Aprendizagem 2023/24

Homework IV – Group 029

(ist1103000, ist1103179)

I. Pen-and-paper

1

EM clustering algorithm

↳ E-step (I) ⟶ expectation

→ M-step (II) ⟶ maximigation

I.

$$\cdot\, p(y_2,y_3 \mid c_k = 1) = \mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\cdot\pi)^{D/2}} \cdot \frac{1}{|\Sigma_k|^{1/2}} \cdot \exp\left(-\tfrac{1}{2}\cdot (x-\mu_k)^T \Sigma_k^{-1} \cdot (x-\mu_k)\right)$$

$$\cdot\, p(c_k = 1, x_n) = \pi_k \cdot p(y_2,y_3 \mid c_k = 1) \cdot p(y_1 \mid c_k = 1)$$

$$\cdot\, p(x_n) = \sum_{k=1}^{k} p(c_k = 1, x_n)$$

$$\cdot\, \gamma(c_{nk}) = p(c_k = 1 \mid x_n) = \frac{p(c_k = 1, x_n)}{p(x_n)}$$

II.

$$\cdot\, N_k = \sum_{n=1}^{N} \gamma(c_{nk})$$

$$\cdot\, \mu_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(c_{nk}) \cdot x_n$$

$$\cdot\, \Sigma_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(c_{nk}) \cdot (x_n - \mu_k) \cdot (x_n - \mu_k)^T$$

$$\cdot\, \pi_k = p(c_k = 1) = \frac{N_k}{N}$$

1. Expectation

Given observations $x = \left\{ \begin{pmatrix} 1 \\ 0.6 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 0 \\ -0.4 \\ 0.8 \end{pmatrix}, \begin{pmatrix} 0 \\ 0.2 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 \\ 0.4 \\ -0.1 \end{pmatrix} \right\}$:

$$\boxed{\begin{aligned} \gamma(c_{11}) &= P(c_1 \mid x_1) \\ \gamma(c_{12}) &= P(c_2 \mid x_1) \end{aligned}} \quad \boxed{\begin{aligned} \gamma(c_{21}) &= P(c_1 \mid x_2) \\ \gamma(c_{22}) &= P(c_2 \mid x_2) \end{aligned}} \quad \boxed{\begin{aligned} \gamma(c_{31}) &= P(c_1 \mid x_3) \\ \gamma(c_{32}) &= P(c_2 \mid x_3) \end{aligned}}$$

$$\boxed{\begin{aligned} \gamma(c_{41}) &= P(c_1 \mid x_4) \\ \gamma(c_{42}) &= P(c_2 \mid x_4) \end{aligned}}$$

$$P(c_1, x_1) = \underbrace{\pi_1}_{0.5} \cdot P(y_2 = 0.6, y_3 = 0.1 \mid c_1 = 1) \cdot \underbrace{P(y_1 = 1 \mid c_1 = 1)}_{0.3}$$

Bernoulli Distribution

$$P(X = x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases}$$

$$P(y_2 = 0.6, y_3 = 0.1 \mid c_1 = 1) = \frac{1}{2\pi} \cdot \frac{1}{|\Sigma_1|^{1/2}} \exp\left(-\tfrac{1}{2} \begin{bmatrix} -0.4 & -0.9 \end{bmatrix} \Sigma_1^{-1} \begin{bmatrix} -0.4 \\ -0.9 \end{bmatrix}\right)$$

$$\begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.4 \\ -0.9 \end{bmatrix} \qquad |\Sigma_1|^{1/2} = 1.93649 \qquad \Sigma_1^{-1} = \frac{1}{3.75} \begin{bmatrix} 2 & -0.5 \\ -0.5 & 2 \end{bmatrix} = \begin{bmatrix} 0.53333 & -0.13333 \\ -0.13333 & 0.53333 \end{bmatrix}$$

And thus $P(y_2=0.6, y_3=0.1 \mid c_1=1) = \frac{1}{2\pi} \cdot \frac{1}{1.93649} \cdot \exp(-0.21067) \simeq$

