Aprendizagem 2024/2025
Homework III – Group 115
(ist106485, ist107219)

Part I: Pen and paper

Consider the bivariate observations $\{\mathbf{x}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix}0\\2\end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix}3\\-1\end{bmatrix}\}$ and the multivariate Gaussian mixture given by

$$\mu_1 = \begin{bmatrix}2\\-1\end{bmatrix}, \quad \mu_2 = \begin{bmatrix}1\\1\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}, \quad \pi_1 = 0.5, \quad \pi_2 = 0.5$$

1. Perform two epochs of the EM clustering algorithm and determine the new parameters.

   (a) **Epoch 1**: After the initialized parameters above, we can begin the Expectation step by computing the weights of each observation and cluster like this:

   $$\gamma_{ki} = P(c_k \mid \mathbf{x}_i) = \frac{P(\mathbf{x}_i \mid c_k)P(c_k)}{P(\mathbf{x}_i)} = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_i \mid \mu_k, \mathbf{\Sigma}_k)}{\sum_{z \in Z} \pi_z \cdot \mathcal{N}(\mathbf{x}_i \mid \mu_z, \mathbf{\Sigma}_z)}$$

   where Z is the set of clusters

   i. $\mathbf{x}_1$

   $$\begin{aligned}(\text{Cluster 1}) : \gamma_{11} &= \frac{0.5 \cdot \mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1)}{0.5\mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1) + 0.5\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)}\\ &= \frac{0.5 \cdot 0.029}{0.5 \cdot 0.029 + 0.5 \cdot 0.062}\\ &= \boxed{0.319}\end{aligned}$$

   $$\begin{aligned}(\text{Cluster 2}) : \gamma_{21} &= \frac{0.5 \cdot \mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)}{0.5\mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1) + 0.5\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)}\\ &= \frac{0.5 \cdot 0.062}{0.5 \cdot 0.029 + 0.5 \cdot 0.062}\\ &= \boxed{0.681}\end{aligned}$$

ii. $\mathbf{x}_2$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{12} &= \frac{0.5 \cdot \mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1)}{0.5\mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1) + 0.5\mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.5 \cdot 0.005}{0.5 \cdot 0.005 + 0.5 \cdot 0.048} \\
&= \boxed{0.094}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{22} &= \frac{0.5 \cdot \mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)}{0.5\mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1) + 0.5\mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.5 \cdot 0.048}{0.5 \cdot 0.005 + 0.5 \cdot 0.048} \\
&= \boxed{0.906}
\end{aligned}$$

iii. $\mathbf{x}_3$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{13} &= \frac{0.5 \cdot \mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1)}{0.5\mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1) + 0.5\mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.5 \cdot 0.036}{0.5 \cdot 0.036 + 0.5 \cdot 0.011} \\
&= \boxed{0.766}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{23} &= \frac{0.5 \cdot \mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)}{0.5\mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1) + 0.5\mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.5 \cdot 0.011}{0.5 \cdot 0.036 + 0.5 \cdot 0.011} \\
&= \boxed{0.234}
\end{aligned}$$

Therefore the computed responsabilities are given by this matrix: $\mathbf{\Gamma} = \begin{bmatrix} 0.319 & 0.094 & 0.766 \\ 0.681 & 0.906 & 0.234 \end{bmatrix}$

(b) **Epoch 1**: Now for the maximization step where for each $\mathbf{c}_k$ we will recalculate the components of the mixture:

i. Cluster 1

$$
\begin{aligned}
N_1 &= \sum_i \gamma_{1i} \\
&= 0.319 + 0.094 + 0.766 \\
&= \boxed{1.179}
\end{aligned}
$$

