_i - \boldsymbol{\mu}_2^{\text{new}})(\mathbf{x}_i - \boldsymbol{\mu}_2^{\text{new}})^T \\
&= \begin{bmatrix} 0.947 & -1.039 \\ -1.039 & 1.364 \end{bmatrix}
\end{aligned}$$

(b) **Epoch 2**

i. **E-step**

Calculate Gaussian densities for each data point:

$$\mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = 0.119 \qquad \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = 0.149$$
$$\mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = 0.001 \qquad \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = 0.209$$
$$\mathcal{N}(\mathbf{x}_3|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = 0.346 \qquad \mathcal{N}(\mathbf{x}_3|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = 0.011$$

Responsibilities:

$$\gamma_{11} = 0.342 \qquad \gamma_{12} = 0.658$$
$$\gamma_{21} = 0.004 \qquad \gamma_{22} = 0.996$$
$$\gamma_{31} = 0.953 \qquad \gamma_{32} = 0.047$$

ii. **M-step**

$$N_1 = \sum_{i=1}^{3} \gamma_{i1} = 1.299 \qquad \pi_1^{\text{new}} = \frac{N_1}{3} = 0.433$$
$$N_2 = \sum_{i=1}^{3} \gamma_{i2} = 1.701 \qquad \pi_2^{\text{new}} = \frac{N_2}{3} = 0.567$$

Update means:

$$\boldsymbol{\mu}_1^{\text{new}} = \frac{1}{N_1} \sum_{i=1}^{3} \gamma_{i1} \mathbf{x}_i = \begin{bmatrix} 2.465 \\ -0.728 \end{bmatrix}$$
$$\boldsymbol{\mu}_2^{\text{new}} = \frac{1}{N_2} \sum_{i=1}^{3} \gamma_{i2} \mathbf{x}_i = \begin{bmatrix} 0.469 \\ 1.144 \end{bmatrix}$$

Update covariances:

$$\begin{aligned}
\boldsymbol{\Sigma}_1^{\text{new}} &= \frac{1}{N_1} \sum_{i=1}^{3} \gamma_{i1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{\text{new}})(\mathbf{x}_i - \boldsymbol{\mu}_1^{\text{new}})^T \\
&= \begin{bmatrix} 0.793 & -0.407 \\ -0.407 & 0.215 \end{bmatrix} \\
\boldsymbol{\Sigma}_2^{\text{new}} &= \frac{1}{N_2} \sum_{i=1}^{3} \gamma_{i2} (\mathbf{x}_i - \boldsymbol{\mu}_2^{\text{new}})(\mathbf{x}_i - \boldsymbol{\mu}_2^{\text{new}})^T \\
&= \begin{bmatrix} 0.414 & -0.619 \\ -0.619 & 1.061 \end{bmatrix}
\end{aligned}$$

2. Using the final parameters computed in previous question:

   (a) Perform a hard assignment of observations to clusters under a MAP assumption.

   Under a MAP assumption:

$$p(c_k|x_i) \propto p(x_i|c_k) \cdot p(c_k) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \pi_k$$

   Thus:

$$\begin{aligned}
&p(c_1|x_1) \propto 0.564 \cdot 0.433 = 0.244 && p(c_2|x_1) \propto 0.304 \cdot 0.567 = 0.173 \\
&p(c_1|x_2) \propto 0.000 \cdot 0.433 = 0.000 && p(c_2|x_2) \propto 0.473 \cdot 0.567 = 0.268 \\
&p(c_1|x_3) \propto 1.903 \cdot 0.433 = 0.824 && p(c_2|x_3) \propto 0.000 \cdot 0.567 = 0.000 \\
&\{x_1, x_3\} \in C_1 && \{x_2\} \in C_2
\end{aligned}$$

(b) Compute the silhouette of the larger cluster using the Euclidean distance.

