Machine Learning - Homework 4

Pedro Curvo (ist1102716) | Salvador Torpes (ist1102474)

1st Term - 23/24

Pen and Paper Exercises

Dataset

In the following exercise our goal is to consider a Bayesian Clustering model in order to separate the observations into 2 different clusters:

$$x_1 = \begin{bmatrix} 1 \\ 0.6 \\ 0.1 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 0 \\ -0.4 \\ 0.8 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 0 \\ 0.2 \\ 0.5 \end{bmatrix} \qquad x_4 = \begin{bmatrix} 1 \\ 0.4 \\ -0.1 \end{bmatrix}$$

We are working with 3 different variables (y_1, y_1, y_3) for each observation. In addition, we are assuming:

1. $\{y_1\} \perp \{y_2, y_3\}$
2. y_1 follows a Bernoulli distribution with parameter p: $y_1 \sim \text{Bernoulli}(p)$

$$P(y_1 = 1) = p \qquad P(y_1 = 0) = 1 - p$$

3. y_2 and y_3 follow a multivariate gaussian distribution with parameters $\vec{\mu}$ and Σ: $y_2, y_3 \sim \mathcal{N}(\vec{\mu}, \Sigma)$

$$P(\vec{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right)$$

$$\vec{x} = (y_2, y_3)$$

1st Question

Computing the Responsibilities

First, we need to compute the responsibility of each cluster for each observation. The responsibility γ_{ki} is defined as the probability of belonging to cluster k for observation i:

$$\gamma_{ki} = P(c_k|\vec{x_i}) =_{\text{Bayes}} \frac{P(\vec{x_i}|c_k)P(c_k)}{P(\vec{x_i})} = \frac{P(\vec{x_i}|c_k)P(c_k)}{\sum_{j=1}^{K} P(\vec{x_i}|c_j)P(c_j)}$$

$$\sum_{j=1}^{K} P(\vec{x_i}|c_j)P(c_j) = P(\vec{x_i})$$

Where c_k is the cluster k, K is the number of clusters and $\vec{x_i}$ is the observation i. In addition, the probability of belonging to cluster k, $P(c_k)$, is represented by the mixing coefficient π_k:

$$P(c_k) = \pi_k$$

In order to compute the responsibilities, we're told to use the following parameters for each cluster's y_1 and y_2, y_3 distributions:

Cluster	π_k	Parameters for $\{y_1\}$	Parameters for $\{y_2, y_3\}$
1 (c_1)	$P(c_1) = \pi_1 = 0.5$	$p_1 = P(y_1 = 1) = 0.3$	$\{y_2, y_3\} \sim \mathcal{N}\left(\vec{\mu}_1 = \begin{bmatrix}1\\1\end{bmatrix}, \Sigma_1 = \begin{bmatrix}2 & 0.5\\0.5 & 2\end{bmatrix}\right)$
2 (c_2)	$P(c_2) = \pi_2 = 0.5$	$p_2 = P(y_1 = 1) = 0.7$	$\{y_2, y_3\} \sim \mathcal{N}\left(\vec{\mu}_2 = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_2 = \begin{bmatrix}1.5 & 1\\1 & 1.5\end{bmatrix}\right)$

Table 1: Initial Parameters for each cluster

Responsibilities for $\vec{x_1}$

First of all, we will compute $P(\vec{x_1}|c_k)P(c_k)$ for clusters c_1 and c_2:

$$P(\vec{x_1}|c_1)P(c_1) = P(y_1 = 1|c_1)P(\{y_2, y_3\} = \{0.6, 0.1\}|c_1)\pi_1 = 0.3 \cdot 0.06658 \cdot 0.5 = 0.00999$$

$$P(\vec{x_1}|c_2)P(c_2) = P(y_1 = 1|c_2)P(\{y_2, y_3\} = \{0.6, 0.1\}|c_2)\pi_2 = 0.7 \cdot 0.11962 \cdot 0.5 = 0.04187$$