$\simeq 0.06658$

We can now compute $\boxed{P(c_1, x_1) = 0.5 \times 0.06658 \times 0.3 = 9.987 \times 10^{-3}}$

$$P(c_2, x_1) = \underbrace{\pi_2}_{0.5} \cdot P(y_2=0.6, y_3=0.1 \mid c_2=1) \cdot \underbrace{P(y_1=1 \mid c_2=1)}_{0.7}$$

$$P(y_2=0.6, y_3=0.1 \mid c_2=1) = \frac{1}{2\pi} \cdot \frac{1}{1.11803} \exp\left(-\frac{1}{2} \begin{bmatrix} 0.6 & 0.1 \end{bmatrix} \begin{bmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix}\right) = *$$

$$\begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix} \quad |\Sigma_2|^{\frac{1}{2}} = 1.11803 \quad \Sigma_2^{-1} = \begin{bmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{bmatrix}$$

$* = 0.11962$

$\boxed{P(c_2, x_1) = 0.5 \times 0.11962 \times 0.7 = 0.04187}$

$$P(c_1, x_2) = \underbrace{\pi_1}_{0.5} \cdot P(y_2=-0.4, y_3=0.8 \mid c_1=1) \cdot \underbrace{P(y_1=0 \mid c_1=1)}_{0.7}$$

$$P(y_2=-0.4, y_3=0.8 \mid c_1=1) = \frac{1}{2\pi} \cdot \frac{1}{1.93649} \cdot \underbrace{\exp\left(-\frac{1}{2} \begin{bmatrix} -1.4 & -0.2 \end{bmatrix} \begin{bmatrix} 0.53333 & -0.13333 \\ -0.17333 & 0.53333 \end{bmatrix} \begin{bmatrix} -1.4 \\ -0.2 \end{bmatrix}\right)}_{**}$$

$$\begin{bmatrix} -0.4 \\ 0.8 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.4 \\ -0.2 \end{bmatrix}$$

$** = 0.05005$

$\boxed{P(c_1, x_2) = 0.5 \times 0.05005 \times 0.7 = 0.01752}$

$$P(c_2, x_2) = \underbrace{\pi_2}_{0.5} \cdot P(y_2=-0.4, y_3=0.8 \mid c_2=1) \cdot \underbrace{P(y_1=0 \mid c_2=1)}_{0.3}$$

$$P(y_2=-0.4, y_3=0.8 \mid c_2=1) = \frac{1}{2\pi} \cdot \frac{1}{1.11803} \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} -0.4 & 0.8 \end{bmatrix} \begin{bmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{bmatrix} \begin{bmatrix} -0.4 \\ 0.8 \end{bmatrix}\right) =$$

$= 0.06819$

$\boxed{P(c_2, x_2) = 0.5 \times 0.06819 \times 0.3 = 0.01023}$

$$P(c_1, x_3) = \underbrace{\pi_1}_{0.5} \cdot P(y_2=0.2, y_3=0.5 \mid c_1=1) \cdot \underbrace{P(y_1=0 \mid c_1=1)}_{0.7}$$

$$P(y_2=0.2, y_3=0.5 \mid c_1=1) = \frac{1}{2\pi} \cdot \frac{1}{1.93649} \cdot \exp\left(-\frac{1}{2}\begin{bmatrix}-0.8 & -0.5\end{bmatrix}\begin{bmatrix}0.53333 & -0.13333 \\ -0.13333 & 0.53333\end{bmatrix}\begin{bmatrix}-0.8 \\ -0.5\end{bmatrix}\right) =$$

$$= 0.06837$$

$$\boxed{P(c_1, x_3) = 0.5 \times 0.06837 \times 0.7 = 0.02393}$$

$$P(c_2, x_3) = \underbrace{\pi_2}_{0.5} \cdot P(y_2=0.2, y_3=0.5 \mid c_2=1) \cdot \underbrace{P(y_1=0 \mid c_2=1)}_{0.3}$$

$$P(y_2=0.2, y_3=0.5 \mid c_2=1) = \frac{1}{2\pi} \cdot \frac{1}{1.11803} \cdot \exp\left(-\frac{1}{2}\begin{bmatrix}0.2 & 0.5\end{bmatrix}\begin{bmatrix}1.2 & -0.8 \\ -0.8 & 1.2\end{bmatrix}\begin{bmatrix}0.2 \\ 0.5\end{bmatrix}\right) =$$

$$= 0.12958$$

$$\boxed{P(c_2, x_3) = 0.5 \times 0.12958 \times 0.3 = 0.01944}$$

$$P(c_1, x_4) = \underbrace{\pi_1}_{0.5} \cdot P(y_2=0.4, y_3=-0.1 \mid c_1=1) \cdot \underbrace{P(y_1=1 \mid c_1=1)}_{0.3}$$

