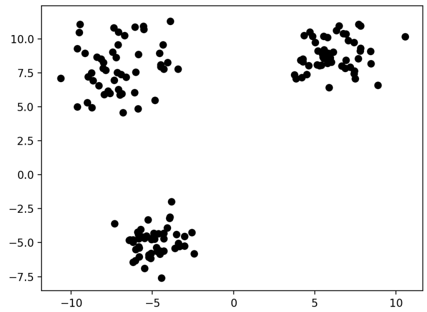
P09 Clustering

Luis Sa-Couto[1] and Andreas Wichert[2]

INESC-ID, Instituto Superior Tecnico, Universidade de Lisboa
{luis.sa.couto,andreas.wichert}@tecnico.ulisboa.pt

1 K-means clustering

The k-means clustering algorithm finds structure in a set of unlabeled data points. More specifically, we choose a number of clusters k and the algorithm tries to group the data points by proximity. Let us go through an example. The following figure presents the data set.



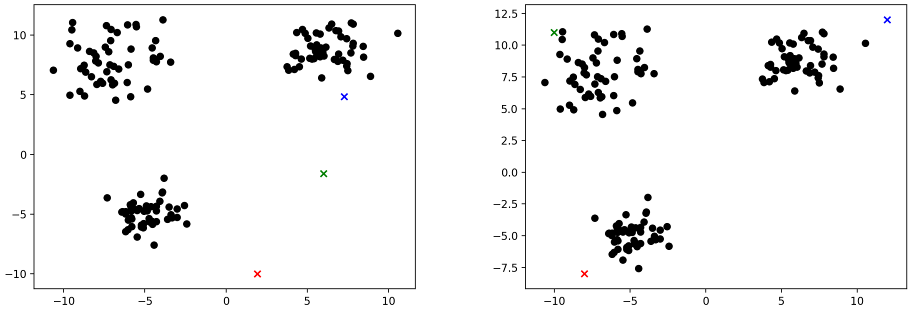
Analyzing it, we can see that there seem to be three distinct groups of points. For that reason, we will use $k = 3$. This is a key parameter. However, it is not always easy to make this choice. For one the problem can be a high dimensional one so we cannot plot the points and see. Furthermore, the points may be much closer so it is not obvious where a group starts and the other begins. Or even worse, there may be no groups at all.

After choosing the number of clusters, we can start the algorithm.

Step 0: Initialize the clusters

The algorithm starts by initializing each of the k clusters. There are many ways to do this each with its own advantages and disadvantages. A typical approach is to choose k random points from the data set or even to choose k random points from the whole space.

Initialization plays a key role in the procedure so we will procede our example with two different initializations.



For each initialization, we have our initial centroids of each cluster μ^1, μ^2 and μ^3. We distinguish them by using different colors: red, green and blue.

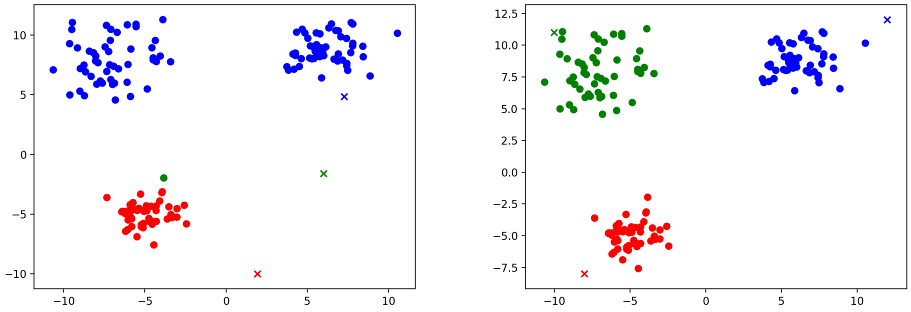
Having initialized the procedure, we can now start to iterate towards a solution. The iterative procedure is made by three steps.

Step 1: Assign points to clusters

K-means is deterministic in the sense that each point belongs to one and only one cluster. However, it is helpful to use probabilities to describe the assignment of a point to a cluster. More specifically, we say that the probability that point $\mathbf{x}^{(n)}$ belongs to cluster c, $p\left(C = c \mid \mathbf{x}^{(n)}\right) = 1$ if point $\mathbf{x}^{(n)}$ is closer to centroid μ^c than to any other and $p\left(C = c \mid \mathbf{x}^{(n)}\right) = 0$ otherwise. This logic can be written with the indicator function as follows.

$$p\left(C = c \mid \mathbf{x}^{(n)}\right) = \mathbb{I}\left[c = \arg \min_{l \in \{1,2,\ldots,k\}} \left\|\mathbf{x}^{(n)} - \mu^l\right\|\right]$$

Applying this logic to our ongoing examples, we get:



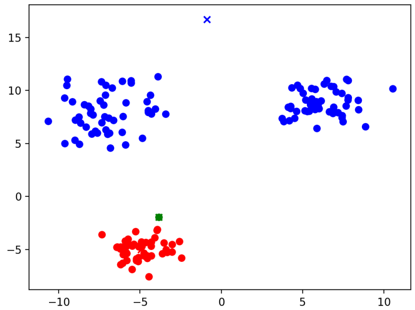
Having assigned all points to a cluster, we can move on to the next step.

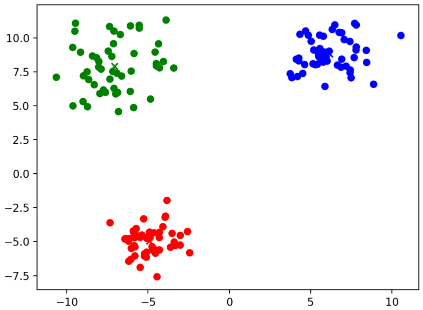
Step 2: Adjust cluster centers

Each cluster centroid is updated to the mean of the points that belong to it:

$$\mu^c = \frac{\sum_{n=1}^{N} p\left(C = c \mid \mathbf{x}^{(n)}\right) \mathbf{x}^{(n)}}{\sum_{n=1}^{N} p\left(C = c \mid \mathbf{x}^{(n)}\right)}$$

As we can see below, the centroids move closer to the points that were assigned to them.





At this point, we can see the impact of initialization. Whereas the example on the right as already identified the three groups correctly, the left one is still far from it. In fact, it may never even find them.

Step 3: Check convergence

If no centroid changed position the algorithm terminates. Otherwise, we go back to step 1.

1) Consider the following training data without labels:

$$\left\{ \mathbf{x}^{(1)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\}$$

Also, consider the following initialization centroids for $k = 2$ clusters $\mu^1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and $\mu^2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

a) Apply the k-means clustering algorithm until convergence.

Solution:
Let us start the first iteration.
Step 1: Assign points to clusters
Let us start with $\mathbf{x}^{(1)}$:

$$\left\|\mathbf{x}^{(1)}-\mu^1\right\|_2^2=\left\|\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}2\\0\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}-2\\0\end{pmatrix}\right\|_2^2=4$$

$$\left\|\mathbf{x}^{(1)}-\mu^2\right\|_2^2=\left\|\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}2\\1\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}-2\\-1\end{pmatrix}\right\|_2^2=5$$

$$\arg\min_{c\in\{1,2\}}\left\|\mathbf{x}^{(1)}-\mu^c\right\|_2^2=\arg\min_{c\in\{1,2\}}\{4,5\}=1$$

Now $\mathbf{x}^{(2)}$:

$$\left\|\mathbf{x}^{(2)}-\mu^1\right\|_2^2=\left\|\begin{pmatrix}1\\0\end{pmatrix}-\begin{pmatrix}2\\0\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}-1\\0\end{pmatrix}\right\|_2^2=1$$

$$\left\|\mathbf{x}^{(2)}-\mu^2\right\|_2^2=\left\|\begin{pmatrix}1\\0\end{pmatrix}-\begin{pmatrix}2\\1\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}-1\\-1\end{pmatrix}\right\|_2^2=2$$

$$\arg\min_{c\in\{1,2\}}\left\|\mathbf{x}^{(2)}-\mu^c\right\|_2^2=\arg\min_{c\in\{1,2\}}\{1,2\}=1$$

Now $\mathbf{x}^{(3)}$:

$$\left\|\mathbf{x}^{(3)}-\mu^1\right\|_2^2=\left\|\begin{pmatrix}0\\2\end{pmatrix}-\begin{pmatrix}2\\0\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}-2\\2\end{pmatrix}\right\|_2^2=8$$

$$\left\|\mathbf{x}^{(3)}-\mu^2\right\|_2^2=\left\|\begin{pmatrix}0\\2\end{pmatrix}-\begin{pmatrix}2\\1\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}-2\\1\end{pmatrix}\right\|_2^2=5$$

$$\arg\min_{c\in\{1,2\}}\left\|\mathbf{x}^{(3)}-\mu^c\right\|_2^2=\arg\min_{c\in\{1,2\}}\{8,5\}=2$$

Now $\mathbf{x}^{(4)}$:

$$\left\|\mathbf{x}^{(4)}-\mu^1\right\|_2^2=\left\|\begin{pmatrix}2\\2\end{pmatrix}-\begin{pmatrix}2\\0\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}0\\2\end{pmatrix}\right\|_2^2=4$$

$$\left\|\mathbf{x}^{(4)}-\mu^2\right\|_2^2=\left\|\begin{pmatrix}2\\2\end{pmatrix}-\begin{pmatrix}2\\1\end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix}0\\1\end{pmatrix}\right\|_2^2=1$$

$$\arg\min_{c\in\{1,2\}}\left\|\mathbf{x}^{(4)}-\mu^c\right\|_2^2=\arg\min_{c\in\{1,2\}}\{4,1\}=2$$

Step 2: Compute new clusters

Cluster 1has two elements: $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. The new centroid can be computing by averaging the two cluster elements.