And then, we can compute γ_{ki}:

$$\gamma_{11} = P(c_1|\vec{x_1}) = \frac{P(\vec{x_1}|c_1)P(c_1)}{P(\vec{x_1})} = \frac{P(\vec{x_1}|c_1)P(c_1)}{\sum_{j=1}^{2} P(\vec{x_1}|c_j)P(c_j)} = \frac{0.00999}{0.00999 + 0.04187} = 0.19259$$

$$\gamma_{21} = P(c_2|\vec{x_1}) = \frac{P(\vec{x_1}|c_2)P(c_2)}{P(\vec{x_1})} = \frac{P(\vec{x_1}|c_2)P(c_2)}{\sum_{j=1}^{2} P(\vec{x_1}|c_j)P(c_j)} = \frac{0.04187}{0.00999 + 0.04187} = 0.80741$$

Responsibilities for $\vec{x}_2$

First of all, we will compute $P(\vec{x}_2|c_k)P(c_k)$ for clusters c_1 and c_2:

$$P(\vec{x}_2|c_1)P(c_1) = P(y_1 = 0|c_1)P(\{y_2, y_3\} = \{-0.4, 0.8\}|c_1)\pi_1 = 0.7 \cdot 0.05005 \cdot 0.5 = 0.01752$$

$$P(\vec{x}_2|c_2)P(c_2) = P(y_1 = 0|c_2)P(\{y_2, y_3\} = \{-0.4, 0.8\}|c_2)\pi_2 = 0.3 \cdot 0.06819 \cdot 0.5 = 0.01023$$

And then, we can compute γ_{ki}:

$$\gamma_{12} = P(c_1|\vec{x}_2) = \frac{P(\vec{x}_2|c_1)P(c_1)}{P(\vec{x}_2)} = \frac{P(\vec{x}_2|c_1)P(c_1)}{\sum_{j=1}^{2} P(\vec{x}_2|c_j)P(c_j)} = \frac{0.01752}{0.01752 + 0.01023} = 0.63135$$

$$\gamma_{22} = P(c_2|\vec{x}_2) = \frac{P(\vec{x}_2|c_2)P(c_2)}{P(\vec{x}_2)} = \frac{P(\vec{x}_2|c_2)P(c_2)}{\sum_{j=1}^{2} P(\vec{x}_2|c_j)P(c_j)} = \frac{0.01023}{0.01752 + 0.01023} = 0.36865$$

Responsibilities for $\vec{x}_3$

First of all, we will compute $P(\vec{x}_3|c_k)P(c_k)$ for clusters c_1 and c_2:

$$P(\vec{x}_3|c_1)P(c_1) = P(y_1 = 0|c_1)P(\{y_2, y_3\} = \{0.2, 0.5\}|c_1)\pi_1 = 0.7 \cdot 0.06837 \cdot 0.5 = 0.02393$$

$$P(\vec{x}_3|c_2)P(c_2) = P(y_1 = 0|c_2)P(\{y_2, y_3\} = \{0.2, 0.5\}|c_2)\pi_2 = 0.3 \cdot 0.12958 \cdot 0.5 = 0.01944$$

And then, we can compute γ_{ki}:

$$\gamma_{13} = P(c_1|\vec{x}_3) = \frac{P(\vec{x}_3|c_1)P(c_1)}{P(\vec{x}_3)} = \frac{P(\vec{x}_3|c_1)P(c_1)}{\sum_{j=1}^{2} P(\vec{x}_3|c_j)P(c_j)} = \frac{0.02393}{0.02393 + 0.01944} = 0.55181$$

$$\gamma_{23} = P(c_2|\vec{x}_3) = \frac{P(\vec{x}_3|c_2)P(c_2)}{P(\vec{x}_3)} = \frac{P(\vec{x}_3|c_2)P(c_2)}{\sum_{j=1}^{2} P(\vec{x}_3|c_j)P(c_j)} = \frac{0.01944}{0.02393 + 0.01944} = 0.44819$$

Responsibilities for $\vec{x}_4$

First of all, we will compute $P(\vec{x}_4|c_k)P(c_k)$ for clusters c_1 and c_2:

$$P(\vec{x}_4|c_1)P(c_1) = P(y_1 = 1|c_1)P(\{y_2, y_3\} = \{0.4, -0.1\}|c_1)\pi_1 = 0.3 \cdot 0.05905 \cdot 0.5 = 0.00886$$

$$P(\vec{x}_4|c_2)P(c_2) = P(y_1 = 1|c_2)P(\{y_2, y_3\} = \{0.4, -0.1\}|c_2)\pi_2 = 0.7 \cdot 0.12450 \cdot 0.5 = 0.04358$$