$$P(y_2=0.4, y_3=-0.1 \mid c_1=1) = \frac{1}{2\pi} \cdot \frac{1}{1.93649} \cdot \exp\left(-\frac{1}{2}\begin{bmatrix}-0.6 & -1.1\end{bmatrix}\begin{bmatrix}0.53333 & -0.13333 \\ -0.13333 & 0.53333\end{bmatrix}\begin{bmatrix}-0.6 \\ -1.1\end{bmatrix}\right)$$

$$= 0.05905$$

$$\boxed{P(c_1, x_4) = 0.5 \times 0.05905 \times 0.3 = 8.8575 \times 10^{-3}}$$

$$P(c_2, x_4) = \underbrace{\pi_2}_{0.5} \cdot P(y_2=0.4, y_3=-0.1 \mid c_2=1) \cdot \underbrace{P(y_1=1 \mid c_2=1)}_{0.7}$$

$$P(y_2=0.4, y_3=-0.1 \mid c_2=1) = \frac{1}{2\pi} \cdot \frac{1}{1.11803} \cdot \exp\left(-\frac{1}{2}\begin{bmatrix}0.4 & -0.1\end{bmatrix}\begin{bmatrix}1.2 & -0.8 \\ -0.8 & 1.2\end{bmatrix}\begin{bmatrix}0.4 \\ -0.1\end{bmatrix}\right) =$$

$$= 0.12450$$

$$\boxed{P(c_2, x_4) = 0.5 \times 0.12450 \times 0.7 = 0.04358}$$

Taking the following values into consideration,

$$\boxed{P(x_1) = 9.987 \times 10^{-3} + 0.04187 = 0.05186}$$

$$\boxed{P(x_2) = 0.01752 + 0.01023 = 0.02775}$$

$$\boxed{P(x_3) = 0.02393 + 0.01944 = 0.04337}$$

$$\boxed{P(x_4) = 8.8575 \times 10^{-3} + 0.04358 = 0.05244}$$

we can now compute:

$$\gamma(c_{11}) = P(c_1|x_1) = \frac{P(c_1, x_1)}{P(x_1)} = \frac{9.987\times10^{-3}}{0.05186} \simeq 0.19258$$

$$\gamma(c_{12}) = P(c_2|x_1) = \frac{P(c_2, x_1)}{P(x_1)} = \frac{0.04187}{0.05186} \simeq 0.80737$$

$$\gamma(c_{21}) = \frac{0.01752}{0.02775} \simeq 0.63135$$

$$\gamma(c_{22}) = \frac{0.01023}{0.02775} \simeq 0.36865$$

$$\gamma(c_{31}) = \frac{0.02393}{0.04337} \simeq 0.55176$$

$$\gamma(c_{32}) = \frac{0.01944}{0.04337} \simeq 0.44824$$

$$\gamma(c_{41}) = \frac{8.8575\times10^{-3}}{0.05244} \simeq 0.16891$$

$$\gamma(c_{42}) = \frac{0.04358}{0.05244} \simeq 0.83105$$

2. Maximization

we evaluate

$$N_k = \sum_{n=1}^{N} \gamma(c_{nk})$$

$$N_1 = 0.19258 + 0.63135 + 0.55176 + 0.16891 = 1.5446$$

$$N_2 = 0.80737 + 0.36865 + 0.44824 + 0.83105 = 2.45531$$

we determine the mean values (and $p(y_1=1|c_1)$)

$$\mu_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(c_{nk}) \cdot x_n$$

$$\mu_1 = \frac{1}{1.5446}\left(0.19258 \cdot \begin{pmatrix}1\\0.6\\0.1\end{pmatrix} + 0.63135\begin{pmatrix}0\\-0.4\\0.8\end{pmatrix} + 0.55176\begin{pmatrix}0\\0.2\\0.5\end{pmatrix} + 0.16891\begin{pmatrix}1\\0.4\\-0.1\end{pmatrix}\right) = \begin{bmatrix}0.23403\\0.02649\\0.50714\end{bmatrix} \rightarrow \begin{array}{l} P(y_1=1|c_1) = 0.23403 \\ \mu_1 = \begin{bmatrix}0.02649\\0.50714\end{bmatrix}\end{array}$$