$$\mu^1=\frac{\mathbf{x}^{(1)}+\mathbf{x}^{(2)}}{2}=\frac{1}{2}\left[\begin{pmatrix}0\\0\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}\right]=\frac{1}{2}\begin{pmatrix}1\\0\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}$$

Cluster 2 has two elements: $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$. The new centroid can be computing by averaging the two cluster elements.

$$\mu^2 = \frac{\mathbf{x}^{(3)} + \mathbf{x}^{(4)}}{2} = \frac{1}{2}\left[\begin{pmatrix}0\\2\end{pmatrix} + \begin{pmatrix}2\\2\end{pmatrix}\right] = \frac{1}{2}\begin{pmatrix}2\\4\end{pmatrix} = \begin{pmatrix}1\\2\end{pmatrix}$$

Step 3: Verify convergence
Change in cluster 2:

$$\|\mu^1_{old} - \mu^1\|_2^2 = \left\|\begin{pmatrix}2\\0\end{pmatrix} - \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}\frac{3}{2}\\0\end{pmatrix}\right\|_2^2 = \frac{9}{4}$$

$$\|\mu^2_{old} - \mu^2\|_2^2 = \left\|\begin{pmatrix}2\\1\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}1\\-1\end{pmatrix}\right\|_2^2 = 2$$

Centroids changed. We need to do another iteration.
Step 1: Assign points to clusters
Let us start with $\mathbf{x}^{(1)}$:

$$\left\|\mathbf{x}^{(1)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}-\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \frac{1}{4}$$

$$\left\|\mathbf{x}^{(1)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}-1\\-2\end{pmatrix}\right\|_2^2 = 5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(1)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \left\{\frac{1}{4}, 5\right\} = 1$$

Now $\mathbf{x}^{(2)}$:

$$\left\|\mathbf{x}^{(2)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}1\\0\end{pmatrix} - \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \frac{1}{4}$$

$$\left\|\mathbf{x}^{(2)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}1\\0\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}0\\-2\end{pmatrix}\right\|_2^2 = 4$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(2)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \left\{\frac{1}{4}, 4\right\} = 1$$

Now $\mathbf{x}^{(3)}$:

$$\left\|\mathbf{x}^{(3)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0\\2\end{pmatrix} - \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}-\frac{1}{2}\\2\end{pmatrix}\right\|_2^2 = \frac{9}{2}$$

$$\left\|\mathbf{x}^{(3)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0\\2\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}-1\\0\end{pmatrix}\right\|_2^2 = 1$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(3)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \left\{\frac{9}{2}, 1\right\} = 2$$

Now $\mathbf{x}^{(4)}$:

$$\left\|\mathbf{x}^{(4)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}\frac{3}{2}\\2\end{pmatrix}\right\|_2^2 = \frac{11}{2}$$

$$\left\|\mathbf{x}^{(4)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}1\\0\end{pmatrix}\right\|_2^2 = 1$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(4)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \left\{\frac{11}{2}, 1\right\} = 2$$

Step 2: Compute new clusters

Cluster 1has two elements: $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. The new centroid can be computing by averaging the two cluster elements.

$$\mu^1 = \frac{\mathbf{x}^{(1)} + \mathbf{x}^{(2)}}{2} = \frac{1}{2}\left[\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}1\\0\end{pmatrix}\right] = \frac{1}{2}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}$$

Cluster 2 has two elements: $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$. The new centroid can be computing by averaging the two cluster elements.

$$\mu^2 = \frac{\mathbf{x}^{(3)} + \mathbf{x}^{(4)}}{2} = \frac{1}{2}\left[\begin{pmatrix}0\\2\end{pmatrix} + \begin{pmatrix}2\\2\end{pmatrix}\right] = \frac{1}{2}\begin{pmatrix}2\\4\end{pmatrix} = \begin{pmatrix}1\\2\end{pmatrix}$$

Step 3: Verify convergence

Change in cluster 2:

$$\|\mu^1_{old} - \mu^1\|_2^2 = \left\|\begin{pmatrix}\frac{1}{2}\\0\end{pmatrix} - \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}0\\0\end{pmatrix}\right\|_2^2 = 0$$

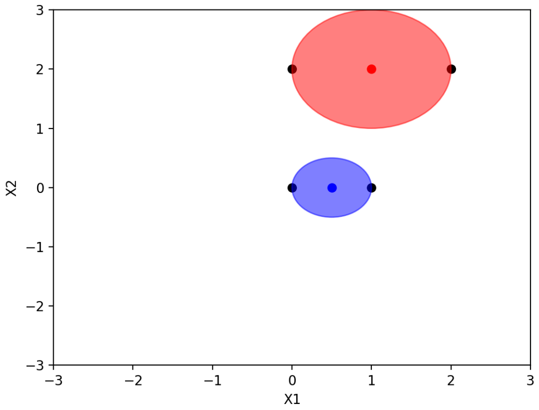
$$\|\mu^2_{old} - \mu^2\|_2^2 = \left\|\begin{pmatrix}1\\2\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}0\\0\end{pmatrix}\right\|_2^2 = 0$$

No centroid change, so the alogrithm converged.

---

b) Plot the data points and draw the clusters.

Solution:



---

2) Consider the following training data without labels:

$$\left\{ \mathbf{x}^{(1)} = \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \end{pmatrix}, \mathbf{x}^{(5)} = \begin{pmatrix} 0.0 \\ 1.0 \\ 0.0 \end{pmatrix}, \mathbf{x}^{(6)} = \begin{pmatrix} 3.0 \\ 2.0 \\ 1.0 \end{pmatrix} \right\}$$

Everytime you need to initialize k clusters, do it by taking the first k points of the dataset and using them as centroids.

a) For $k = 2$ perform k-means clustering until convergence.

Solution:

Let us start the first iteration.

$$\mu^1 = \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix}$$

$$\mu^2 = \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}$$

Step 1: Assign points to clusters.

Since there is only one cluster all points are assigned to it.

$$\left\| \mathbf{x}^{(1)} - \mu^1 \right\|_2^2 = \left\| \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} \right\|_2^2 = 0$$

$$\left\|\mathbf{x}^{(1)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 129$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(1)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{0, 129\} = 1$$

$$\left\|\mathbf{x}^{(2)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 129$$

$$\left\|\mathbf{x}^{(2)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 0$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(2)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{129, 0\} = 2$$

$$\left\|\mathbf{x}^{(3)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 13$$

$$\left\|\mathbf{x}^{(3)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 66$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(3)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{13, 66\} = 1$$

$$\left\|\mathbf{x}^{(4)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 2$$

$$\left\|\mathbf{x}^{(4)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 137$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(4)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{2, 137\} = 1$$

$$\left\|\mathbf{x}^{(5)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 2$$

$$\left\|\mathbf{x}^{(5)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 129$$

$$\arg \min_{c \in \{1,2\}} \left\| \mathbf{x}^{(5)} - \mu^c \right\|_2^2 = \arg \min_{c \in \{1,2\}} \{2, 129\} = 1$$

$$\left\| \mathbf{x}^{(6)} - \mu^1 \right\|_2^2 = \left\| \begin{pmatrix} 3.0 \\ 2.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} \right\|_2^2 = 9$$

$$\left\| \mathbf{x}^{(6)} - \mu^2 \right\|_2^2 = \left\| \begin{pmatrix} 3.0 \\ 2.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right\|_2^2 = 70$$

$$\arg \min_{c \in \{1,2\}} \left\| \mathbf{x}^{(6)} - \mu^c \right\|_2^2 = \arg \min_{c \in \{1,2\}} \{9, 70\} = 1$$

Step 2: Compute new centroids
The new centroids can be computing by averaging the cluster elements.

$$\mu^1 = \frac{\sum_{n=1}^{6} p\left(C = 1 \mid \mathbf{x}^{(n)}\right) \mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = 1 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{5} \left[\begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} + \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} + \begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \end{pmatrix} + \begin{pmatrix} 0.0 \\ 1.0 \\ 0.0 \end{pmatrix} + \begin{pmatrix} 3.0 \\ 2.0 \\ 1.0 \end{pmatrix} \right] = \frac{1}{5} \begin{pmatrix} 7 \\ 6 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix}$$

$$\mu^2 = \frac{\sum_{n=1}^{6} p\left(C = 2 \mid \mathbf{x}^{(n)}\right) \mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = 2 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{1} \left[\begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right] = \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}$$

Step 3: Verify convergence
Change in cluster 2:

$$\left\| \mu^1_{old} - \mu^1 \right\|_2^2 = \left\| \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 = 1.33$$

$$\left\| \mu^2_{old} - \mu^2 \right\|_2^2 = \left\| \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right\|_2^2 = 0.0$$

Centroids changed. We need to do another iteration.
Step 1: Assign points to clusters.
Since there is only one cluster all points are assigned to it.

$$\left\| \mathbf{x}^{(1)} - \mu^1 \right\|_2^2 = \left\| \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 = 1.76$$

$$\left\| \mathbf{x}^{(1)} - \mu^2 \right\|_2^2 = \left\| \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right\|_2^2 = 129$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(1)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{1.76, 129\} = 1$$

$$\left\|\mathbf{x}^{(2)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix}\right\|_2^2 = 102.76$$

$$\left\|\mathbf{x}^{(2)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 0$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(2)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{102.76, 0\} = 2$$

$$\left\|\mathbf{x}^{(3)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix}\right\|_2^2 = 5.96$$

$$\left\|\mathbf{x}^{(3)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 66$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(3)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{5.96, 66\} = 1$$

$$\left\|\mathbf{x}^{(4)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix}\right\|_2^2 = 3.76$$

$$\left\|\mathbf{x}^{(4)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 137$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(4)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{3.76, 137\} = 1$$

$$\left\|\mathbf{x}^{(5)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 0.0 \\ 1.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix}\right\|_2^2 = 2.16$$

$$\left\|\mathbf{x}^{(5)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 0.0 \\ 1.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 129$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(5)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{2.16, 129\} = 1$$

$$\left\|\mathbf{x}^{(6)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}\frac{7}{5}\\\frac{6}{5}\\\frac{2}{5}\end{pmatrix}\right\|_2^2 = 3.56$$

$$\left\|\mathbf{x}^{(6)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 70$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(6)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{3.56, 70\} = 1$$