$$
\begin{aligned}
\mu_1 &= \frac{1}{N_1} \sum_i \gamma_{1i} \cdot \mathbf{x}_i \\
&= \frac{1}{1.179} \left(\begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0.319 \\ 0.094 \\ 0.766 \end{bmatrix} \right) \\
&= \boxed{\begin{bmatrix} 2.220 \\ -0.490 \end{bmatrix}}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_1 &= \frac{1}{N_1} \sum_i \gamma_{1i} \cdot (\mathbf{x}_i - \mu_1) \cdot (\mathbf{x}_i - \mu_1)^T \\
&= \frac{1}{1.179} (0.319 \cdot \begin{bmatrix} -1.220 \\ 0.490 \end{bmatrix} \cdot \begin{bmatrix} -1.220 & 0.495 \end{bmatrix} + 0.094 \cdot \begin{bmatrix} -2.220 \\ 2.490 \end{bmatrix} \cdot \begin{bmatrix} -2.220 & 2.490 \end{bmatrix} \\
&\quad + 0.766 \cdot \begin{bmatrix} 0.78 \\ -0.51 \end{bmatrix} \cdot \begin{bmatrix} 0.78 & -0.51 \end{bmatrix}) \\
&= \boxed{\begin{bmatrix} 1.191 & -0.861 \\ -0.861 & 0.728 \end{bmatrix}}
\end{aligned}
$$

$$
\begin{aligned}
\pi_1 &= \frac{N_1}{N_1 + N_2} \quad (N_2 \text{ is calculated below}) \\
&= \frac{1.179}{1.179 + 1.821} \\
&= \boxed{0.393}
\end{aligned}
$$

ii. Cluster 2

$$
\begin{aligned}
N_2 &= \sum_i \gamma_{2i} \\
&= 0.681 + 0.906 + 0.234 \\
&= \boxed{1.821}
\end{aligned}
$$

$$
\begin{aligned}
\mu_2 &= \frac{1}{N_2} \sum_i \gamma_{2i} \cdot \mathbf{x}_i \\
&= \frac{1}{1.821} \left(\begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0.681 \\ 0.906 \\ 0.234 \end{bmatrix} \right) \\
&= \boxed{\begin{bmatrix} 0.759 \\ 0.867 \end{bmatrix}}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_2 &= \frac{1}{N_2} \sum_i \gamma_{2i} \cdot (\mathbf{x}_i - \mu_2) \cdot (\mathbf{x}_i - \mu_2)^T \\
&= \frac{1}{1.821} (0.681 \cdot \begin{bmatrix} 0.241 \\ -0.867 \end{bmatrix} \cdot \begin{bmatrix} 0.241 & -0.867 \end{bmatrix} + 0.906 \cdot \begin{bmatrix} -0.759 \\ 1.133 \end{bmatrix} \cdot \begin{bmatrix} -0.759 & 1.133 \end{bmatrix} \\
&\quad + 0.234 \cdot \begin{bmatrix} 2.241 \\ -1.867 \end{bmatrix} \cdot \begin{bmatrix} 2.241 & -1.867 \end{bmatrix}) \\
&= \boxed{\begin{bmatrix} 0.954 & -1.044 \\ -1.044 & 1.368 \end{bmatrix}}
\end{aligned}
$$

$$
\begin{aligned}
\pi_2 &= \frac{N_2}{N_1 + N_2} \\
&= \frac{1.821}{1.179 + 1.821} \\
&= \boxed{0.607}
\end{aligned}
$$

(c) **Epoch 2**: E-step

The updated gaussian mixture after the 1st iteration is the following:

$$
\mu_1 = \begin{bmatrix} 2.220 \\ -0.490 \end{bmatrix}, \ \mu_2 = \begin{bmatrix} 0.759 \\ 0.867 \end{bmatrix}
$$

$$
\Sigma_1 = \begin{bmatrix} 1.191 & -0.861 \\ -0.861 & 0.728 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0.954 & -1.044 \\ -1.044 & 1.368 \end{bmatrix}
$$

$$
\pi_1 = 0.393, \ \pi_2 = 0.607
$$

Let us calculate the responsabilities once more:

i. $\mathbf{x}_1$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{11} &= \frac{0.393 \cdot \mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1)}{0.393\mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1) + 0.607\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.393 \cdot 0.116}{0.393 \cdot 0.116 + 0.607 \cdot 0.149} \\
&= \boxed{0.335}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{21} &= \frac{0.607 \cdot \mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)}{0.393\mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1) + 0.607\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.607 \cdot 0.149}{0.393 \cdot 0.118 + 0.607 \cdot 0.149} \\
&= \boxed{0.665}
\end{aligned}$$

ii. $\mathbf{x}_2$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{12} &= \frac{0.393 \cdot \mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1)}{0.393\mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1) + 0.607\mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.393 \cdot 0.001}{0.395 \cdot 0.001 + 0.607 \cdot 0.207} \\
&= \boxed{0.003}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{22} &= \frac{0.607 \cdot \mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)}{0.393\mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1) + 0.607\mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.607 \cdot 0.207}{0.395 \cdot 0.001 + 0.607 \cdot 0.207} \\
&= \boxed{0.997}
\end{aligned}$$

iii. $\mathbf{x}_3$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{13} &= \frac{0.393 \cdot \mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1)}{0.393\mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1) + 0.607\mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.393 \cdot 0.343}{0.395 \cdot 0.343 + 0.607 \cdot 0.012} \\
&= \boxed{0.949}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{23} &= \frac{0.607 \cdot \mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)}{0.393\mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1) + 0.607\mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.607 \cdot 0.012}{0.395 \cdot 0.343 + 0.607 \cdot 0.012} \\
&= \boxed{0.051}
\end{aligned}$$

Therefore the computed responsabilities are given by this matrix: $\mathbf{\Gamma} = \begin{bmatrix} 0.335 & 0.003 & 0.949 \\ 0.665 & 0.997 & 0.051 \end{bmatrix}$

(d) **Epoch 2**: M-step

i. Cluster 1

$$N_1 = \sum_i \gamma_{1i}$$
$$= 0.335 + 0.003 + 0.949$$
$$= \boxed{1.287}$$

$$\mu_1 = \frac{1}{N_1} \sum_i \gamma_{1i} \cdot \mathbf{x}_i$$
$$= \frac{1}{1.287} \left(\begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0.335 \\ 0.003 \\ 0.949 \end{bmatrix} \right)$$
$$= \boxed{\begin{bmatrix} 2.472 \\ -0.733 \end{bmatrix}}$$

$$\Sigma_1 = \frac{1}{N_1} \sum_i \gamma_{1i} \cdot (\mathbf{x}_i - \mu_1) \cdot (\mathbf{x}_i - \mu_1)^T$$
$$= \frac{1}{1.287} \cdot (0.335 \cdot \begin{bmatrix} -1.472 \\ 0.733 \end{bmatrix} \cdot \begin{bmatrix} -1.472 & 0.733 \end{bmatrix} + 0.003 \cdot \begin{bmatrix} -2.472 \\ 2.733 \end{bmatrix} \cdot \begin{bmatrix} -2.472 & 2.733 \end{bmatrix}$$
$$+ 0.949 \cdot \begin{bmatrix} 0.528 \\ -0.267 \end{bmatrix} \cdot \begin{bmatrix} 0.528 & -0.267 \end{bmatrix})$$
$$= \boxed{\begin{bmatrix} 0.784 & -0.401 \\ -0.401 & 0.210 \end{bmatrix}}$$

$$\pi_1 = \frac{N_1}{N_1 + N_2} \quad (N_2 \text{ is calculated below})$$
$$= \frac{1.287}{1.287 + 1.713}$$
$$= \boxed{0.429}$$

ii. Cluster 2

$$
\begin{aligned}
N_2 &= \sum_i \gamma_{2i} \\
&= 0.665 + 0.997 + 0.051 \\
&= \boxed{1.713}
\end{aligned}
$$