$$a(x_1) = d(x_1, x_3) = 2 \qquad a(x_3) = d(x_3, x_1) = 2$$
$$b(x_1) = d(x_1, x_2) = 1 \qquad b(x_3) = d(x_3, x_2) = 3$$
$$s(x_1) = \frac{b(x_1)}{a(x_1)} - 1 = -0.5 \qquad s(x_3) = 1 - \frac{a(x_1)}{b(x_1)} = 0.333$$

$$s(c_1) = \frac{s(x_1) + s(x_3)}{2} = -0.083$$

Part II: Programming

1. MinMaxScaler

(a) Sum of Squared errors againts the number of clusters in the K-means algorithm

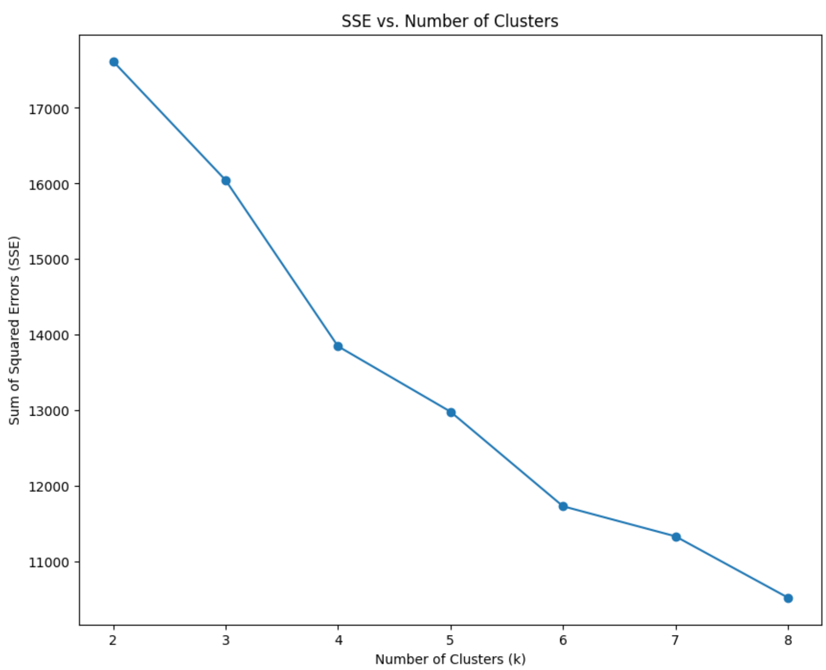


Figure 1: Sum of Squared errors vs Number of Clusters (k)

(b) There should be 4 customer segments (clusters) based on the plot. This is because of the slope in the inertia curve demonstrating an elbow (slope change) at the third clustering solution (k = 4).

Inertia decreases as the number of clusters k increases because more clusters allow data points to be grouped closer to their respective centroids. However, this trend weakens throughout the plot, the decrease in inertia becomes less significant, indicating diminishing returns with a larger number of smaller clusters, which might imply a loss in the generalization ability of that clustering solution, and be more computationally intensive.

(c) K-modes would be a better clustering approach for this dataset. Out of the first 8 features, 6 are categorical variables (marital, education, default, housing, loan, contact) and only 2 are numerical (age, balance). K-modes is generally better at handling categorical data, while k-means is generally better at handling numerical data. Additionally, One-Hot encoding (done with `get_dummies()`) is not necessary with k-modes.

K-modes uses a matching dissimilarity measure for categorical variables, which is more appropriate than the Euclidean distance used in k-means. This makes the clustering more meaningful for categorical data. K-modes also represents cluster centers using modes rather than means, which makes more sense for categorical data.

2. StandardScaler

(a) Variability explained by top 2 components:
22.76%

(b) We cannot clearly separate the three clusters, the inter-cluster distance is not considerable and there is a great deal of overlap between clusters so there is not a clear separation.