Step 2: Compute new centroids
The new centroids can be computing by averaging the cluster elements.

$$\mu^1 = \frac{\sum_{n=1}^{6} p\left(C = c \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = c \mid \mathbf{x}^{(n)}\right)} = \frac{1}{5}\left[\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} + \begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} + \begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} + \begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} + \begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix}\right] = \frac{1}{5}\begin{pmatrix}7\\6\\2\end{pmatrix} = \begin{pmatrix}\frac{7}{5}\\\frac{6}{5}\\\frac{2}{5}\end{pmatrix}$$

$$\mu^2 = \frac{\sum_{n=1}^{6} p\left(C = 2 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = 2 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{1}\left[\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right] = \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}$$

Step 3: Verify convergence
Change in cluster 2:

$$\left\|\mu^1_{old} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}\frac{7}{5}\\\frac{6}{5}\\\frac{2}{5}\end{pmatrix} - \begin{pmatrix}\frac{7}{5}\\\frac{6}{5}\\\frac{2}{5}\end{pmatrix}\right\|_2^2 = 0.0$$

$$\left\|\mu^2_{old} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 0.0$$

No centroid change, so the alogrithm converged.

---

b) For $k = 3$ perform k-means clustering until convergence.

Solution:
Let us start the first iteration.

$$\mu^1 = \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}$$

$$\mu^2 = \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}$$

$$\mu^3 = \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix}$$

Step 1: Assign points to clusters.
Since there is only one cluster all points are assigned to it.

$$\left\|\mathbf{x}^{(1)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix}\right\|_2^2 = 0$$

$$\left\|\mathbf{x}^{(1)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 129$$

$$\left\|\mathbf{x}^{(1)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix}\right\|_2^2 = 13$$

$$\arg\min_{c \in \{1,2\}} \left\|\mathbf{x}^{(1)} - \mu^c\right\|_2^2 = \arg\min_{c \in \{1,2\}} \{0, 129, 13\} = 1$$

$$\left\|\mathbf{x}^{(2)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix}\right\|_2^2 = 129$$

$$\left\|\mathbf{x}^{(2)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 0$$

$$\left\|\mathbf{x}^{(2)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix}\right\|_2^2 = 66$$

$$\arg\min_{c \in \{1,2\}} \left\|\mathbf{x}^{(2)} - \mu^c\right\|_2^2 = \arg\min_{c \in \{1,2\}} \{129, 0, 66\} = 2$$

$$\left\|\mathbf{x}^{(3)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix}\right\|_2^2 = 13$$

$$\left\|\mathbf{x}^{(3)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix}\right\|_2^2 = 66$$

$$\left\|\mathbf{x}^{(3)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix}\right\|_2^2 = 0$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(3)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{13, 66, 0\} = 3$$

$$\left\|\mathbf{x}^{(4)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 2$$

$$\left\|\mathbf{x}^{(4)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 137$$

$$\left\|\mathbf{x}^{(4)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix}\right\|_2^2 = 19$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(4)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{2, 137, 19\} = 1$$

$$\left\|\mathbf{x}^{(5)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 2$$

$$\left\|\mathbf{x}^{(5)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 129$$

$$\left\|\mathbf{x}^{(5)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix}\right\|_2^2 = 13$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(5)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{2, 129, 13\} = 1$$

$$\left\|\mathbf{x}^{(6)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix}\right\|_2^2 = 9$$

$$\left\|\mathbf{x}^{(6)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 70$$

$$\left\|\mathbf{x}^{(6)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix}\right\|_2^2 = 2$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(6)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{9, 70, 2\} = 3$$

Step 2: Compute new centroids
The new centroids can be computing by averaging the cluster elements.

$$\mu^1 = \frac{\sum_{n=1}^6 p\left(C=1 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^6 p\left(C=1 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{3}\left[\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} + \begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} + \begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix}\right] = \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}$$

$$\mu^2 = \frac{\sum_{n=1}^6 p\left(C=2 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^6 p\left(C=2 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{1}\left[\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right] = \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}$$

$$\mu^3 = \frac{\sum_{n=1}^6 p\left(C=3 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^6 p\left(C=3 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{2}\left[\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} + \begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix}\right] = \frac{1}{2}\begin{pmatrix}6\\5\\1\end{pmatrix} = \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}$$

Step 3: Verify convergence
Change in cluster 2:

$$\left\|\mu^1_{old} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}1\\0\\0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 0.67$$

$$\left\|\mu^2_{old} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 0.0$$

$$\left\|\mu^3_{old} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 0.5$$

Centroids changed. We need to do another iteration.
Step 1: Assign points to clusters.
Since there is only one cluster all points are assigned to it.

$$\left\|\mathbf{x}^{(1)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 0.67$$

$$\left\|\mathbf{x}^{(1)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 129$$

$$\left\|\mathbf{x}^{(1)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 10.5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(1)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{0.67, 129, 10.5\} = 1$$

$$\left\|\mathbf{x}^{(2)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 131$$

$$\left\|\mathbf{x}^{(2)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 0$$

$$\left\|\mathbf{x}^{(2)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 67.5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(2)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{131, 0, 67.5\} = 2$$

$$\left\|\mathbf{x}^{(3)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 14.33$$

$$\left\|\mathbf{x}^{(3)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 66$$

$$\left\|\mathbf{x}^{(3)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 0.5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(3)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{14.33, 66, 0.5\} = 3$$

$$\left\|\mathbf{x}^{(4)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 0.67$$

$$\left\|\mathbf{x}^{(4)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 137$$

$$\left\|\mathbf{x}^{(4)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 15.5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(4)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{0.67, 137, 15.5\} = 1$$

$$\left\|\mathbf{x}^{(5)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 0.67$$

$$\left\|\mathbf{x}^{(5)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 129$$

$$\left\|\mathbf{x}^{(5)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 11.5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(5)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{0.67, 129, 11.5\} = 1$$

$$\left\|\mathbf{x}^{(6)} - \mu^1\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}\right\|_2^2 = 10.33$$

$$\left\|\mathbf{x}^{(6)} - \mu^2\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right\|_2^2 = 70$$

$$\left\|\mathbf{x}^{(6)} - \mu^3\right\|_2^2 = \left\|\begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix} - \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}\right\|_2^2 = 0.5$$

$$\arg\min_{c\in\{1,2\}} \left\|\mathbf{x}^{(6)} - \mu^c\right\|_2^2 = \arg\min_{c\in\{1,2\}} \{10.33, 70, 0.5\} = 3$$

Step 2: Compute new centroids
The new centroids can be computing by averaging the cluster elements.

$$\mu^1 = \frac{\sum_{n=1}^{6} p\left(C = 1 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = 1 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{3}\left[\begin{pmatrix}1.0\\0.0\\0.0\end{pmatrix} + \begin{pmatrix}0.0\\0.0\\1.0\end{pmatrix} + \begin{pmatrix}0.0\\1.0\\0.0\end{pmatrix}\right] = \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}\frac{1}{3}\\\frac{1}{3}\\\frac{1}{3}\end{pmatrix}$$

$$\mu^2 = \frac{\sum_{n=1}^{6} p\left(C = 2 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = 2 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{1}\left[\begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}\right] = \begin{pmatrix}8.0\\8.0\\4.0\end{pmatrix}$$

$$\mu^3 = \frac{\sum_{n=1}^{6} p\left(C = 3 \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^{6} p\left(C = 3 \mid \mathbf{x}^{(n)}\right)} = \frac{1}{2}\left[\begin{pmatrix}3.0\\3.0\\0.0\end{pmatrix} + \begin{pmatrix}3.0\\2.0\\1.0\end{pmatrix}\right] = \frac{1}{2}\begin{pmatrix}6\\5\\1\end{pmatrix} = \begin{pmatrix}3\\\frac{5}{2}\\\frac{1}{2}\end{pmatrix}$$

Step 3: Verify convergence
Change in cluster 2:

$$\left\|\mu_{old}^1 - \mu^1\right\|_2^2 = \left\| \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|_2^2 = 0.0$$

$$\left\|\mu_{old}^2 - \mu^2\right\|_2^2 = \left\| \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right\|_2^2 = 0.0$$

$$\left\|\mu_{old}^3 - \mu^3\right\|_2^2 = \left\| \begin{pmatrix} 3 \\ \frac{5}{2} \\ \frac{1}{2} \end{pmatrix} - \begin{pmatrix} 3 \\ \frac{5}{2} \\ \frac{1}{2} \end{pmatrix} \right\|_2^2 = 0.0$$

Centroids changed. We need to do another iteration.
No centroid change, so the alogrithm converged.

---

c) Which k provides a better clustering in terms of sum of intra-cluster euclidean distances.