$$
\begin{aligned}
\mu_2 &= \frac{1}{N_2} \sum_i \gamma_{2i} \cdot \mathbf{x}_i \\
&= \frac{1}{1.713} \left(\begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0.665 \\ 0.997 \\ 0.051 \end{bmatrix} \right) \\
&= \boxed{\begin{bmatrix} 0.478 \\ 1.134 \end{bmatrix}}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_2 &= \frac{1}{N_2} \sum_i \gamma_{2i} \cdot (\mathbf{x}_i - \mu_2) \cdot (\mathbf{x}_i - \mu_2)^T \\
&= \frac{1}{1.713} (0.665 \cdot \begin{bmatrix} 0.522 \\ -1.134 \end{bmatrix} \cdot \begin{bmatrix} 0.522 & -1.134 \end{bmatrix} + 0.997 \cdot \begin{bmatrix} -0.478 \\ 0.866 \end{bmatrix} \cdot \begin{bmatrix} -0.478 & 0.866 \end{bmatrix} \\
&\quad + 0.051 \cdot \begin{bmatrix} 2.522 \\ -2.134 \end{bmatrix} \cdot \begin{bmatrix} 2.522 & -2.134 \end{bmatrix}) \\
&= \boxed{\begin{bmatrix} 0.428 & -0.631 \\ -0.631 & 1.071 \end{bmatrix}}
\end{aligned}
$$

$$
\begin{aligned}
\pi_2 &= \frac{N_2}{N_1 + N_2} \\
&= \frac{1.713}{1.287 + 1.713} \\
&= \boxed{0.571}
\end{aligned}
$$

2. Using the final parameters computed in the previous question:

$$
\mu_1 = \begin{bmatrix} 2.472 \\ -0.733 \end{bmatrix}, \ \mu_2 = \begin{bmatrix} 0.478 \\ 1.134 \end{bmatrix}
$$

$$
\Sigma_1 = \begin{bmatrix} 0.784 & -0.401 \\ -0.401 & 0.210 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0.428 & -0.631 \\ -0.631 & 1.071 \end{bmatrix}
$$

$$
\pi_1 = 0.429, \ \pi_2 = 0.571
$$

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

i. First we must calculate the posterior probabilities once again for each observation and cluster using Bayes' Theorem given the final parameters of the previous question

A. $\mathbf{x}_1$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{11} &= \frac{0.429 \cdot \mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1)}{0.429\mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1) + 0.571\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.429 \cdot 0.619}{0.429 \cdot 0.619 + 0.571 \cdot 0.294} \\
&= \boxed{0.613}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{21} &= \frac{0.571 \cdot \mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)}{0.429\mathcal{N}(\mathbf{x}_1 \mid \mu_1, \mathbf{\Sigma}_1) + 0.571\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.571 \cdot 0.294}{0.433 \cdot 0.619 + 0.571 \cdot 0.294} \\
&= \boxed{0.387}
\end{aligned}$$

B. $\mathbf{x}_2$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{12} &= \frac{0.429 \cdot \mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1)}{P(x_2)} \\
&= \frac{0.429 \cdot 0}{P(x_2)} \\
&= \boxed{0}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{22} &= \frac{0.567 \cdot \mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)}{0.433\mathcal{N}(\mathbf{x}_2 \mid \mu_1, \mathbf{\Sigma}_1) + 0.567\mathcal{N}(\mathbf{x}_2 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.571 \cdot 0.453}{0.429 \cdot 0 + 0.571 \cdot 0.453} \\
&= \boxed{1}
\end{aligned}$$