$$\mu_2 = \frac{1}{2.45531}\left(0.80737\begin{pmatrix}1\\0.6\\0.1\end{pmatrix} + 0.36865\begin{pmatrix}0\\-0.4\\0.8\end{pmatrix} + 0.44824\begin{pmatrix}0\\0.2\\0.5\end{pmatrix} + 0.83105\begin{pmatrix}1\\0.4\\-0.1\end{pmatrix}\right) = \begin{bmatrix}0.66730\\0.30914\\0.21043\end{bmatrix} \rightarrow \begin{array}{l} P(y_1=1|c_2) = 0.66730 \\ \mu_2 = \begin{bmatrix}0.30914\\0.21043\end{bmatrix}\end{array}$$

and the new covariance matrix

$$\Sigma_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(c_{nk}) \cdot (x_n - \mu_k) \cdot (x_n - \mu_k)^T$$

$$\Sigma_1 = \frac{1}{1.5446}\Big(0.19258 \cdot \begin{pmatrix} 0.6 - 0.02649 \\ 0.1 - 0.50714 \end{pmatrix}\begin{pmatrix} 0.6-0.02649 & 0.1-0.50714 \end{pmatrix} +$$

$$+ 0.63135 \begin{pmatrix} -0.4 - 0.02649 \\ 0.8 - 0.50714 \end{pmatrix}\begin{pmatrix} -0.4-0.02649 & 0.8-0.50714 \end{pmatrix} +$$

$$+ 0.55176 \begin{pmatrix} 0.2 - 0.02649 \\ 0.5 - 0.50714 \end{pmatrix}\begin{pmatrix} 0.2-0.02649 & 0.5-0.50714 \end{pmatrix} +$$

$$+ 0.16891 \begin{pmatrix} 0.4 - 0.02649 \\ -0.1 - 0.50714 \end{pmatrix}\begin{pmatrix} 0.4-0.02649 & -0.1-0.50714 \end{pmatrix}\Big) =$$

$$= \begin{bmatrix} 0.14137 & -0.10541 \\ -0.10541 & 0.09605 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} 0.10829 & -0.08865 \\ -0.08865 & 0.10412 \end{bmatrix}$$

and the new mixing parameter is

$$\pi_k = p(c_k = 1) = \frac{N_k}{N}$$

$$\pi_1 = p(c_1 = 1) = \frac{1.5446}{4} \simeq 0.38615$$

$$\pi_2 = p(c_2 = 1) = \frac{2.45531}{4} \simeq 0.61383$$

2

$$x_{new} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \end{pmatrix}$$

$\boxed{\text{posteriors: } p(c_1 | x_{new}) \text{ and } p(c_2 | x_{new})}$

$$p(c_1, x_{new}) = \underbrace{\pi_1}_{0.38615} \cdot p(y_2 = 0.3, y_3 = 0.7 \mid c_1 = 1) \cdot \underbrace{p(y_1 = 1 \mid c_1 = 1)}_{0.23403}$$

$$P(y_2=0.3, y_3=0.7 \mid C_1=1) = \frac{1}{2\pi}\frac{1}{|\Sigma_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}\cdot(x-\mu_1)\right) = *$$

$$|\Sigma_1|^{\frac{1}{2}} = 0.04967 \quad \begin{bmatrix}0.3\\0.7\end{bmatrix} - \begin{bmatrix}0.02649\\0.50714\end{bmatrix} = \begin{bmatrix}0.27351\\0.19286\end{bmatrix}$$

$$\Sigma_1^{-1} = \begin{bmatrix}38.92887 & 42.72246\\42.72246 & 57.29698\end{bmatrix}$$

$$* = \frac{1}{2\pi}\cdot\frac{1}{0.04967}\cdot\exp(-4.77524) \simeq 0.02703$$

and thus $P(C_1, x_{new}) = 0.38615 \times 0.02703 \times 0.23403 =$
$= 2.44272 \times 10^{-3}$

$$P(C_2, x_{new}) = \underbrace{\pi_2}_{0.61383}\cdot P(y_2=0.3, y_3=0.7 \mid C_2=1)\cdot \underbrace{P(y_1=1 \mid C_2=1)}_{0.66730}$$

$$P(y_2=0.3, y_3=0.7 \mid C_2=1) = \frac{1}{2\pi}\frac{1}{|\Sigma_2|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}\cdot(x-\mu_2)\right) =$$
$$= \frac{1}{2\pi}\cdot\frac{1}{0.05845}\cdot\exp(-3.68380) \simeq 0.06842$$

