Aprendizagem 2024/2025

Homework IV – Group 034

(ist199417, ist1106481)

I. Pen-and-paper

1)

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

1) Inicialização

$u_1 = \begin{bmatrix}2\\-1\end{bmatrix}$ $\Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}$ $\pi_1 = 0.5$

$u_2 = \begin{bmatrix}1\\1\end{bmatrix}$ $\Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$ $\pi_2 = 0.5$

2) Expetativa

$\gamma_{ki} = p(c_k|x_i) = \frac{p(c_k, x_i)}{\sum_j p(c_j, x_i)}, p(c_k, x_i) = N(x_i|u_k, \Sigma_k) \cdot \pi_k$

Recapitulando, para duas dimensões:

$N(x_i|u_k, \Sigma_k) = \frac{1}{2\pi|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-u_k)^T\Sigma_k^{-1}(x_i-u_k)}$

Apresentam-se os valores de $p(c_k, x_i)$ na tabela seguinte:

k \ i	1	2	3
1	0.015	0.002	0.018
2	0.031	0.024	0.005

Com estes valores, construímos a tabela com os valores de γ_{ki}:

k \ i	1	2	3
1	0.322	0.092	0.770
2	0.678	0.908	0.230

3) Maximização

$N_k = \sum_{i=1}^{n} \gamma_{ki}$

$u_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ki} \cdot x_i$

$$\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ki} \cdot (x_i - u_k) \cdot (x_i - u_k)^T$$

$$\pi_k = p(c_k) = \frac{N_k}{N}$$

Calculando estes novos valores, ficamos em:

$N_1 = 1.183$ $\quad$ $N_2 = 1.817$

$u_1 = \begin{bmatrix} 2.223 \\ -0.496 \end{bmatrix}$ $\quad$ $u_2 = \begin{bmatrix} 0.754 \\ 0.873 \end{bmatrix}$

$\Sigma_1 = \begin{bmatrix} 1.182 & -0.849 \\ -0.849 & 0.714 \end{bmatrix}$ $\quad$ $\Sigma_2 = \begin{bmatrix} 0.947 & -1.039 \\ -1.039 & 1.364 \end{bmatrix}$

$\pi_1 = 0.394$ $\quad$ $\pi_2 = 0.606$

- Fazendo a segunda iteração do algoritmo:

2) Expetativa

Apresentam-se os valores de $p(c_k, x_i)$ na tabela seguinte:

k \ i	1	2	3
1	0.047	0.000	0.137
2	0.090	0.127	0.007

Com estes valores, construímos a tabela com os valores de γ_{ki}:

k \ i	1	2	3
1	0.342	0.004	0.953
2	0.658	0.996	0.047

3) Maximização

Calculando estes novos valores, ficamos em:

$N_1 = 1.299$ $\quad$ $N_2 = 1.701$

$u_1 = \begin{bmatrix} 2.465 \\ -0.728 \end{bmatrix}$ $\quad$ $u_2 = \begin{bmatrix} 0.469 \\ 1.144 \end{bmatrix}$

$\Sigma_1 = \begin{bmatrix} 0.793 & -0.407 \\ -0.407 & 0.215 \end{bmatrix}$ $\quad$ $\Sigma_2 = \begin{bmatrix} 0.414 & -0.619 \\ -0.619 & 1.061 \end{bmatrix}$

$\pi_1 = 0.433$ $\quad$ $\pi_2 = 0.567$

2)

a)

$$h_{MAP} = \underset{k}{argmax}\, p(c_k|x_i)$$

Tal como na questão anterior, construímos a tabela dos valores $p(c_k|x_i)$.

k \ i	1	2	3
1	0.244	0	0.824
2	0.173	0.268	0

Com estes valores, construímos a tabela com os valores de γ_{ki}:

k \ i	1	2	3
1	0.586	0	1
2	0.414	1	0

Logo, fazendo um hard assignment, temos:

$\widehat{x_1} = c_1$

$\widehat{x_2} = c_2$

$\widehat{x_3} = c_1$

b)