C. $\mathbf{x}_3$

$$\begin{aligned}
\text{(Cluster 1)} : \gamma_{13} &= \frac{0.429 \cdot \mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1)}{0.429\mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1) + 0.571\mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.429 \cdot 2.148}{0.429 \cdot 2.148 + 0.571 \cdot 0} \\
&= \boxed{1}
\end{aligned}$$

$$\begin{aligned}
\text{(Cluster 2)} : \gamma_{23} &= \frac{0.571 \cdot \mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)}{0.429\mathcal{N}(\mathbf{x}_3 \mid \mu_1, \mathbf{\Sigma}_1) + 0.571\mathcal{N}(\mathbf{x}_3 \mid \mu_2, \mathbf{\Sigma}_2)} \\
&= \frac{0.571 \cdot 0}{P(x_3)} \\
&= \boxed{0}
\end{aligned}$$

ii. Now we just need to assess each observation under MAP assumption:

$$f(\mathbf{x}_i) \longleftarrow \arg\max_k P(c_k \mid \mathbf{x}_i) = \arg\max_k \gamma_{ki}$$

$$\begin{aligned} x_1 : \gamma_{21} < \gamma_{11} &\implies \textbf{Cluster 1} \\ x_2 : \gamma_{22} > \gamma_{12} &\implies \textbf{Cluster 2} \\ x_3 : \gamma_{23} < \gamma_{13} &\implies \textbf{Cluster 1} \end{aligned}$$

(b) Compute the silhouette of the largest cluster (the one that has more observations assigned to it) using the Euclidean distance.

i. The largest cluster is **Cluster 1**. So let us compute the silhouette score for each observation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

ii. Calculating the average intra-cluster distance $a(i)$:

$$\begin{aligned} a(1) = d(x_1, x_3) &= \sqrt{(1-3)^2 + (0+1)^2} = \sqrt{5} = \boxed{2.236} \\ a(3) = d(x_3, x_1) &= a(1) = \sqrt{5} = \boxed{2.236} \end{aligned}$$

iii. Calculating the average inter-cluster distance $b(i)$:

$$\begin{aligned} b(1) = d(x_1, x_2) &= \sqrt{(1-0)^2 + (0-2)^2} = \sqrt{5} = \boxed{2.236} \\ b(3) = d(x_3, x_2) &= \sqrt{(3-0)^2 + (-1-2)^2} = \sqrt{18} = \boxed{4.243} \end{aligned}$$

iv. Finally, the silhouette score of Cluster 1

$$\begin{aligned} s(1) &= \frac{b(1) - a(1)}{\max\{a(1), b(1)\}} \\ &= \frac{2.236 - 2.236}{2.236} \\ &= \boxed{0} \end{aligned}$$

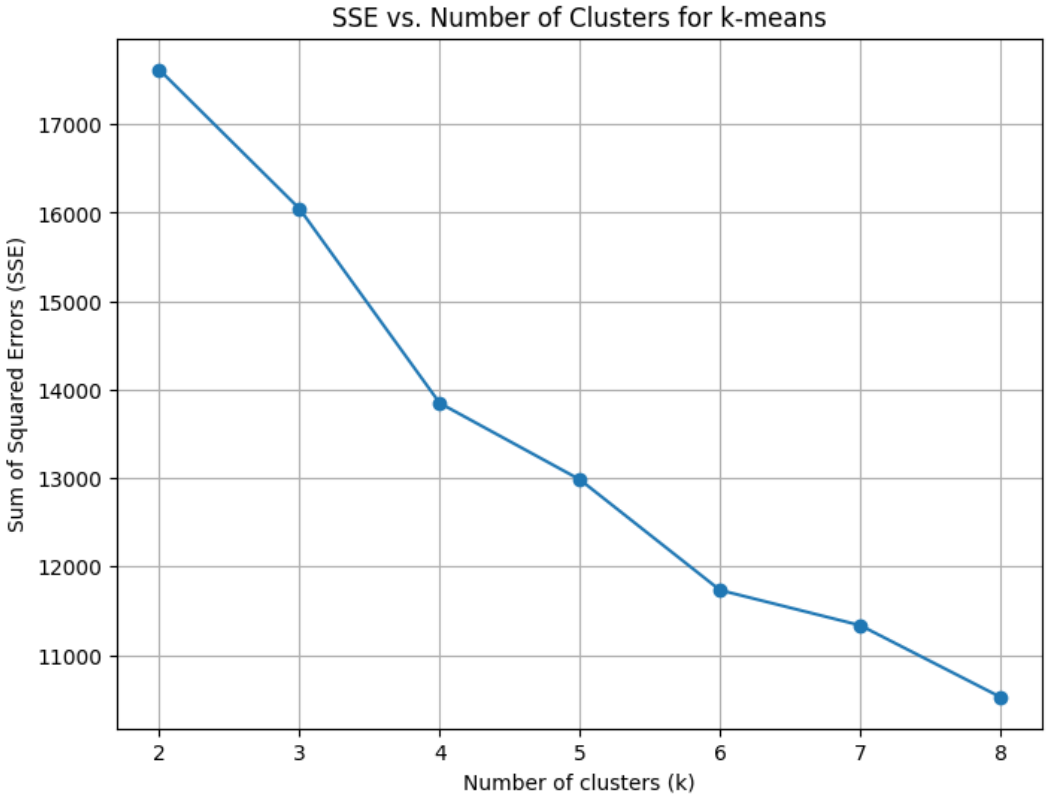
$$\begin{aligned} s(3) &= \frac{b(3) - a(3)}{\max\{a(3), b(3)\}} \\ &= \frac{4.243 - 2.236}{4.243} \\ &= \boxed{0.473} \end{aligned}$$