And then, we can compute γ_{ki}:

$$\gamma_{14} = P(c_1|\vec{x}_4) = \frac{P(\vec{x}_4|c_1)P(c_1)}{P(\vec{x}_4)} = \frac{P(\vec{x}_4|c_1)P(c_1)}{\sum_{j=1}^{2} P(\vec{x}_4|c_j)P(c_j)} = \frac{0.00886}{0.00886 + 0.04358} = 0.16892$$

$$\gamma_{24} = P(c_2|\vec{x}_4) = \frac{P(\vec{x}_4|c_2)P(c_2)}{P(\vec{x}_4)} = \frac{P(\vec{x}_4|c_2)P(c_2)}{\sum_{j=1}^{2} P(\vec{x}_4|c_j)P(c_j)} = \frac{0.04358}{0.00886 + 0.04358} = 0.83108$$

Responsibilities

$\gamma_{11} = 0.19259$	$\gamma_{12} = 0.63135$	$\gamma_{13} = 0.55181$	$\gamma_{14} = 0.16892$
$\gamma_{21} = 0.80741$	$\gamma_{22} = 0.36865$	$\gamma_{23} = 0.44819$	$\gamma_{24} = 0.83108$

M-Step

In the M-Step (Maximization Step), we will compute the new parameters for each cluster (we need to update the parameters in the table 1).

New Parameters for Cluster c_1

$$\vec{\mu}_{1_{\text{new}}} = \frac{\sum_{i=1}^{4} \gamma_{1i} \vec{x_i}}{\sum_{i=1}^{4} \gamma_{1i}} = \begin{bmatrix} 0.02651 & 0.50713 \end{bmatrix}$$

$$\Sigma_{1_{\text{new}}} = \frac{\sum_{i=1}^{4} \gamma_{1i} \left(\vec{x_i} - \vec{\mu}_{1_{\text{new}}}\right) \left(\vec{x_i} - \vec{\mu}_{1_{\text{new}}}\right)^T}{\sum_{i=1}^{4} \gamma_{1i}} = \begin{bmatrix} 0.14137 & -0.10541 \\ -0.10541 & 0.09605 \end{bmatrix}$$

$$p_{1_{\text{new}}} = \frac{\sum_{i=1}^{4} \gamma_{1i} \vec{x_i}}{\sum_{i=1}^{4} \gamma_{1i}} = 0.23404$$

New Parameters for Cluster c_2

$$\vec{\mu}_{2_{\text{new}}} = \frac{\sum_{i=1}^{4} \gamma_{2i} \vec{x_i}}{\sum_{i=1}^{4} \gamma_{2i}} = \begin{bmatrix} 0.30914 & 0.21042 \end{bmatrix}$$

$$\Sigma_{2_{\text{new}}} = \frac{\sum_{i=1}^{4} \gamma_{2i} \left(\vec{x_i} - \vec{\mu}_{2_{\text{new}}}\right) \left(\vec{x_i} - \vec{\mu}_{2_{\text{new}}}\right)^T}{\sum_{i=1}^{4} \gamma_{2i}} = \begin{bmatrix} 0.10829 & -0.08865 \\ -0.08865 & 0.10412 \end{bmatrix}$$

$$p_{2_{\text{new}}} = \frac{\sum_{i=1}^{4} \gamma_{2i} \vec{x_i}}{\sum_{i=1}^{4} \gamma_{2i}} = 0.66732$$