$a(x_1) = a(x_3) = d(x_1, x_3) = \sqrt{(1-3)^2 + (0+1)^2} = \sqrt{5}$

$b(x_1) = d(x_1, x_2) = \sqrt{(1-0)^2 + (0-2)^2} = \sqrt{5}$

$b(x_3) = d(x_3, x_2) = \sqrt{(3-0)^2 + (-1-2)^2} = 3\sqrt{2}$

Como $a(x_1) \geq b(x_1)$:

$$s(x_1) = \frac{b(x_1)}{a(x_1)} - 1 = \frac{\sqrt{5}}{\sqrt{5}} - 1 = 0$$

Como $a(x_3) < b(x_3)$:

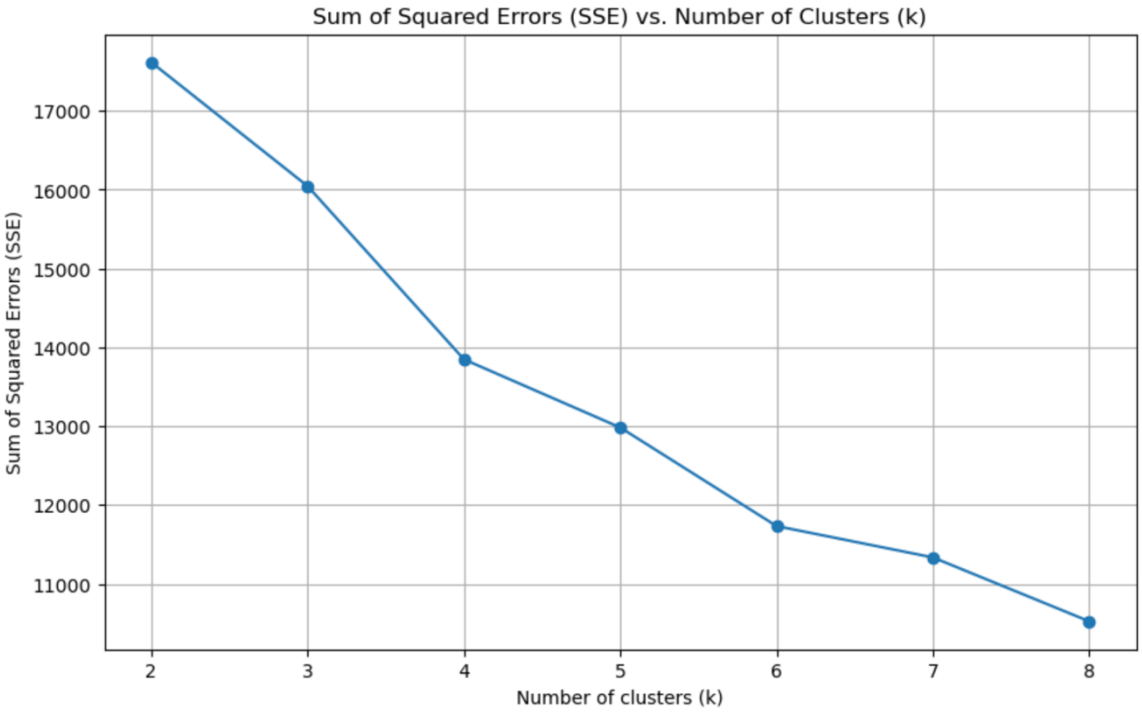
$$s(x_3) = 1 - \frac{a(x_3)}{b(x_3)} = 1 - \frac{\sqrt{5}}{3\sqrt{2}} = 0.473$$

$$s(c_1) = \frac{s(x_1) + s(x_3)}{2} = 0.236$$

II. Programming and critical analysis

1)

a)



b)

O "elbow method" consiste em identificar o ponto no gráfico onde a taxa de redução do SSE desacelera significativamente. Esse ponto geralmente indica o número ideal de clusters, pois adicionar mais clusters além desse ponto traz retornos decrescentes em termos de redução do SSE.

Observando o gráfico, a queda acentuada no SSE ocorre entre 2 e 4 clusters. Após 4 clusters, a redução no SSE se torna mais lenta, indicando que o "elbow point" parece estar em 4 clusters.

c)

O método k-means é eficaz para dados numéricos e contínuos, pois ele minimiza a variação dentro dos clusters usando a média dos dados. No entanto, o k-modes é uma variante do k-means projetada especificamente para lidar com dados categóricos, onde os pontos de dados podem ser representados pela moda (valores mais frequentes) em vez de médias.