$$\begin{aligned} s(\text{Cluster 1}) &= \frac{s(1) + s(3)}{2} \\ &= \boxed{0.237} \end{aligned}$$

Part II: Programming

1. Normalize the data using MinMaxScaler:

   (a) Using sklearn, apply k-means clustering (without targets) on the normalized data with $k = \{2, 3, 4, 5, 6, 7, 8\}$, `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.



   (b) 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.

   Answer: In this plot, we can see a noticeable decrease in SSE up to around k=4, after which the SSE reduction rate slows down. This suggests that 4 clusters might be an appropriate choice, as adding more clusters beyond this point provides diminishing returns in terms of reducing SSE. This choice balances cluster compactness with simplicity, likely representing the underlying customer segments effectively without adding unnecessary complexity.

   (c) Would k-modes be a better clustering approach? Explain why based on the dataset features.

   Answer: The dataset includes both categorical and numerical features, with a significant focus on categorical data such as job, marital status, education, and various financial behavior indicators. K-means clustering, which minimizes the Euclidean distance between data points and cluster centroids, is primarily suited for numerical data. This reliance on distance calculations makes K-means inadequate for effectively clustering datasets with many categorical attributes, as it fails to interpret the relationships among these variables appropriately.

In contrast, K-modes clustering is specifically designed for categorical data. It uses a dissimilarity measure based on the matching of categorical values, grouping data points by the frequency of each category. Rather than calculating means, K-modes employs modes to define cluster centroids, aligning more closely with the dataset's characteristics. This approach enhances the algorithm's ability to form meaningful clusters without being affected by outliers, a common issue with K-means.