New table with the updated parameters

Cluster	π_k	Parameters for $\{y_1\}$	Parameters for $\{y_2, y_3\}$
1 (c_1)	$P(c_1) = \pi_1 = 0.5$	$p_1 = 0.23404$	$\{y_2, y_3\} \sim \mathcal{N}\left(\vec{\mu}_1 = \begin{bmatrix} 0.02651 \\ 0.50713 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 0.14137 & -0.10541 \\ -0.10541 & 0.09605 \end{bmatrix}\right)$
2 (c_2)	$P(c_2) = \pi_2 = 0.5$	$p_2 = 0.66732$	$\{y_2, y_3\} \sim \mathcal{N}\left(\vec{\mu}_2 = \begin{bmatrix} 0.30914 \\ 0.21042 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.10829 & -0.08865 \\ -0.08865 & 0.10412 \end{bmatrix}\right)$

Table 2: New Parameters for each cluster

2nd Question

We now have a new observation $\vec{x}_{\text{new}}$ and want to compute the probability of belonging to each cluster. The new observation is:

$$\vec{x}_{\text{new}} = \begin{bmatrix} 1 \\ 0.3 \\ 0.7 \end{bmatrix}$$

First of all, we will compute $P(\vec{x}_{\text{new}}|c_k)P(c_k)$ for clusters c_1 and c_2:

$$P(\vec{x}_{\text{new}}|c_1)P(c_1) = P(y_1 = 1|c_1)P(\{y_2, y_3\} = \{0.3, 0.7\}|c_1)\pi_1 = 0.23404 \cdot 0.98904 \cdot 0.5 = 0.08939$$

$$P(\vec{x}_{\text{new}}|c_2)P(c_2) = P(y_1 = 1|c_2)P(\{y_2, y_3\} = \{0.3, 0.7\}|c_2)\pi_2 = 0.66732 \cdot 1.42292 \cdot 0.5 = 0.58286$$

And then, we can compute γ_{ki}:

$$\gamma_{1_{\text{new}}} = P(c_1|\vec{x}_{\text{new}}) = \frac{P(\vec{x}_{\text{new}}|c_1)P(c_1)}{P(\vec{x}_{\text{new}})} = \frac{0.08939}{0.08939 + 0.58286} = 0.13297$$

$$\gamma_{2_{\text{new}}} = P(c_2|\vec{x}_{\text{new}}) = \frac{P(\vec{x}_{\text{new}}|c_2)P(c_2)}{P(\vec{x}_{\text{new}})} = \frac{0.58286}{0.08939 + 0.58286} = 0.86703$$

Answer We can conclude that the observation $\vec{x}_{\text{new}}$ belongs to cluster c_2 because $\gamma_{2_{\text{new}}} > \gamma_{1_{\text{new}}}$.

3rd Question

Along the first and second questions, we have worked under a soft assignment approach. In this question, we will work under a hard assignment approach: we will assign each observation to the cluster with the highest probability instead of assigning each observation to each cluster with a certain probability. We first need to determine the cluster for each observation, using the parameters from table 2:

Cluster for $\vec{x}_1$

$$P(\vec{x}_1|c_1)P(c_1) = P(y_1 = 1|c_1)P(\{y_2, y_3\} = \{0.6, 0.1\}|c_1)\pi_1 = 0.23404 \cdot 0.98904 \cdot 0.5 = 0.08939$$

$$P(\vec{x}_1|c_2)P(c_2) = P(y_1 = 1|c_2)P(\{y_2, y_3\} = \{0.6, 0.1\}|c_2)\pi_2 = 0.66732 \cdot 1.42292 \cdot 0.5 = 0.58286$$

$$\gamma_{11} = P(c_1|\vec{x}_1) = \frac{P(\vec{x}_1|c_1)P(c_1)}{P(\vec{x}_1)} = \frac{0.08939}{0.08939 + 0.58286} = 0.13297$$

$$\gamma_{21} = P(c_2|\vec{x}_1) = \frac{P(\vec{x}_1|c_2)P(c_2)}{P(\vec{x}_1)} = \frac{0.58286}{0.08939 + 0.58286} = 0.86703$$

We can conclude that the observation $\vec{x}_1$ belongs to cluster c_2 because $\gamma_{21} > \gamma_{11}$.