$$|\Sigma_2|^{\frac{1}{2}} \simeq 0.05845 \quad \begin{bmatrix}0.3\\0.7\end{bmatrix} - \begin{bmatrix}0.30914\\0.21043\end{bmatrix} = \begin{bmatrix}-9.14\times10^{-3}\\0.48957\end{bmatrix}$$

$$\Sigma_2^{-1} = \begin{bmatrix}30.47713 & 25.94888\\25.94888 & 31.69774\end{bmatrix}$$

and thus $P(C_2, x_{new}) = 0.61383 \times 0.06842 \times 0.66730 \simeq$
$\simeq 0.02803$

$$P(x_{new}) = P(C_1, x_{new}) + P(C_2, x_{new}) =$$
$$= 2.44272\times10^{-3} + 0.02803 \simeq 0.03047$$

$$\boxed{\begin{aligned} P(C_1 \mid x_{new}) &= \frac{2.44272\times10^{-3}}{0.03047} \simeq 0.08017\\ P(C_2 \mid x_{new}) &= \frac{0.02803}{0.03047} \simeq 0.91992\end{aligned}}$$

3

under a ML assumption, we can assign each observation to a cluster by computing $\arg\max_{c_k} P(x \mid c_k)$:

$$P(x_1 \mid c_1) = \frac{P(c_1, x_1)}{P(c_1)} = \frac{9.987 \times 10^{-3}}{0.38615} = 0.02586$$

$$P(x_1 \mid c_2) = \frac{0.04187}{0.61383} = 0.06821$$

$\rightarrow$ x_1 should be assigned to c_2

$$P(x_2 \mid c_1) = \frac{0.01752}{0.38615} = 0.04537$$

$$P(x_2 \mid c_2) = \frac{0.01023}{0.61383} = 0.01667$$

$\rightarrow$ x_2 should be assigned to c_1

$$P(x_3 \mid c_1) = \frac{0.02393}{0.38615} = 0.06197$$

$$P(x_3 \mid c_2) = \frac{0.01944}{0.61383} = 0.03167$$

$\rightarrow$ x_3 should be assigned to c_1

$$P(x_4 \mid c_1) = \frac{8.8575 \times 10^{-3}}{0.38615} = 0.02294$$

$$P(x_4 \mid c_2) = \frac{0.04358}{0.61383} = 0.07100$$

$\rightarrow$ x_4 should be assigned to c_2

thus, clusters $= \{ c_1 = \{x_2, x_3\}, c_2 = \{x_1, x_4\} \}$

Silhouette score (for both clusters)

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}, \quad a(i) = \frac{1}{\text{size of cluster} - 1} \sum_{j \text{ in the same cluster as } i} d(i,j),$$

$$b(i) = \min_{c \neq \text{cluster of } i} \left(\frac{1}{\text{size of cluster } c} \sum_{j \text{ in cluster } c} d(i,j) \right)$$

manhattan distance

$$d(x_1,x_2) = \underbrace{|1-0| + |0.6-(-0.4)| + |0.1-0.8|}_{\text{manhattan distance}} =$$
$$= 2.7$$

$d(x_1,x_3) = 1.8$
$d(x_1,x_4) = 0.4$
$d(x_2,x_3) = 0.9$
$d(x_2,x_4) = 2.7$
$d(x_3,x_4) = 1.8$

$$S(c_1) = \sum_{x_i \in c_1} s(x_i)$$

$$s(x_2) = \frac{b(x_2) - a(x_2)}{\max\{a(x_2), b(x_2)\}} = 1 - \frac{a(x_2)}{b(x_2)} = 1 - \frac{0.9}{2.7} = \frac{2}{3}$$