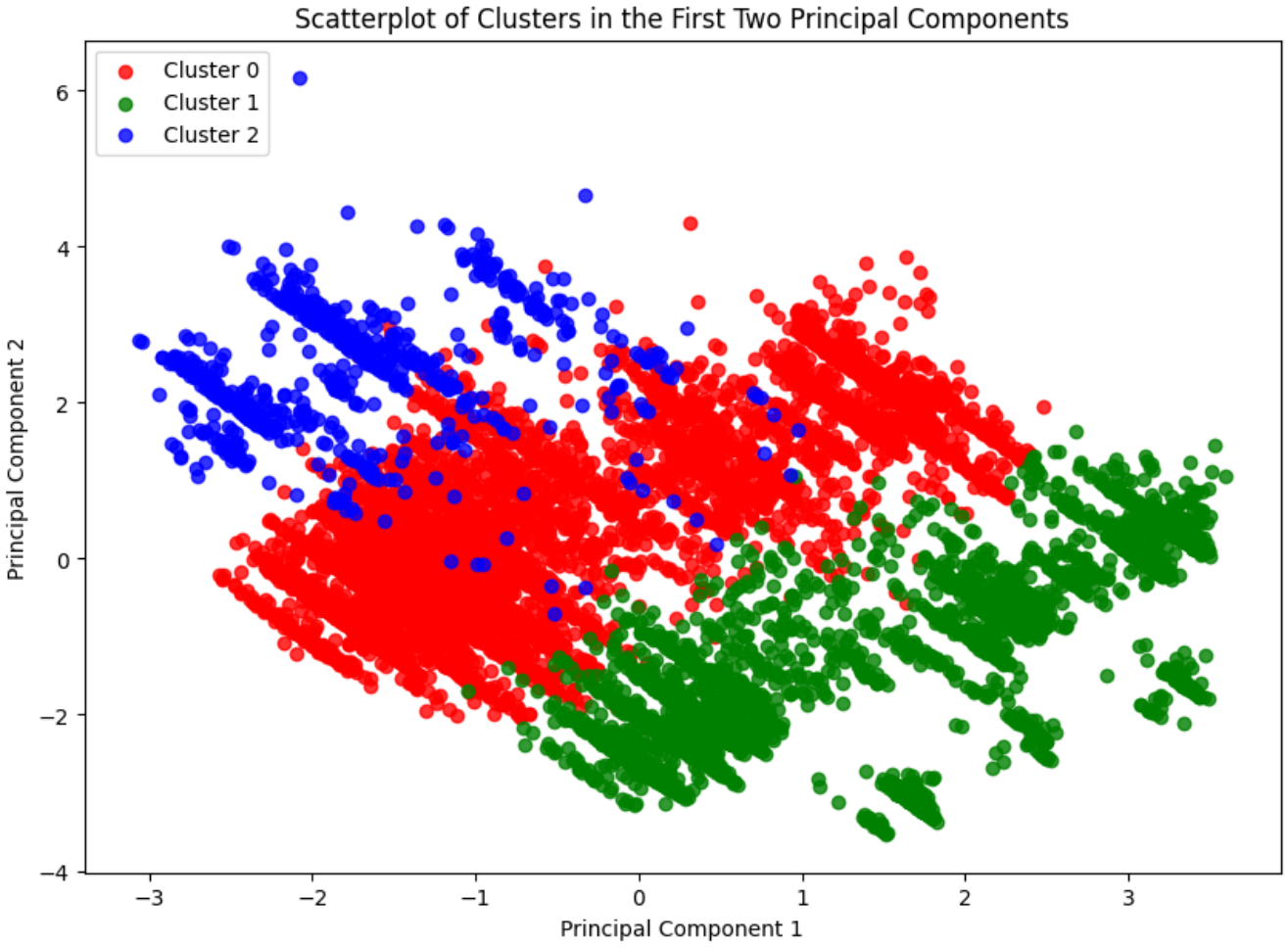
In conclusion, given the dataset's mixed data types and emphasis on categorical features, K-modes is the more suitable clustering approach.

2. Normalize the data using StandardScaler:

   (a) Apply PCA to the data. How much variability is explained by the top 2 components?

   Answer: Upon examination of the variance explained by the principal components, we observe that the first two components collectively account for 22.76% of the total variance within the dataset, with the first component explaining 11.68% and the second component contributing an additional 11.08%.

   (b) Apply k-means clustering with $k = 3$ and `random_state=42` (all other arguments as default) and use the original 8 features. Next, provide a scatterplot according to the first 2 principal components. Can we clearly separate the clusters? Justify.