Solution:
For $k = 2$:

$$\begin{aligned} D_{intra-cluster} &= \sum_{n=1}^{6} \sum_{c=1}^{2} p\left(C = c \mid \mathbf{x}^{(n)}\right) \left\|\mathbf{x}^{(n)} - \mu^c\right\|_2^2 \\ &= \left\| \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right\|_2^2 + \\ &+ \left\| \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 3 \\ \frac{5}{2} \\ \frac{1}{2} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 + \\ &+ \left\| \begin{pmatrix} 0.0 \\ 1.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} 3.0 \\ 2.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} 3 \\ \frac{5}{2} \\ \frac{1}{2} \end{pmatrix} \right\|_2^2 = \\ &= 1.76 + 0 + 5.96 + 3.76 + 2.16 + 3.56 = \\ &= 17.2 \end{aligned}$$

For $k = 3$:

$$\begin{aligned}
E &= \sum_{n=1}^{6} \sum_{c=1}^{2} p\left(C = c \mid \mathbf{x}^{(n)}\right) \left\| \mathbf{x}^{(n)} - \mu^{c} \right\|_2^2 \\
&= \left\| \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 8.0 \\ 8.0 \\ 4.0 \end{pmatrix} \right\|_2^2 + \\
&+ \left\| \begin{pmatrix} 3.0 \\ 3.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|_2^2 + \\
&+ \left\| \begin{pmatrix} 0.0 \\ 1.0 \\ 0.0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} 3.0 \\ 2.0 \\ 1.0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|_2^2 = \\
&= 0.67 + 0 + 0.5 + 0.67 + 0.67 + 0.5 = \\
&= 3.0
\end{aligned}$$

So, $k = 3$ has more tightly packed clusters which is in general better.

---

d) Which k provides a better clustering in terms of mean inter-cluster centroid distance.

Solution:
The mean distance between centroids can be computed as follows.

$$D_{inter-cluster} = \sum_{i=1}^{k} \sum_{j=1}^{k} \left\| \mu^{i} - \mu^{j} \right\|_2^2$$

So, for $k = 2$:

$$\begin{aligned}
D_{inter-cluster} &= \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \left\| \mu^{i} - \mu^{j} \right\|_2^2 \\
&= \frac{1}{4} \left(0 + 102.76 + 0 + 102.76\right) \\
&= 51.39
\end{aligned}$$

For $k = 3$:

$$
\begin{aligned}
D_{inter-cluster} &= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\mu^i - \mu^j\right\|_2^2 \\
&= \frac{1}{9}\left(0 + 131 + 11.83 + 131 + 0 + 67.5 + 11.83 + 67.5 + 0\right) \\
&= 46.67
\end{aligned}
$$

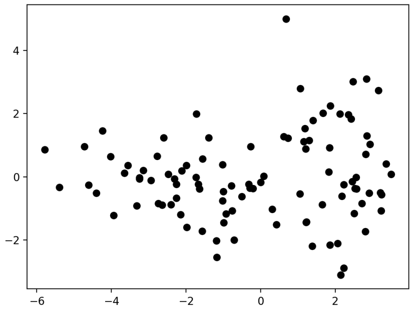
Looks like $k = 3$ has better separated clusters which is in general better.

2 Expectation-Maximization Clustering

Like k-means, EM-clustering finds structure in a set of unlabeled data points. However, it does so in a probabilistic manner. As in k-means the goal is to assign a probability $\mathrm{p}\big(C = c \mid \mathbf{x}^{(n)}\big)$ that a point belongs to a cluster. Yet, unlike k-means, in EM we allow for non-deterministic distributions. This means that a point does not necessarly belong to one cluster. It is a member of every cluster but with different degrees of probability.

To estimate this posterior probability, we use Bayes rule to decompose it into a prior $p\left(C = c\right)$ and likelihood $p\left(\mathbf{x}^{(n)} \mid C = c\right)$. So, we will use this two probabilities for each cluster as we go along.

To make things more specific, let us go through an example. The following figure presents the data set.

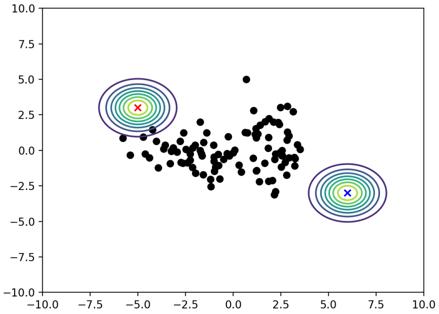


Analyzing it, we can see that there seem to be two distinct groups of points. For that reason, we will use $k = 2$ clusters. This is a key parameter. However, just like in k-means, it is not always easy to make this choice.

After choosing the number of clusters, we can start the algorithm.

Step 0: Initialize the clusters

The algorithm starts by initializing the likelihood and prior of each of the k clusters. In our example we will use two-dimensional Gaussians as likelihoods. However, the procedure can be generalized to other distributions. So, we start with uniform priors for each cluster ($p\left(C=1\right)=p\left(C=2\right)=0.5$) and, for the likelihoods, we randomly choose the means and use identity matrix covariances. The following figure plots the likelihoods.



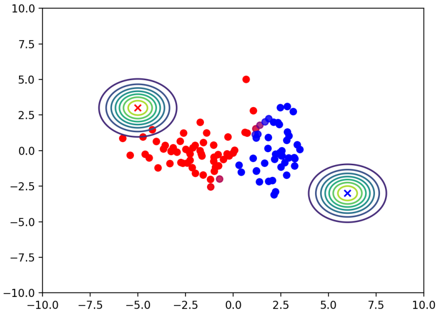
Having initialized the procedure, we can now start to iterate towards a solution. The iterative procedure is made by three steps.

Expectation Step: Assign points to clusters

With Bayes rule, we can write the posterior probability of each point belonging to each cluster as follows.

$$\mathrm{p}\Big(C=c \mid \mathbf{x}^{(n)}\Big) = \frac{\mathrm{p}\big(\mathbf{x}^{(n)} \mid C=c;\theta^{c}\big)\mathrm{p}(C=c)}{\mathrm{p}\big(\mathbf{x}^{(n)}\big)}$$

Applying this logic to our ongoing example, we get the figure below. Note that the strength of the color represents the probability that a point belongs to that cluster.



Having assigned probabilities to all points for all cluster, we can move on to the next step.

Maximization Step: Adjust cluster parameters

The posterior probability of a cluster for a given point represents how much that point contributes to estimate the distributions of that cluster. So, for instance, since we would estimate the prior of a cluster by counting the number of points in that cluster and dividing by the total number of points, we, instead sum and normalize the contributions of each point to that cluster as follows.

$$p\left(C=c\right)=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}{\sum_{l=1}^{k}\sum_{n=1}^{N}p\left(C=l\mid\mathbf{x}^{(n)}\right)}$$

Specifically, if, for a given point, $p\left(C=1\mid\mathbf{x}^{(n)}\right)=1$ then this point counts totally for cluster one and nothing to all other clusters. The same reasoning can be applied, in general, for the likelihoods. So, if the likelihood is defined by a parameter θ^{c}, then we estimated by averaging the contribution of each point to this parameter, weighted by the posterior probability that each point belongs to the cluster:

$$\theta^{c}=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)Contribution\left(\theta,\mathbf{x}^{(n)}\right)}{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}$$

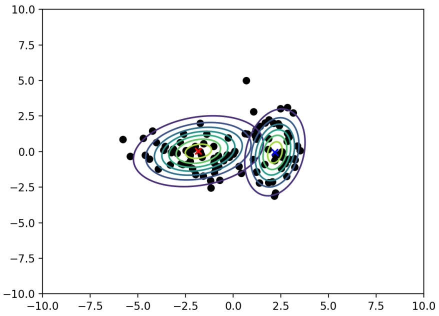
For example, for a multivariate gaussian, we have two parameters. For the mean, the contribution of each point is itself $x_i^{(n)}$. So, it can be estimated as:

$$\mu_i^c=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)Contribution\left(\mu_i,\mathbf{x}^{(n)}\right)}{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)\left[x_i^{(n)}\right]}{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}$$

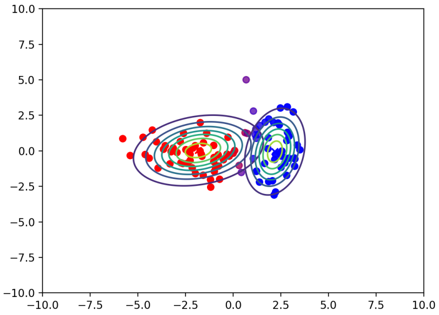
For position ij of the covariance matrix, the contribution of each point is the deviation from the dimensional means of its dimensions $\left(\mu_i^c-x_i^{(n)}\right)\left(\mu_j^c-x_j^{(n)}\right)$. So, it can be estimated as follows:

$$\Sigma_{ij}^c=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)Contribution\left(\Sigma_{ij},\mathbf{x}^{(n)}\right)}{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)\left[\left(\mu_i^c-x_i^{(n)}\right)\left(\mu_j^c-x_j^{(n)}\right)\right]}{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}$$

As we can see below, the clusters are now more suited to the points that were assigned to them.



This terminates the M-step. Afterwards, the algorithm checks for convergence by seeing if any parameters changed. If they did not, then it terminates. If they did, then it goes back to the E-step. For curiosity, the next E-step would yield the follwoing probabilities.



1) Consider the following training data with boolean features:

$$\left\{ \mathbf{x}^{(1)} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{x}^{(5)} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \right\}$$

We want to model the data with three clusters. Initialize all priors uniformly and initialize using the following table:

	$p(x_1 = 1 \mid C = c)$	$p(x_2 = 1 \mid C = c)$	$p(x_3 = 1 \mid C = c)$	$p(x_4 = 1 \mid C = c)$
$c = 1$	0.8	0.5	0.1	0.1
$c = 2$	0.1	0.5	0.4	0.8
$c = 3$	0.1	0.1	0.9	0.2

Assume all features are conditionally independent given the cluster.

a) Perform one expectation maximization iteration.