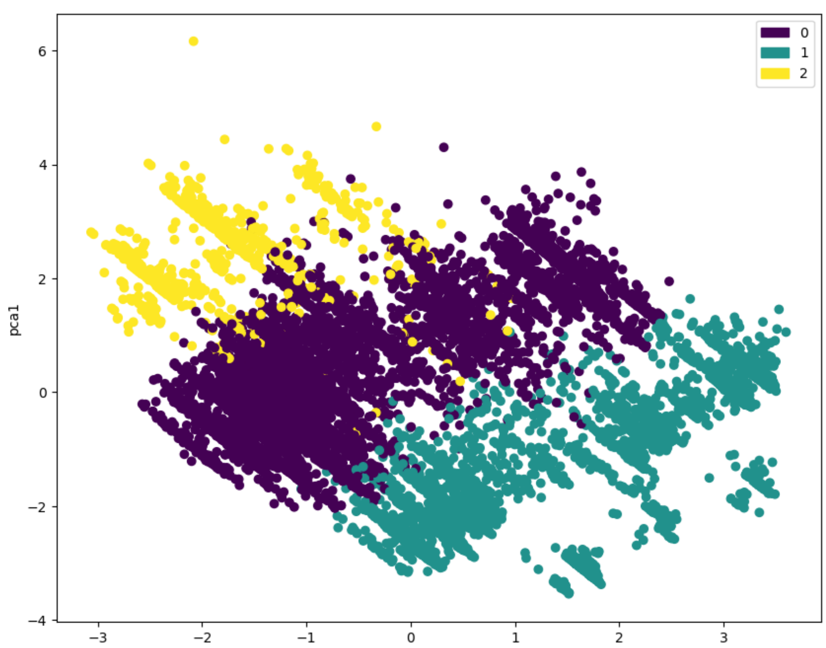
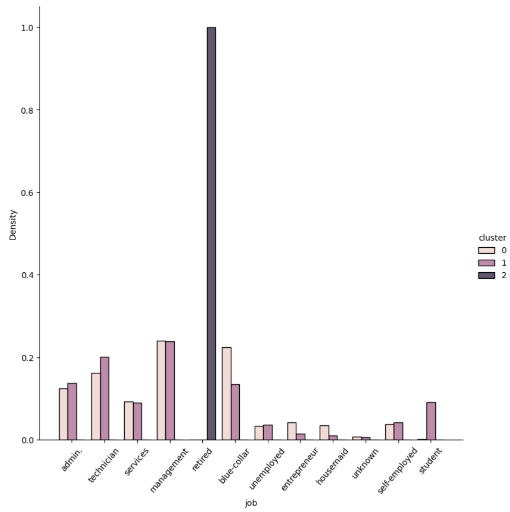
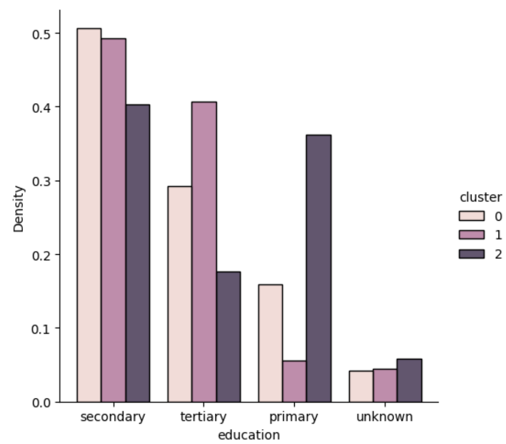


Figure 2: Scatter plot according to the first 2 PCs by cluster.

(c) The first thing that is more immediately clear is that the entirety of cluster 2 is comprised of *retired* people, and all retired people belong to cluster 2, and that cluster 2 observations have relatively lower education, being the one with the greatest proportion of having only primary education and the one with smallest proportion of tertiary and secondary education, which might suggest it's an older population.



(a) Cluster conditional job density distribution plot.



(b) Cluster conditional education density distribution plot.

Cluster 0 and 1 are much more similar to each other with the notable differences that regarding jobs, pratically every student belongs to cluster 1, while cluster 0 has more blue-collar, entrepeneur and house-maid positions, which might suggest that cluster 1 is generally younger. Also notable that cluster 1 has a proportionally higher level of education.