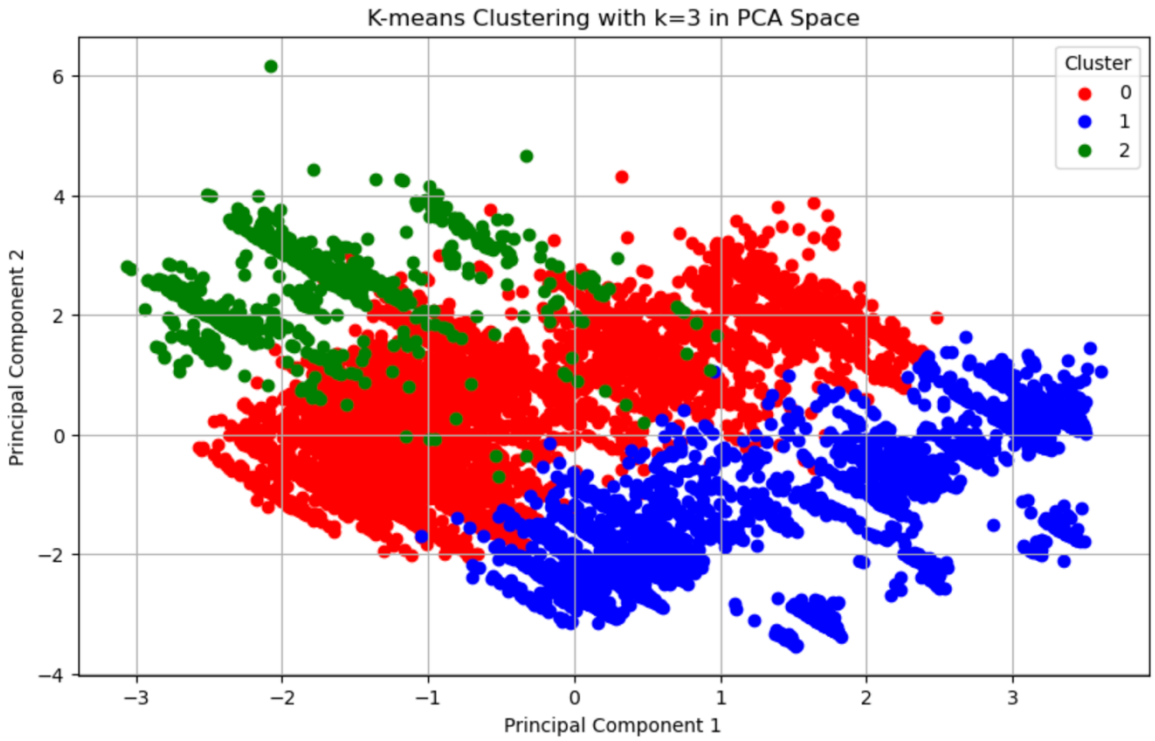
Se o conjunto de dados contém muitas variáveis categóricas (como ocupação, educação, ou outros atributos qualitativos), o k-modes poderia ser uma escolha melhor. Ele é mais apropriado para lidar com dados em que a média não é significativa, o que é o caso para variáveis não numéricas.

2)

a)

PCA explained variance (first two components): [0.117 0.111]

b)