Solution:

The question tells us that all features are conditionally independent given the cluster. So, we can write the likelihoods as follows:

$$p\left(\mathbf{x} \mid C=1\right) = p\left(x_1 \mid C=1\right) p\left(x_2 \mid C=1\right) p\left(x_3 \mid C=1\right) p\left(x_4 \mid C=1\right)$$

Furthermore, the questions tells us that all distributions are initialized uniformly. So, we will have priors:

$$p\left(C=1\right) = \frac{1}{3}$$

$$p\left(C=2\right) = \frac{1}{3}$$

$$p\left(C=3\right) = \frac{1}{3}$$

And for the likelihoods:

	$p\left(x_1=1 \mid C=c\right)$	$p\left(x_2=1 \mid C=c\right)$	$p\left(x_3=1 \mid C=c\right)$	$p\left(x_4=1 \mid C=c\right)$
$c=1$	0.8	0.5	0.1	0.1
$c=2$	0.1	0.5	0.4	0.8
$c=3$	0.1	0.1	0.9	0.2

Which means:

	$p\left(x_1=0 \mid C=c\right)$	$p\left(x_2=0 \mid C=c\right)$	$p\left(x_3=0 \mid C=c\right)$	$p\left(x_4=0 \mid C=c\right)$
$c=1$	0.2	0.5	0.9	0.9
$c=2$	0.9	0.5	0.6	0.2
$c=3$	0.9	0.9	0.1	0.8

Each iteration has two steps. Let us do one:

E-Step: Assign each point to the cluster that yields higher posterior

- For $\mathbf{x}^{(1)}$:
  - For cluster $C=1$:
    - Prior: $p\left(C=1\right) = \frac{1}{3}$
    - Likelihood:
      $p\left(\mathbf{x}^{(1)} \mid C=1\right) =$
      $= p\left(x_1^{(1)}=1 \mid C=1\right) p\left(x_2^{(1)}=0 \mid C=1\right) p\left(x_3^{(1)}=0 \mid C=1\right) p\left(x_4^{(1)}=0 \mid C=1\right) =$
      $0.8 \times 0.5 \times 0.9 \times 0.9 = 0.324$

        - Joint Probability: $p\left(C=1 \mid \mathbf{x}^{(1)}\right)=p\left(C=1\right)p\left(\mathbf{x}^{(1)} \mid C=1\right)=\frac{1}{3}\times 0.324=0.108$
    - For cluster $C=2$:
        - Prior: $p\left(C=2\right)=\frac{1}{3}$
        - Likelihood:
          $p\left(\mathbf{x}^{(1)} \mid C=2\right)=$
          $=p\left(x_1^{(1)}=1 \mid C=2\right)p\left(x_2^{(1)}=0 \mid C=2\right)p\left(x_3^{(1)}=0 \mid C=2\right)p\left(x_4^{(1)}=0 \mid C=2\right)=$
          $0.1\times 0.5\times 0.6\times 0.2=0.006$
        - Joint Probability: $p\left(C=2,\mathbf{x}^{(1)}\right)=p\left(C=2\right)p\left(\mathbf{x}^{(1)} \mid C=2\right)=\frac{1}{3}\times 0.006=0.002$
    - For cluster $C=3$:
        - Prior: $p\left(C=3\right)=\frac{1}{3}$
        - Likelihood:
          $p\left(\mathbf{x}^{(1)} \mid C=3\right)=$
          $=p\left(x_1^{(1)}=1 \mid C=3\right)p\left(x_2^{(1)}=0 \mid C=3\right)p\left(x_3^{(1)}=0 \mid C=3\right)p\left(x_4^{(1)}=0 \mid C=3\right)=$
          $0.1\times 0.9\times 0.1\times 0.8=0.0072$
        - Joint Probability: $p\left(C=3,\mathbf{x}^{(1)}\right)=p\left(C=3\right)p\left(\mathbf{x}^{(1)} \mid C=3\right)=\frac{1}{3}\times 0.0072=0.0024$
    - So, we can compute the normalized posteriors for each cluster:
        - $C=1$: $p\left(C=1 \mid \mathbf{x}^{(1)}\right)=\frac{p\left(C=1,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)+p\left(C=3,\mathbf{x}^{(1)}\right)}=\frac{0.108}{0.108+0.002+0.0024}=0.961$
        - $C=2$: $p\left(C=2 \mid \mathbf{x}^{(1)}\right)=\frac{p\left(C=2,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)+p\left(C=3,\mathbf{x}^{(1)}\right)}=\frac{0.002}{0.108+0.002+0.0024}=0.018$
        - $C=3$: $p\left(C=3 \mid \mathbf{x}^{(1)}\right)=\frac{p\left(C=3,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)+p\left(C=3,\mathbf{x}^{(1)}\right)}=\frac{0.0024}{0.108+0.002+0.0024}=0.021$
- For $\mathbf{x}^{(2)}$:
    - For cluster $C=1$:
        - Prior: $p\left(C=1\right)=\frac{1}{3}$
        - Likelihood:
          $p\left(\mathbf{x}^{(2)} \mid C=1\right)=$
          $=p\left(x_1^{(2)}=0 \mid C=1\right)p\left(x_2^{(2)}=1 \mid C=1\right)p\left(x_3^{(2)}=1 \mid C=1\right)p\left(x_4^{(2)}=1 \mid C=1\right)=$
          $0.2\times 0.5\times 0.1\times 0.1=0.001$
        - Joint Probability: $p\left(C=1 \mid \mathbf{x}^{(1)}\right)=p\left(C=1\right)p\left(\mathbf{x}^{(1)} \mid C=1\right)=\frac{1}{3}\times 0.001=0.0003$
    - For cluster $C=2$:
        - Prior: $p\left(C=2\right)=\frac{1}{3}$
        - Likelihood:
          $p\left(\mathbf{x}^{(2)} \mid C=2\right)=$
          $=p\left(x_1^{(2)}=0 \mid C=2\right)p\left(x_2^{(2)}=1 \mid C=2\right)p\left(x_3^{(2)}=1 \mid C=2\right)p\left(x_4^{(2)}=1 \mid C=2\right)=$
          $0.9\times 0.5\times 0.4\times 0.8=0.144$
        - Joint Probability: $p\left(C=2,\mathbf{x}^{(1)}\right)=p\left(C=2\right)p\left(\mathbf{x}^{(1)} \mid C=2\right)=\frac{1}{3}\times 0.144=0.048$

  - For cluster $C = 3$:
    - Prior: $p\left(C = 3\right) = \frac{1}{3}$
    - Likelihood:
      $p\left(\mathbf{x}^{(2)} \mid C = 3\right) =$
      $= p\left(x_1^{(2)} = 0 \mid C = 3\right) p\left(x_2^{(2)} = 1 \mid C = 3\right) p\left(x_3^{(2)} = 1 \mid C = 3\right) p\left(x_4^{(2)} = 1 \mid C = 3\right) =$
      $0.9 \times 0.1 \times 0.9 \times 0.2 = 0.0162$
    - Joint Probability: $p\left(C = 3, \mathbf{x}^{(1)}\right) = p\left(C = 3\right) p\left(\mathbf{x}^{(1)} \mid C = 3\right) = \frac{1}{3} \times 0.0162 = 0.0054$
  - So, we can compute the normalized posteriors for each cluster:
    - $C = 1$: $p\left(C = 1 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=1,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)+p\left(C=3,\mathbf{x}^{(2)}\right)} = \frac{0.0003}{0.0003+0.048+0.0054} = 0.006$
    - $C = 2$: $p\left(C = 2 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=2,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)+p\left(C=3,\mathbf{x}^{(2)}\right)} = \frac{0.048}{0.0003+0.048+0.0054} = 0.894$
    - $C = 3$: $p\left(C = 3 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=3,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)+p\left(C=3,\mathbf{x}^{(2)}\right)} = \frac{0.0054}{0.0003+0.048+0.0054} = 0.100$
- For $\mathbf{x}^{(3)}$:
  - For cluster $C = 1$:
    - Prior: $p\left(C = 1\right) = \frac{1}{3}$
    - Likelihood:
      $p\left(\mathbf{x}^{(3)} \mid C = 1\right) =$
      $= p\left(x_1^{(3)} = 0 \mid C = 1\right) p\left(x_2^{(3)} = 1 \mid C = 1\right) p\left(x_3^{(3)} = 0 \mid C = 1\right) p\left(x_4^{(3)} = 1 \mid C = 1\right) =$
      $0.2 \times 0.5 \times 0.9 \times 0.1 = 0.009$
    - Joint Probability: $p\left(C = 1, \mathbf{x}^{(3)}\right) = p\left(C = 1\right) p\left(\mathbf{x}^{(3)} \mid C = 1\right) = \frac{1}{3} \times 0.009 = 0.003$
  - For cluster $C = 2$:
    - Prior: $p\left(C = 2\right) = \frac{1}{3}$
    - Likelihood:
      $p\left(\mathbf{x}^{(3)} \mid C = 2\right) =$
      $= p\left(x_1^{(3)} = 0 \mid C = 2\right) p\left(x_2^{(3)} = 1 \mid C = 2\right) p\left(x_3^{(3)} = 0 \mid C = 2\right) p\left(x_4^{(3)} = 1 \mid C = 2\right) =$
      $0.9 \times 0.5 \times 0.6 \times 0.8 = 0.216$
    - Joint Probability: $p\left(C = 2, \mathbf{x}^{(3)}\right) = p\left(C = 2\right) p\left(\mathbf{x}^{(3)} \mid C = 2\right) = \frac{1}{3} \times 0.216 = 0.072$
  - For cluster $C = 3$:
    - Prior: $p\left(C = 3\right) = \frac{1}{3}$
    - Likelihood:
      $p\left(\mathbf{x}^{(3)} \mid C = 3\right) =$
      $= p\left(x_1^{(3)} = 0 \mid C = 3\right) p\left(x_2^{(3)} = 1 \mid C = 3\right) p\left(x_3^{(3)} = 0 \mid C = 3\right) p\left(x_4^{(3)} = 1 \mid C = 3\right) =$
      $0.9 \times 0.1 \times 0.1 \times 0.2 = 0.0018$
    - Joint Probability: $p\left(C = 3, \mathbf{x}^{(3)}\right) = p\left(C = 3\right) p\left(\mathbf{x}^{(3)} \mid C = 3\right) = \frac{1}{3} \times 0.0018 = 0.0006$
  - So, we can compute the normalized posteriors for each cluster:
    - $C = 1$: $p\left(C = 1 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=1,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)+p\left(C=3,\mathbf{x}^{(3)}\right)} = \frac{0.003}{0.003+0.072+0.0006} = 0.0397$