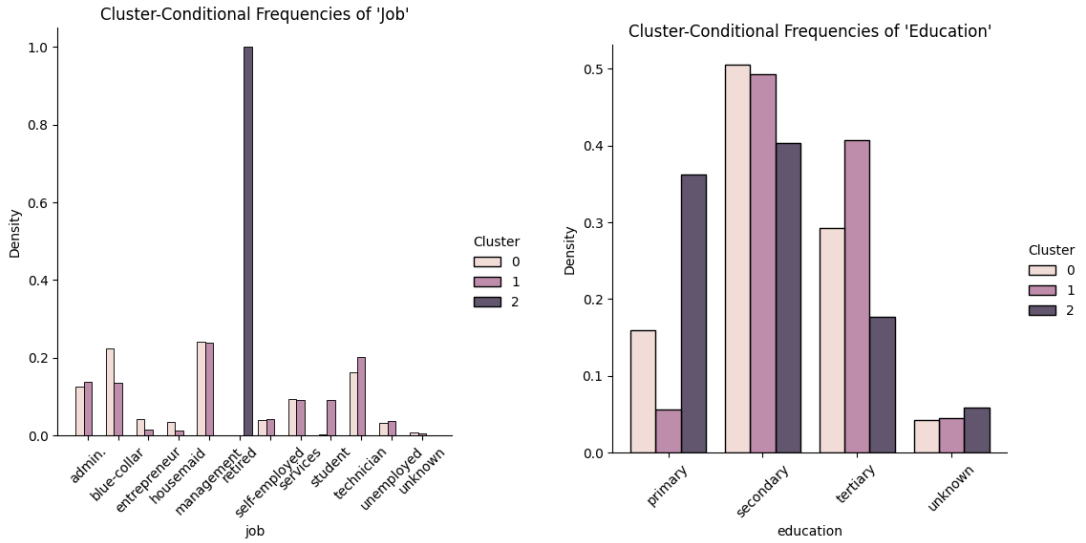


Answer: The scatterplot of the first two principal components shows significant overlap among the three clusters identified by K-means, indicating that these components do not provide sufficient separation between clusters.

This outcome can be attributed to the distinct objectives of K-means and PCA: while K-means seeks to minimize intra-cluster distances, PCA optimizes for maximum variance without regard for the spatial arrangement of clusters.

Moreover, the original eight features may lack strong discriminatory power with respect to the target variable (term deposit subscription). Since these features primarily contribute to overall variance rather than class separability, the clustering is weakly reflected in the PCA-transformed space.

(c) Plot the cluster conditional features of the frequencies of "job" and "education" according to the clusters obtained in the previous question (2b.). Use `sns.distplot` with `multiple="dodge", stat='density', shrink=0.8` and `common_norm=False`. Describe the main differences between the clusters in no more than half a page.



(a) (b)

Answer: Based on the analysis of the clusters, we observe that Cluster 0 is mostly characterized by blue-collar workers, entrepeneurs and housemaids. The predominant education level in this cluster is secondary education, suggesting that most individuals have completed high school but may not have pursued further academic qualifications. This reflects a workforce engaged in hands-on, labor-intensive roles, often requiring practical skills.

In contrast, Cluster 1 features a demographic that includes students and technicians, indicating a younger population actively pursuing education or entry-level technical positions. The most prominent education level here is tertiary education, which suggests that many individuals in this cluster are either pursuing or have completed post-secondary qualifications. This difference suggests a changing labor market that emphasizes the importance of education.

Cluster 2, on the other hand, is uniquely defined by retired individuals, indicating a segment of the population that is no longer part of the active workforce. The predominant education level in this cluster is primary education, suggesting that many individuals have had limited formal education, which may influence their employment history and opportunities throughout their careers.

Interestingly, there are similar job distributions between Cluster 0 and Cluster 1 for positions such as manager, self-employed, administration, services, and unemployed. This overlap implies that, despite their differing primary roles, these clusters may share common career paths and job markets. The prominent education level among these shared roles is also secondary education, indicating a similar pool of educational attainment between the two clusters.