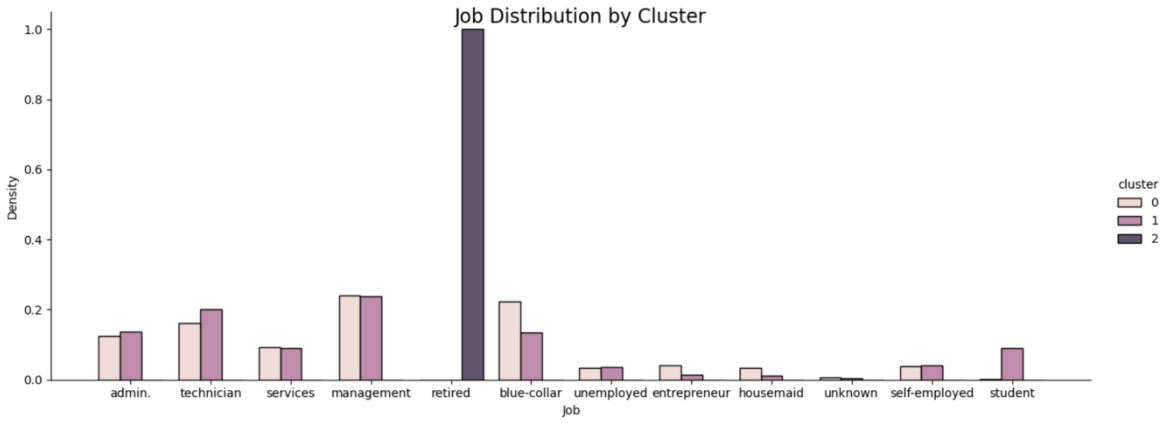


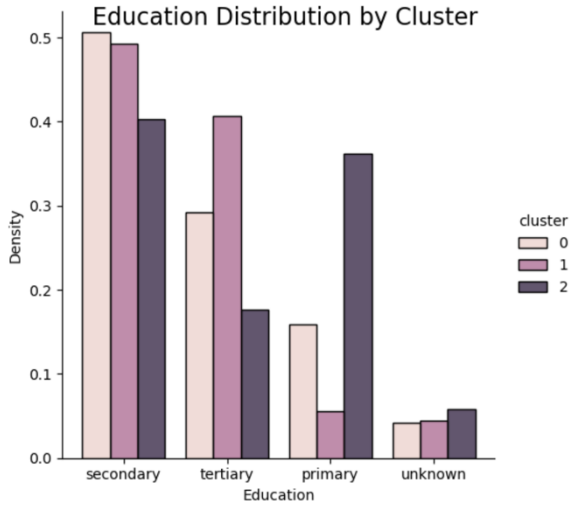
Os dois primeiros componentes principais explicam apenas cerca de 11,7% e 11,1% da variância dos dados, respectivamente. Isso significa que essas componentes capturam uma pequena fração da variabilidade total dos dados, sugerindo que pode haver mais informação relevante em outras componentes principais. Em termos práticos, isso limita a capacidade de separação visual dos clusters neste gráfico.

Os clusters 0, 1 e 2 (representados por cores diferentes) apresentam uma sobreposição considerável. Não há uma divisão nítida entre eles, indicando que os dados se distribuem de forma mista nas duas primeiras componentes principais. Apesar de alguma separação entre as cores, especialmente entre os clusters 1 e 2, a sobreposição entre os clusters sugere que o valor de k=3 não leva a uma separação muito clara na dimensão das duas primeiras componentes principais.

Com base na baixa variância explicada pelas duas primeiras componentes principais e na considerável sobreposição entre clusters, não é possível separar claramente os clusters apenas com essas duas componentes. Isso indica que uma análise de clustering completa pode exigir mais componentes principais para uma segmentação mais precisa.

c)





Observemos o gráfico relativamente à profissão.
O cluster 0 tem uma maior concentração de indivíduos em profissões operacionais, como "blue-collar", seguido de uma leve representação em diversas outras categorias de emprego.
O cluster 1 apresenta uma distribuição mais variada, incluindo profissões como "admin" e "student", sugerindo que este grupo pode incluir clientes mais jovens ou em início de carreira.
O cluster 2 tem uma proporção significativa de indivíduos reformados ("retired"), indicando que ele pode representar uma categoria demográfica mais velha ou de clientes inativos no mercado de trabalho.

Observemos o gráfico relativamente à educação.
O cluster 0 tem uma predominância de clientes com nível secundário de educação, o que pode indicar um perfil socioeconómico específico dentro deste grupo.
O cluster 1 apresenta uma maior variedade, com uma proporção considerável de clientes com nível terciário de educação, o que pode indicar um grupo mais jovem e com formação mais alta, como visto para a profissão.
O cluster 2 também possui uma representação alta de clientes com nível secundário, mas há uma presença significativa de clientes com nível primário, o que pode reforçar o perfil demográfico mais velho observado anteriormente.