$a(x_2) = 0.9$
$b(x_2) = \frac{1}{2}(2.7 + 2.7) = 2.7$

$$s(x_3) = 1 - \frac{a(x_3)}{b(x_3)} = 1 - \frac{0.9}{1.8} = \frac{1}{2}$$

$a(x_3) = 0.9$
$b(x_3) = \frac{1}{2}(1.8 + 1.8) = 1.8$

$$\boxed{S(c_1) = s(x_2) + s(x_3) = \frac{2}{3} + \frac{1}{2} = \frac{7}{6} = 1.1(6)}$$

$$S(c_2) = \sum_{x_i \in c_2} s(x_i)$$

$$s(x_1) = 1 - \frac{a(x_1)}{b(x_1)} = 1 - \frac{0.4}{2.25} = \frac{37}{45}$$

$a(x_1) = 0.4$
$b(x_1) = \frac{1}{2}(2.7 + 1.8) = 2.25$

$$s(x_4) = 1 - \frac{a(x_4)}{b(x_4)} = 1 - \frac{0.4}{2.25} = \frac{37}{45}$$

$$a(x_4) = 0.4$$

$$b(x_4) = \frac{1}{2}(2.7 + 1.8) = 2.25$$

$$\boxed{s(c_2) = s(x_1) + s(x_4) = 2 \times \frac{37}{45} = 1.6(4)}$$

4) Given the fact that purity is the number of correctly matched class and cluster labels divided by the number of total datapoints and that

$$\frac{1}{4} \sum_{k=1}^{2} (\arg\max(|c_k \cap g_j|) = 0.75$$

we can infer that $0{,}75 \times 4 = 3$ is the maximum number $\sum_{k=1}^{2} (\arg\max(|c_k \cap g_j|))$ can be.

Since clusters $= \{ c_1 = \{x_2, x_3\}, c_2 = \{x_1, x_4\} \}$,

$\arg\max(|c_k \cap g_j|) \in \{1, 2\}$.

Thus, $\sum_{k=1}^{2} (\arg\max(|c_k \cap g_j|))$ can only be 3 if

number of observations per class

$$\sum_{k=1}^{2} (\arg\max(|c_k \cap g_j|)) = \arg\max_{c_1}(1, 1) + \arg\max_{c_2}(0, 2) =$$

$= 2 + 1 = 3$ (there's also the possibility of swapping these values, but what matters is that two observations of one cluster are part of the same class and the other two of the remaining cluster are part of different classes.)

Taking this into consideration, we can conclude that the number of possible classes is ≥ 2.

II. Programming and critical analysis

1. Code:

```
import pandas as pd
from scipy.io.arff import loadarff

import warnings
warnings.filterwarnings("ignore")

data = loadarff('column_diagnosis.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')
df.head()

from sklearn import metrics, cluster
from sklearn.preprocessing import MinMaxScaler
import numpy as np

def purity_score(y_true, y_pred):
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(confusion_matrix, axis=0)) /
np.sum(confusion_matrix)

X = df.drop("class", axis=1)
y = df["class"]
X_normalized = MinMaxScaler().fit_transform(X)
kmeans_list = []

for k in [2, 3, 4, 5]:
    print("k =", k)
    kmeans = cluster.KMeans(n_clusters=k, random_state=0).fit(X_normalized)
    kmeans_list.append(kmeans)
    print("\tSilhouette Score =", metrics.silhouette_score(X_normalized,
kmeans.labels_))
    print("\tPurity =", purity_score(y, kmeans.labels_))
```

```
k = 2
	Silhouette Score = 0.36081773371884557
	Purity = 0.6290322580645161

k = 3
	Silhouette Score = 0.29579055730002257
	Purity = 0.667741935483871

k = 4
	Silhouette Score = 0.2686566721650703
```

```
Purity = 0.6612903225806451
```

```
k = 5
    Silhouette Score = 0.24328260038805272
    Purity = 0.6741935483870968
```

2.

i. Code:

```python
from sklearn.decomposition import PCA

learnt_pca = PCA(n_components=2)
learnt_pca.fit(X_normalized)
X_trans = learnt_pca.transform(X_normalized)

top2_explained_variance = learnt_pca.explained_variance_ratio_
print(top2_explained_variance[0]+top2_explained_variance[1])
```

```
0.77137397434354
```