Cluster for $\vec{x}_2$

$$P(\vec{x}_2|c_1)P(c_1) = P(y_1 = 0|c_1)P(\{y_2, y_3\} = \{-0.4, 0.8\}|c_1)\pi_1 = 0.76596 \cdot 1.65326 \cdot 0.5 = 0.48902$$

$$P(\vec{x}_2|c_2)P(c_2) = P(y_1 = 0|c_2)P(\{y_2, y_3\} = \{-0.4, 0.8\}|c_2)\pi_2 = 0.33268 \cdot 0.26673 \cdot 0.5 = 0.05447$$

$$\gamma_{12} = P(c_1|\vec{x}_2) = \frac{P(\vec{x}_2|c_1)P(c_1)}{P(\vec{x}_2)} = \frac{0.48902}{0.48902 + 0.05447} = 0.89978$$

$$\gamma_{22} = P(c_2|\vec{x}_2) = \frac{P(\vec{x}_2|c_2)P(c_2)}{P(\vec{x}_2)} = \frac{0.05447}{0.48902 + 0.05447} = 0.10022$$

We can conclude that the observation $\vec{x}_2$ belongs to cluster c_1 because $\gamma_{12} > \gamma_{22}$.

Cluster for $\vec{x}_3$

$$P(\vec{x}_3|c_1)P(c_1) = P(y_1 = 0|c_1)P(\{y_2, y_3\} = \{0.2, 0.5\}|c_1)\pi_1 = 0.76596 \cdot 1.87753 \cdot 0.5 = 0.55535$$

$$P(\vec{x}_3|c_2)P(c_2) = P(y_1 = 0|c_2)P(\{y_2, y_3\} = \{0.2, 0.5\}|c_2)\pi_2 = 0.33268 \cdot 1.36519 \cdot 0.5 = 0.27879$$

$$\gamma_{13} = P(c_1|\vec{x}_3) = \frac{P(\vec{x}_3|c_1)P(c_1)}{P(\vec{x}_3)} = \frac{0.55535}{0.55535 + 0.27879} = 0.66578$$

$$\gamma_{23} = P(c_2|\vec{x}_3) = \frac{P(\vec{x}_3|c_2)P(c_2)}{P(\vec{x}_3)} = \frac{0.27879}{0.55535 + 0.27879} = 0.33422$$

We can conclude that the observation $\vec{x}_3$ belongs to cluster c_1 because $\gamma_{13} > \gamma_{23}$.

Cluster for $\vec{x}_4$

$$P(\vec{x}_4|c_1)P(c_1) = P(y_1 = 1|c_1)P(\{y_2, y_3\} = \{0.4, -0.1\}|c_1)\pi_1 = 0.23404 \cdot 0.08873 \cdot 0.5 = 0.00802$$

$$P(\vec{x}_4|c_2)P(c_2) = P(y_1 = 1|c_2)P(\{y_2, y_3\} = \{0.4, -0.1\}|c_2)\pi_2 = 0.66732 \cdot 1.08391 \cdot 0.5 = 0.44399$$

$$\gamma_{14} = P(c_1|\vec{x}_4) = \frac{P(\vec{x}_4|c_1)P(c_1)}{P(\vec{x}_4)} = \frac{0.00802}{0.00802 + 0.44399} = 0.01774$$

$$\gamma_{24} = P(c_2|\vec{x}_4) = \frac{P(\vec{x}_4|c_2)P(c_2)}{P(\vec{x}_4)} = \frac{0.44399}{0.00802 + 0.44399} = 0.98226$$

We can conclude that the observation $\vec{x}_4$ belongs to cluster c_2 because $\gamma_{24} > \gamma_{14}$.

Clusters for each observation

Observation	$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	$\vec{x}_4$
Cluster	c_2	c_1	c_1	c_2

Table 3: Cluster for each observation

Silhouette Coefficient of the Clustering

The silhouette coefficient of the clustering is defined as the mean of the silhouette coefficients of each cluster. The silhouette coefficient of each cluster is defined as the mean of the silhouette coefficients of each observation in the cluster. The silhouette coefficient of each observation is defined as:

$$S(\text{clustering}) = \frac{1}{K}\sum_{i=1}^{K} S(c_i)$$

$$S(c_i) = \frac{1}{N_i}\sum_{j=1}^{N_i} S(\vec{x}_j)$$

$$S(\vec{x}_i) = \frac{b(\vec{x}_i) - a(\vec{x}_i)}{\max\{a(\vec{x}_i), b(\vec{x}_i)\}}$$

Where $a(\vec{x}_i)$ is the mean distance between $\vec{x}_i$ and the other observations in the same cluster and $b(\vec{x}_i)$ is the mean distance between $\vec{x}_i$ and the observations in the other clusters. N_i is the number of observations in the cluster c_i and K is the number of clusters.