        - ∗ $C=2$: $p\left(C=2 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=2,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)+p\left(C=3,\mathbf{x}^{(3)}\right)} = \frac{0.072}{0.003+0.072+0.0006} = 0.9524$
        - ∗ $C=3$: $p\left(C=3 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=3,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)+p\left(C=3,\mathbf{x}^{(3)}\right)} = \frac{0.0006}{0.003+0.072+0.0006} = 0.0079$
- – For $\mathbf{x}^{(4)}$:
    - • For cluster $C=1$:
        - ∗ Prior: $p(C=1) = \frac{1}{3}$
        - ∗ Likelihood:
          $p\left(\mathbf{x}^{(4)} \mid C=1\right) =$
          $= p\left(x_1^{(4)}=0 \mid C=1\right) p\left(x_2^{(4)}=0 \mid C=1\right) p\left(x_3^{(4)}=1 \mid C=1\right) p\left(x_4^{(4)}=0 \mid C=1\right) = 0.2 \times 0.5 \times 0.1 \times 0.9 = 0.009$
        - ∗ Joint Probability: $p\left(C=1,\mathbf{x}^{(4)}\right) = p(C=1)\, p\left(\mathbf{x}^{(4)} \mid C=1\right) = \frac{1}{3} \times 0.009 = 0.003$
    - • For cluster $C=2$:
        - ∗ Prior: $p(C=2) = \frac{1}{3}$
        - ∗ Likelihood:
          $p\left(\mathbf{x}^{(4)} \mid C=2\right) =$
          $= p\left(x_1^{(4)}=0 \mid C=2\right) p\left(x_2^{(4)}=0 \mid C=2\right) p\left(x_3^{(4)}=1 \mid C=2\right) p\left(x_4^{(4)}=0 \mid C=2\right) = 0.9 \times 0.5 \times 0.4 \times 0.2 = 0.036$
        - ∗ Joint Probability: $p\left(C=2,\mathbf{x}^{(4)}\right) = p(C=2)\, p\left(\mathbf{x}^{(4)} \mid C=2\right) = \frac{1}{3} \times 0.036 = 0.012$
    - • For cluster $C=3$:
        - ∗ Prior: $p(C=3) = \frac{1}{3}$
        - ∗ Likelihood:
          $p\left(\mathbf{x}^{(4)} \mid C=3\right) =$
          $= p\left(x_1^{(4)}=0 \mid C=3\right) p\left(x_2^{(4)}=0 \mid C=3\right) p\left(x_3^{(4)}=1 \mid C=3\right) p\left(x_4^{(4)}=0 \mid C=3\right) = 0.9 \times 0.9 \times 0.9 \times 0.8 = 0.5832$
        - ∗ Joint Probability: $p\left(C=3,\mathbf{x}^{(4)}\right) = p(C=3)\, p\left(\mathbf{x}^{(4)} \mid C=3\right) = \frac{1}{3} \times 0.5832 = 0.1944$
    - • So, we can compute the normalized posteriors for each cluster:
        - ∗ $C=1$: $p\left(C=1 \mid \mathbf{x}^{(4)}\right) = \frac{p\left(C=1,\mathbf{x}^{(4)}\right)}{p\left(C=1,\mathbf{x}^{(4)}\right)+p\left(C=2,\mathbf{x}^{(4)}\right)+p\left(C=3,\mathbf{x}^{(4)}\right)} = \frac{0.003}{0.003+0.012+0.1944} = 0.0143$
        - ∗ $C=2$: $p\left(C=2 \mid \mathbf{x}^{(4)}\right) = \frac{p\left(C=2,\mathbf{x}^{(4)}\right)}{p\left(C=1,\mathbf{x}^{(4)}\right)+p\left(C=2,\mathbf{x}^{(4)}\right)+p\left(C=3,\mathbf{x}^{(4)}\right)} = \frac{0.012}{0.003+0.012+0.1944} = 0.0573$
        - ∗ $C=3$: $p\left(C=3 \mid \mathbf{x}^{(4)}\right) = \frac{p\left(C=3,\mathbf{x}^{(4)}\right)}{p\left(C=1,\mathbf{x}^{(4)}\right)+p\left(C=2,\mathbf{x}^{(4)}\right)+p\left(C=3,\mathbf{x}^{(4)}\right)} = \frac{0.1944}{0.003+0.012+0.1944} = 0.9284$
- – For $\mathbf{x}^{(5)}$:
    - • For cluster $C=1$:
        - ∗ Prior: $p(C=1) = \frac{1}{3}$
        - ∗ Likelihood:
          $p\left(\mathbf{x}^{(5)} \mid C=1\right) =$
          $= p\left(x_1^{(5)}=1 \mid C=1\right) p\left(x_2^{(5)}=1 \mid C=1\right) p\left(x_3^{(5)}=0 \mid C=1\right) p\left(x_4^{(5)}=0 \mid C=1\right) = 0.8 \times 0.5 \times 0.9 \times 0.9 = 0.324$

- $*$ Joint Probability: $p\left(C=1,\mathbf{x}^{(5)}\right)=p\left(C=1\right)p\left(\mathbf{x}^{(5)}\mid C=1\right)=\frac{1}{3}\times 0.324=0.108$
- For cluster $C=2$:
  - $*$ Prior: $p\left(C=2\right)=\frac{1}{3}$
  - $*$ Likelihood:
    $p\left(\mathbf{x}^{(5)}\mid C=2\right)=$
    $=p\left(x_1^{(5)}=1\mid C=2\right)p\left(x_2^{(5)}=1\mid C=2\right)p\left(x_3^{(5)}=0\mid C=2\right)p\left(x_4^{(5)}=0\mid C=2\right)=$
    $0.1\times 0.5\times 0.6\times 0.2=0.006$
  - $*$ Joint Probability: $p\left(C=2,\mathbf{x}^{(5)}\right)=p\left(C=2\right)p\left(\mathbf{x}^{(5)}\mid C=2\right)=\frac{1}{3}\times 0.006=0.002$
- For cluster $C=3$:
  - $*$ Prior: $p\left(C=3\right)=\frac{1}{3}$
  - $*$ Likelihood:
    $p\left(\mathbf{x}^{(5)}\mid C=3\right)=$
    $=p\left(x_1^{(5)}=1\mid C=3\right)p\left(x_2^{(5)}=1\mid C=3\right)p\left(x_3^{(5)}=0\mid C=3\right)p\left(x_4^{(5)}=0\mid C=3\right)=$
    $0.1\times 0.1\times 0.1\times 0.8=0.0008$
  - $*$ Joint Probability: $p\left(C=3,\mathbf{x}^{(5)}\right)=p\left(C=3\right)p\left(\mathbf{x}^{(5)}\mid C=3\right)=\frac{1}{3}\times 0.0008=0.000267$
- So, we can compute the normalized posteriors for each cluster:
  - $*$ $C=1$: $p\left(C=1\mid\mathbf{x}^{(5)}\right)=\frac{p\left(C=1,\mathbf{x}^{(5)}\right)}{p\left(C=1,\mathbf{x}^{(5)}\right)+p\left(C=2,\mathbf{x}^{(5)}\right)+p\left(C=3,\mathbf{x}^{(5)}\right)}=\frac{0.108}{0.108+0.002+0.000267}=0.9795$
  - $*$ $C=2$: $p\left(C=2\mid\mathbf{x}^{(5)}\right)=\frac{p\left(C=2,\mathbf{x}^{(5)}\right)}{p\left(C=1,\mathbf{x}^{(5)}\right)+p\left(C=2,\mathbf{x}^{(5)}\right)+p\left(C=3,\mathbf{x}^{(5)}\right)}=\frac{0.002}{0.108+0.002+0.000267}=0.0181$
  - $*$ $C=3$: $p\left(C=3\mid\mathbf{x}^{(5)}\right)=\frac{p\left(C=3,\mathbf{x}^{(5)}\right)}{p\left(C=1,\mathbf{x}^{(5)}\right)+p\left(C=2,\mathbf{x}^{(5)}\right)+p\left(C=3,\mathbf{x}^{(5)}\right)}=\frac{0.000267}{0.108+0.002+0.000267}=0.0024$

M-Step: Re-estimate cluster parameters such that they fit their assigned elements

For each cluster we need to find the new prior and likelihood parameters. For each likelihood, we compute count and normalize occurrences for each cluster, weigthed by the corresponding posterior:

$$p\left(x_l=1\mid C=c\right)=\frac{\sum_{n=1}^{3}p\left(C=c\mid\mathbf{x}^{(n)}\right)\mathbb{I}\left[x_l^{(n)}=1\right]}{\sum_{n=1}^{3}p\left(C=c\mid\mathbf{x}^{(n)}\right)}$$

For the priors we perform a weighted mean of the posteriors:

$$p\left(C=c\right)=\frac{\sum_{n=1}^{N}p\left(C=c\mid\mathbf{x}^{(n)}\right)}{\sum_{l=1}^{k}\sum_{n=1}^{N}p\left(C=l\mid\mathbf{x}^{(n)}\right)}$$

So, let us estimate the new parameters for each cluster.