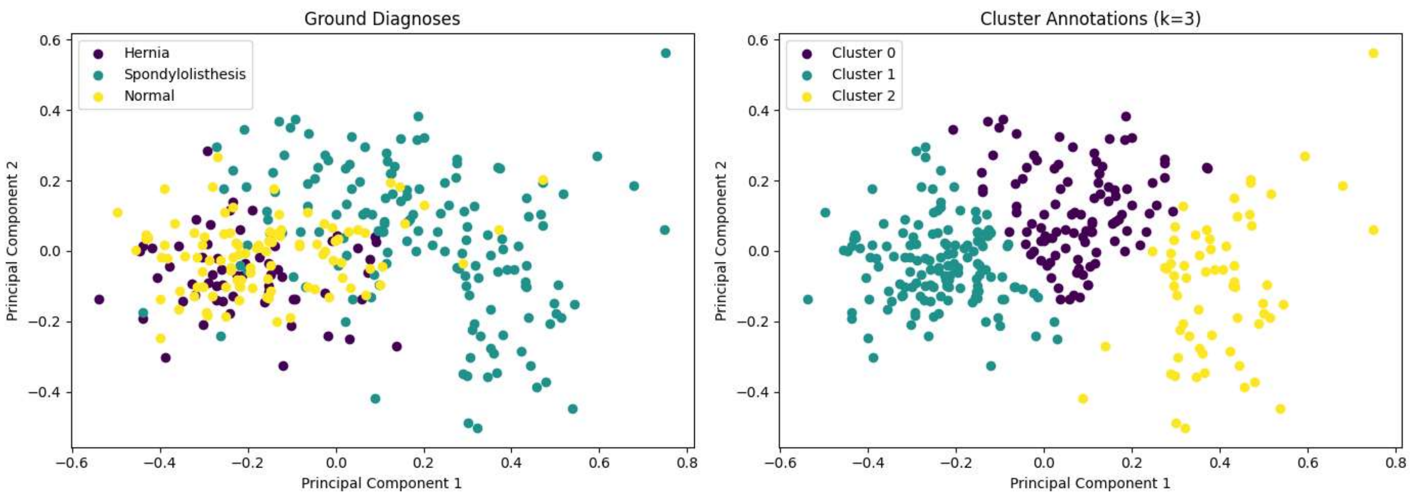
ii. Code:

```python
labels = list(df.columns)
labels.remove("class")

weights_component1 = list(abs(learnt_pca.components_[0]))
weights_component2 = list(abs(learnt_pca.components_[1]))

input_vars_by_relevance1= [x for _, x in sorted(zip(weights_component1,
labels), key=lambda pair: pair[0])]
input_vars_by_relevance2= [x for _, x in sorted(zip(weights_component2,
labels), key=lambda pair: pair[0])]

print("Most relevant imput variables by component (in ascending order
of importance):")
print("- Component 1:", input_vars_by_relevance1)
print("- Component 2:", input_vars_by_relevance2)
```

```
Most relevant imput variables by component (in ascending order of
importance):
- Component 1: ['pelvic_radius', 'degree_spondylolisthesis',
'sacral_slope', 'pelvic_tilt', 'lumbar_lordosis_angle',
'pelvic_incidence']
- Component 2: ['degree_spondylolisthesis', 'lumbar_lordosis_angle',
'pelvic_incidence', 'sacral_slope', 'pelvic_radius', 'pelvic_tilt']
```

3. Code and graphs:

```
import matplotlib.pyplot as plt

X_2d = X_trans
cluster_labels = kmeans_list[1].labels_

class_labels = df['class'].unique()

colors = plt.cm.viridis(np.linspace(0, 1, len(class_labels)))

fig, ax = plt.subplots(1, 2, figsize=(14, 5))

ax[0].set_title('Ground Diagnoses')
ax[0].set_xlabel('Principal Component 1')
ax[0].set_ylabel('Principal Component 2')
for i, label in enumerate(class_labels):
    indices = df['class'] == label
    ax[0].scatter(X_2d[indices, 0], X_2d[indices, 1], label=label,
color=colors[i])
ax[0].legend()

ax[1].set_title('Cluster Annotations (k=3)')
ax[1].set_xlabel('Principal Component 1')
ax[1].set_ylabel('Principal Component 2')
for i in range(len(class_labels)):
    indices = cluster_labels == i
    ax[1].scatter(X_2d[indices, 0], X_2d[indices, 1], label=f'Cluster {i}',
color=colors[i])
ax[1].legend()

plt.tight_layout()
plt.show()
```



4. We can find natural clusters within the population based on the features used for clustering by applying algorithms like K-means to the dataset. These clusters might highlight trends or associations between people who have similar traits. When it comes to health, it's probable that some clusters correspond to people with comparable diseases or risk factors, while other clusters correspond to people who are generally in good health.

   Moreover, clustering can be used to divide the population into various groups, each with a distinct health profile. For instance, people with high-risk health profiles may form one cluster, and people with low-risk profiles may form another. Medical professionals may find this segmentation useful as it enables them to customize interventions or treatments for particular population subgroups.

END