Distances between Observations

In this exercise we are considering the Manhattan distance between observations:

$$d(\vec{x}_i, \vec{x}_j) = \sum_{k=1}^{3} |x_{ik} - x_{jk}|$$

We computed the distances between each observation in the following table:

Observation	$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	$\vec{x}_4$
$\vec{x}_1$	0	2.7	1.8	0.4
$\vec{x}_2$	2.7	0	0.9	2.7
$\vec{x}_3$	1.8	0.9	0	1.8
$\vec{x}_4$	0.4	2.7	1.8	0

Table 4: Distances between observations

Computing $\vec{a}$

We will now compute $\vec{a}$ for each observation. $\vec{a}$ is the mean distance between $\vec{x}_i$ and the other observations in the same cluster:

$$\vec{a} = \begin{bmatrix} a(\vec{x}_1) \\ a(\vec{x}_2) \\ a(\vec{x}_3) \\ a(\vec{x}_4) \end{bmatrix} = \begin{bmatrix} \frac{0.4}{1} \\ \frac{0.9}{1} \\ \frac{0.9}{1} \\ \frac{0.4}{1} \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.9 \\ 0.9 \\ 0.4 \end{bmatrix}$$

Computing $\vec{b}$

We will now compute $\vec{b}$ for each observation. $\vec{b}$ is the mean distance between $\vec{x}_i$ and the observations in the other clusters:

$$\vec{b} = \begin{bmatrix} b(\vec{x}_1) \\ b(\vec{x}_2) \\ b(\vec{x}_3) \\ b(\vec{x}_4 \end{bmatrix} = \begin{bmatrix} \frac{2.7+1.8}{2} \\ \frac{2.7+2.7}{2} \\ \frac{1.8+1.8}{2} \\ \frac{2.7+1.8}{2} \end{bmatrix} = \begin{bmatrix} 2.25 \\ 2.7 \\ 1.8 \\ 2.25 \end{bmatrix}$$

Silhouette Coefficient of each Observation

$$S(\vec{x}_1) = \frac{b(\vec{x}_1) - a(\vec{x}_1)}{\max\{a(\vec{x}_1), b(\vec{x}_1)\}} = \frac{2.25 - 0.4}{\max\{0.4, 2.25\}} = 0.82222$$
$$S(\vec{x}_2) = \frac{b(\vec{x}_2) - a(\vec{x}_2)}{\max\{a(\vec{x}_2), b(\vec{x}_2)\}} = \frac{2.7 - 0.9}{\max\{0.9, 2.7\}} = 0.66667$$
$$S(\vec{x}_3) = \frac{b(\vec{x}_3) - a(\vec{x}_3)}{\max\{a(\vec{x}_3), b(\vec{x}_3)\}} = \frac{1.8 - 0.9}{\max\{0.9, 1.8\}} = 0.5$$
$$S(\vec{x}_4) = \frac{b(\vec{x}_4) - a(\vec{x}_4)}{\max\{a(\vec{x}_4), b(\vec{x}_4)\}} = \frac{2.25 - 0.4}{\max\{0.4, 2.25\}} = 0.82222$$

Silhouette Coefficient of each Cluster

$$S(c_1) = \frac{1}{N_1}\sum_{j=1}^{N_1} S(\vec{x_j}) = \frac{1}{2}(S(\vec{x_2}) + S(\vec{x_3})) = \frac{1}{2}(0.66667 + 0.5) = 0.58333$$

$$S(c_2) = \frac{1}{N_2}\sum_{j=1}^{N_2} S(\vec{x_j}) = \frac{1}{2}(S(\vec{x_1}) + S(\vec{x_4})) = \frac{1}{2}(0.82222 + 0.82222) = 0.82222$$

Silhouette Coefficient of the Clustering

$$S(\text{clustering}) = \frac{1}{K}\sum_{i=1}^{K} S(c_i) = \frac{1}{2}(S(c_1) + S(c_2)) = \frac{1}{2}(0.58333 + 0.82222) = 0.70278$$

Answer The silhouette coefficient of the clustering is 0.70278.

4th Question

The purity of a clustering is defined as:

$$\text{purity} = \frac{1}{N}\sum_{k=1}^{K}\max_{j}(\#(c_k \cap l_j))$$

Where N is the number of observations, K is the number of clusters, c_k is the cluster k and l_j is the class j. $\#(c_k \cap l_j)$ is the number of observations in the cluster c_k and in the class l_j.

We are told that the purity of our clustering is 0.75, therefore:

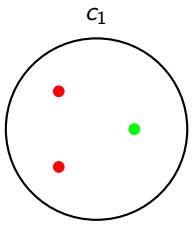
$$0.75 = \frac{1}{4}(\max_{j}(\#(c_1 \cap l_j)) + \max_{j}(\#(c_2 \cap l_j))) \Leftrightarrow$$
$$\Leftrightarrow 3 = \max_{j}(\#(c_1 \cap l_j)) + \max_{j}(\#(c_2 \cap l_j))$$

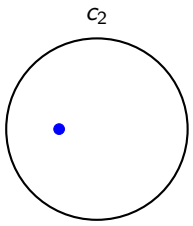
Therefore, we have two options, either $\max_j(\#(c_1 \cap l_j)) = 3$ and $\max_j(\#(c_2 \cap l_j)) = 0$ or $\max_j(\#(c_1 \cap l_j)) = 2$ and $\max_j(\#(c_2 \cap l_j)) = 1$.

1st Option

Considering the case where the maximum number of observations in a class in one cluster is 2 and the maximum number of observations in a class in the other cluster is 1, we have the following options:

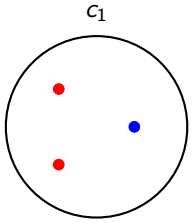
Option 1

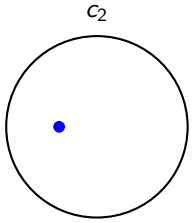




3 classes

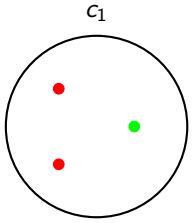
Option 2

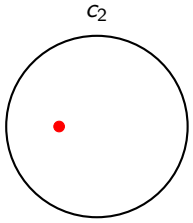




2 classes

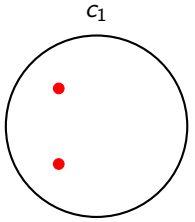
Option 3

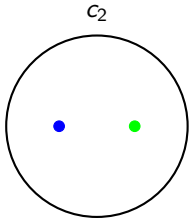




2 classes

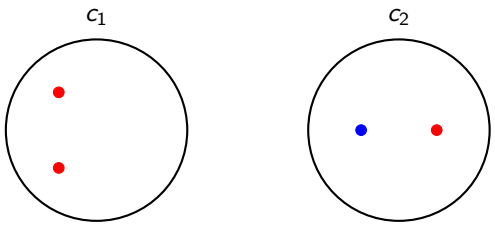
Option 4





3 classes

Option 5



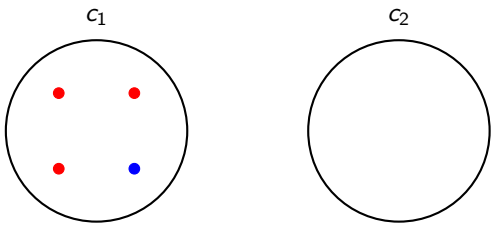
2 classes

In each of these diagrams, different colors represent different classes.

2nd Option

Considering the case where the maximum number of observations in a class in one cluster is 3 and the maximum number of observations in a class in the other cluster is 0, we have the following options:

Option 1



2 classes

Answer: We can conclude that a purity of 0.75 can be obtained if our data is classified in 2 or 3 classes.

Programming and Critical Analysis