- For $C=1$:
  - For the likelihood:
    - $*$ $p\left(x_1=1\mid C=1\right)=\frac{0.961\times 1+0.006\times 0+0.0397\times 0+0.0143\times 0+0.9795\times 1}{0.961+0.006+0.0397+0.0143+0.9795}=0.97$

    - $p\left(x_2 = 1 \mid C = 1\right) = \frac{0.961\times 0+0.006\times 1+0.0397\times 1+0.0143\times 0+0.9795\times 1}{0.961+0.006+0.0397+0.0143+0.9795} = 0.51$
    - $p\left(x_3 = 1 \mid C = 1\right) = \frac{0.961\times 0+0.006\times 1+0.0397\times 0+0.0143\times 1+0.9795\times 0}{0.961+0.006+0.0397+0.0143+0.9795} = 0.01$
    - $p\left(x_4 = 1 \mid C = 1\right) = \frac{0.961\times 0+0.006\times 1+0.0397\times 1+0.0143\times 0+0.9795\times 0}{0.961+0.006+0.0397+0.0143+0.9795} = 0.02$
  - For the prior:
    $p\left(C = 1\right) =$
    $= \frac{0.961+0.006+0.0397+0.0143+0.9795}{(0.961+0.006+0.0397+0.0143+0.9795)+(0.018+0.894+0.9524+0.0573+0.0181)+(0.021+0.1+0.0079+0.9284+0.0024)} =$
    0.4
- For $C = 2$:
  - For the likelihood:
    - $p\left(x_1 = 1 \mid C = 2\right) = \frac{0.018\times 1+0.894\times 0+0.9524\times 0+0.0573\times 0+0.0181\times 1}{0.018+0.894+0.9524+0.0573+0.0181} = 0.0186$
    - $p\left(x_2 = 1 \mid C = 2\right) = \frac{0.018\times 0+0.894\times 1+0.9524\times 1+0.0573\times 0+0.0181\times 1}{0.018+0.894+0.9524+0.0573+0.0181} = 0.9612$
    - $p\left(x_3 = 1 \mid C = 2\right) = \frac{0.018\times 0+0.894\times 1+0.9524\times 0+0.0573\times 1+0.0181\times 0}{0.018+0.894+0.9524+0.0573+0.0181} = 0.4904$
    - $p\left(x_4 = 1 \mid C = 2\right) = \frac{0.018\times 0+0.894\times 1+0.9524\times 1+0.0573\times 0+0.0181\times 0}{0.018+0.894+0.9524+0.0573+0.0181} = 0.9519$
  - For the prior:
    $p\left(C = 2\right) =$
    $= \frac{0.018+0.894+0.9524+0.0573+0.0181}{(0.961+0.006+0.0397+0.0143+0.9795)+(0.018+0.894+0.9524+0.0573+0.0181)+(0.021+0.1+0.0079+0.9284+0.0024)} =$
    0.39
- For $C = 3$:
  - For the likelihood:
    - $p\left(x_1 = 1 \mid C = 3\right) = \frac{0.021\times 1+0.1\times 0+0.0079\times 0+0.9284\times 0+0.0024\times 1}{0.021+0.1+0.0079+0.9284+0.0024} = 0.0221$
    - $p\left(x_2 = 1 \mid C = 3\right) = \frac{0.021\times 0+0.1\times 1+0.0079\times 1+0.9284\times 0+0.0024\times 1}{0.021+0.1+0.0079+0.9284+0.0024} = 0.1041$
    - $p\left(x_3 = 1 \mid C = 3\right) = \frac{0.021\times 0+0.1\times 1+0.0079\times 0+0.9284\times 1+0.0024\times 0}{0.021+0.1+0.0079+0.9284+0.0024} = 0.9705$
    - $p\left(x_4 = 1 \mid C = 3\right) = \frac{0.021\times 0+0.1\times 1+0.0079\times 1+0.9284\times 0+0.0024\times 0}{0.021+0.1+0.0079+0.9284+0.0024} = 0.1018$
  - For the prior:
    $p\left(C = 3\right) =$
    $= \frac{0.021+0.1+0.0079+0.9284+0.0024}{(0.961+0.006+0.0397+0.0143+0.9795)+(0.018+0.894+0.9524+0.0573+0.0181)+(0.021+0.1+0.0079+0.9284+0.0024)} =$
    0.21

Having the new priors and likelihoods we could go for another iteration. However, the exercise only asks for one.

---

b) Verify that after one iteration the probability of the data increased.

Solution:

The probability of the observed data $p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\right)$ can be decomposed into the product of the probability of each point assuming indepent, identically distributed samples:

$$p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\right) = p\left(\mathbf{x}^{(1)}\right) p\left(\mathbf{x}^{(2)}\right) p\left(\mathbf{x}^{(3)}\right) p\left(\mathbf{x}^{(4)}\right) p\left(\mathbf{x}^{(5)}\right)$$

So, we need to compute the probability of each point before and after the EM update.

By the law of total probability we have that:

$$p\left(\mathbf{x}^{(n)}\right) = p\left(\mathbf{x}^{(n)}, C=1\right) + p\left(\mathbf{x}^{(n)}, C=2\right) + p\left(\mathbf{x}^{(n)}, C=3\right)$$

Which can be rewritten as follows:

$$p\left(\mathbf{x}^{(n)}\right) = p\left(C=1\right) p\left(\mathbf{x}^{(n)} \mid C=1\right) + p\left(C=2\right) p\left(\mathbf{x}^{(n)} \mid C=2\right) + p\left(C=3\right) p\left(\mathbf{x}^{(n)} \mid C=3\right)$$

With that in mind let us analyze each scenario. Before the iteration we had:

	$p\left(C=c\right)$
$c=1$	$\frac{1}{3}$
$c=2$	$\frac{1}{3}$
$c=3$	$\frac{1}{3}$

	$p\left(x_1=1 \mid C=c\right)$	$p\left(x_2=1 \mid C=c\right)$	$p\left(x_3=1 \mid C=c\right)$	$p\left(x_4=1 \mid C=c\right)$
$c=1$	0.8	0.5	0.1	0.1
$c=2$	0.1	0.5	0.4	0.8
$c=3$	0.1	0.1	0.9	0.2

- For $\mathbf{x}^{(1)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(1)} \mid C=1\right) = p\left(x_1^{(1)}=1 \mid C=1\right) p\left(x_2^{(1)}=0 \mid C=1\right) p\left(x_3^{(1)}=0 \mid C=1\right) p\left(x_4^{(1)}=0 \mid C=1\right) =$
    0.324
  - Likelihood:
    $p\left(\mathbf{x}^{(1)} \mid C=2\right) = p\left(x_1^{(1)}=1 \mid C=2\right) p\left(x_2^{(1)}=0 \mid C=2\right) p\left(x_3^{(1)}=0 \mid C=2\right) p\left(x_4^{(1)}=0 \mid C=2\right) =$
    0.006
  - Likelihood:
    $p\left(\mathbf{x}^{(1)} \mid C=3\right) = p\left(x_1^{(1)}=1 \mid C=3\right) p\left(x_2^{(1)}=0 \mid C=3\right) p\left(x_3^{(1)}=0 \mid C=3\right) p\left(x_4^{(1)}=0 \mid C=3\right) =$
    0.0072
  - Probability: $p\left(\mathbf{x}^{(1)}\right) = \frac{1}{3}0.324 + \frac{1}{3}0.006 + \frac{1}{3}0.0072 = 0.1124$
- For $\mathbf{x}^{(2)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(2)} \mid C=1\right) = p\left(x_1^{(2)}=0 \mid C=1\right) p\left(x_2^{(2)}=1 \mid C=1\right) p\left(x_3^{(2)}=1 \mid C=1\right) p\left(x_4^{(2)}=1 \mid C=1\right) =$
    0.001
  - Likelihood:
    $p\left(\mathbf{x}^{(2)} \mid C=2\right) = p\left(x_1^{(2)}=0 \mid C=2\right) p\left(x_2^{(2)}=1 \mid C=2\right) p\left(x_3^{(2)}=1 \mid C=2\right) p\left(x_4^{(2)}=1 \mid C=2\right) =$
    $0.9 \times 0.5 \times 0.4 \times 0.8 = 0.144$
  - Likelihood:
    $p\left(\mathbf{x}^{(2)} \mid C=3\right) = p\left(x_1^{(2)}=0 \mid C=3\right) p\left(x_2^{(2)}=1 \mid C=3\right) p\left(x_3^{(2)}=1 \mid C=3\right) p\left(x_4^{(2)}=1 \mid C=3\right) =$
    $0.9 \times 0.1 \times 0.9 \times 0.2 = 0.0162$
  - Probability: $p\left(\mathbf{x}^{(2)}\right) = \frac{1}{3}0.001 + \frac{1}{3}0.144 + \frac{1}{3}0.0162 = 0.0537$
- For $\mathbf{x}^{(3)}$:

  - Likelihood:
    $p\left(\mathbf{x}^{(3)} \mid C=1\right) = p\left(x_1^{(3)}=0 \mid C=1\right) p\left(x_2^{(3)}=1 \mid C=1\right) p\left(x_3^{(3)}=0 \mid C=1\right) p\left(x_4^{(3)}=1 \mid C=1\right) =$
    $0.2 \times 0.5 \times 0.9 \times 0.1 = 0.009$
  - Likelihood:
    $p\left(\mathbf{x}^{(3)} \mid C=2\right) = p\left(x_1^{(3)}=0 \mid C=2\right) p\left(x_2^{(3)}=1 \mid C=2\right) p\left(x_3^{(3)}=0 \mid C=2\right) p\left(x_4^{(3)}=1 \mid C=2\right) =$
    $0.9 \times 0.5 \times 0.6 \times 0.8 = 0.216$
  - Likelihood:
    $p\left(\mathbf{x}^{(3)} \mid C=3\right) = p\left(x_1^{(3)}=0 \mid C=3\right) p\left(x_2^{(3)}=1 \mid C=3\right) p\left(x_3^{(3)}=0 \mid C=3\right) p\left(x_4^{(3)}=1 \mid C=3\right) =$
    $0.9 \times 0.1 \times 0.1 \times 0.2 = 0.0018$
  - Probability: $p\left(\mathbf{x}^{(1)}\right) = \frac{1}{3}0.009 + \frac{1}{3}0.216 + \frac{1}{3}0.0018 = 0.0756$
- For $\mathbf{x}^{(4)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(4)} \mid C=1\right) = p\left(x_1^{(4)}=0 \mid C=1\right) p\left(x_2^{(4)}=0 \mid C=1\right) p\left(x_3^{(4)}=1 \mid C=1\right) p\left(x_4^{(4)}=0 \mid C=1\right) =$
    $0.2 \times 0.5 \times 0.1 \times 0.9 = 0.009$
  - Likelihood:
    $p\left(\mathbf{x}^{(4)} \mid C=2\right) = p\left(x_1^{(4)}=0 \mid C=2\right) p\left(x_2^{(4)}=0 \mid C=2\right) p\left(x_3^{(4)}=1 \mid C=2\right) p\left(x_4^{(4)}=0 \mid C=2\right) =$
    $0.9 \times 0.5 \times 0.4 \times 0.2 = 0.036$
  - Likelihood:
    $p\left(\mathbf{x}^{(4)} \mid C=3\right) = p\left(x_1^{(4)}=0 \mid C=3\right) p\left(x_2^{(4)}=0 \mid C=3\right) p\left(x_3^{(4)}=1 \mid C=3\right) p\left(x_4^{(4)}=0 \mid C=3\right) =$
    $0.9 \times 0.9 \times 0.9 \times 0.8 = 0.5832$
  - Probability: $p\left(\mathbf{x}^{(1)}\right) = \frac{1}{3}0.009 + \frac{1}{3}0.036 + \frac{1}{3}0.5832 = 0.2094$
- For $\mathbf{x}^{(5)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(5)} \mid C=1\right) = p\left(x_1^{(5)}=1 \mid C=1\right) p\left(x_2^{(5)}=1 \mid C=1\right) p\left(x_3^{(5)}=0 \mid C=1\right) p\left(x_4^{(5)}=0 \mid C=1\right) =$
    $0.8 \times 0.5 \times 0.9 \times 0.9 = 0.324$
  - Likelihood:
    $p\left(\mathbf{x}^{(5)} \mid C=2\right) = p\left(x_1^{(5)}=1 \mid C=2\right) p\left(x_2^{(5)}=1 \mid C=2\right) p\left(x_3^{(5)}=0 \mid C=2\right) p\left(x_4^{(5)}=0 \mid C=2\right) =$
    $0.1 \times 0.5 \times 0.6 \times 0.2 = 0.006$
  - Likelihood:
    $p\left(\mathbf{x}^{(5)} \mid C=3\right) = p\left(x_1^{(5)}=1 \mid C=3\right) p\left(x_2^{(5)}=1 \mid C=3\right) p\left(x_3^{(5)}=0 \mid C=3\right) p\left(x_4^{(5)}=0 \mid C=3\right) =$
    $0.1 \times 0.1 \times 0.1 \times 0.8 = 0.0008$
  - Probability: $p\left(\mathbf{x}^{(1)}\right) = \frac{1}{3}0.324 + \frac{1}{3}0.006 + \frac{1}{3}0.0008 = 0.1103$

So, we have that:

$$p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\right) = p\left(\mathbf{x}^{(1)}\right) p\left(\mathbf{x}^{(2)}\right) p\left(\mathbf{x}^{(3)}\right) p\left(\mathbf{x}^{(4)}\right) p\left(\mathbf{x}^{(5)}\right) = 1.054 \times 10^{-5}$$

From a) we have that after one iteration the parameters are as follows.

	$p(C=c)$
$c=1$	0.40
$c=2$	0.39
$c=3$	0.21

	$p\left(x_1=1 \mid C=c\right)$	$p\left(x_2=1 \mid C=c\right)$	$p\left(x_3=1 \mid C=c\right)$	$p\left(x_4=1 \mid C=c\right)$
$c=1$	0.97	0.51	0.01	0.02
$c=2$	0.02	0.96	0.49	0.95
$c=3$	0.02	0.10	0.97	0.10

- For $\mathbf{x}^{(1)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(1)} \mid C=1\right)=p\left(x_1^{(1)}=1 \mid C=1\right) p\left(x_2^{(1)}=0 \mid C=1\right) p\left(x_3^{(1)}=0 \mid C=1\right) p\left(x_4^{(1)}=0 \mid C=1\right)=$
    0.46
  - Likelihood:
    $p\left(\mathbf{x}^{(1)} \mid C=2\right)=p\left(x_1^{(1)}=1 \mid C=2\right) p\left(x_2^{(1)}=0 \mid C=2\right) p\left(x_3^{(1)}=0 \mid C=2\right) p\left(x_4^{(1)}=0 \mid C=2\right)=$
    0.0000204
  - Likelihood:
    $p\left(\mathbf{x}^{(1)} \mid C=3\right)=p\left(x_1^{(1)}=1 \mid C=3\right) p\left(x_2^{(1)}=0 \mid C=3\right) p\left(x_3^{(1)}=0 \mid C=3\right) p\left(x_4^{(1)}=0 \mid C=3\right)=$
    0.000486
  - Probability: $p\left(\mathbf{x}^{(1)}\right)=0.4 \times 0.46+0.39 \times 0.0000204+0.21 \times 0.000486=$
    0.1846
- For $\mathbf{x}^{(2)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(2)} \mid C=1\right)=p\left(x_1^{(2)}=0 \mid C=1\right) p\left(x_2^{(2)}=1 \mid C=1\right) p\left(x_3^{(2)}=1 \mid C=1\right) p\left(x_4^{(2)}=1 \mid C=1\right)=$
    0.00000306
  - Likelihood:
    $p\left(\mathbf{x}^{(2)} \mid C=2\right)=p\left(x_1^{(2)}=0 \mid C=2\right) p\left(x_2^{(2)}=1 \mid C=2\right) p\left(x_3^{(2)}=1 \mid C=2\right) p\left(x_4^{(2)}=1 \mid C=2\right)=$
    0.438
  - Likelihood:
    $p\left(\mathbf{x}^{(2)} \mid C=3\right)=p\left(x_1^{(2)}=0 \mid C=3\right) p\left(x_2^{(2)}=1 \mid C=3\right) p\left(x_3^{(2)}=1 \mid C=3\right) p\left(x_4^{(2)}=1 \mid C=3\right)=$
    0.009506
  - Probability: $p\left(\mathbf{x}^{(2)}\right)=0.4 \times 0.00000306+0.39 \times 0.438+0.21 \times 0.009506=$
    0.1728
- For $\mathbf{x}^{(3)}$:
  - Likelihood:
    $p\left(\mathbf{x}^{(3)} \mid C=1\right)=p\left(x_1^{(3)}=0 \mid C=1\right) p\left(x_2^{(3)}=1 \mid C=1\right) p\left(x_3^{(3)}=0 \mid C=1\right) p\left(x_4^{(3)}=1 \mid C=1\right)=$
    0.000303
  - Likelihood:
    $p\left(\mathbf{x}^{(3)} \mid C=2\right)=p\left(x_1^{(3)}=0 \mid C=2\right) p\left(x_2^{(3)}=1 \mid C=2\right) p\left(x_3^{(3)}=0 \mid C=2\right) p\left(x_4^{(3)}=1 \mid C=2\right)=$
    0.456
  - Likelihood:
    $p\left(\mathbf{x}^{(3)} \mid C=3\right)=p\left(x_1^{(3)}=0 \mid C=3\right) p\left(x_2^{(3)}=1 \mid C=3\right) p\left(x_3^{(3)}=0 \mid C=3\right) p\left(x_4^{(3)}=1 \mid C=3\right)=$
    0.000294
  - Probability: $p\left(\mathbf{x}^{(3)}\right)=0.4 \times 0.000303+0.39 \times 0.456+0.21 \times 0.000294=$
    0.178
- For $\mathbf{x}^{(4)}$:

- Likelihood:
  $p\left(\mathbf{x}^{(4)} \mid C=1\right)=p\left(x_1^{(4)}=0 \mid C=1\right) p\left(x_2^{(4)}=0 \mid C=1\right) p\left(x_3^{(4)}=1 \mid C=1\right) p\left(x_4^{(4)}=0 \mid C=1\right)=$
  0.0001441
- Likelihood:
  $p\left(\mathbf{x}^{(4)} \mid C=2\right)=p\left(x_1^{(4)}=0 \mid C=2\right) p\left(x_2^{(4)}=0 \mid C=2\right) p\left(x_3^{(4)}=1 \mid C=2\right) p\left(x_4^{(4)}=0 \mid C=2\right)=$
  0.00096
- Likelihood:
  $p\left(\mathbf{x}^{(4)} \mid C=3\right)=p\left(x_1^{(4)}=0 \mid C=3\right) p\left(x_2^{(4)}=0 \mid C=3\right) p\left(x_3^{(4)}=1 \mid C=3\right) p\left(x_4^{(4)}=0 \mid C=3\right)=$
  0.77
- Probability: $p\left(\mathbf{x}^{(4)}\right)=0.4 \times 0.0001441+0.39 \times 0.00096+0.21 \times 0.77=$
  0.1621

– For $\mathbf{x}^{(5)}$:

- Likelihood:
  $p\left(\mathbf{x}^{(5)} \mid C=1\right)=p\left(x_1^{(5)}=1 \mid C=1\right) p\left(x_2^{(5)}=1 \mid C=1\right) p\left(x_3^{(5)}=0 \mid C=1\right) p\left(x_4^{(5)}=0 \mid C=1\right)=$
  0.48
- Likelihood:
  $p\left(\mathbf{x}^{(5)} \mid C=2\right)=p\left(x_1^{(5)}=1 \mid C=2\right) p\left(x_2^{(5)}=1 \mid C=2\right) p\left(x_3^{(5)}=0 \mid C=2\right) p\left(x_4^{(5)}=0 \mid C=2\right)=$
  0.0004896
- Likelihood:
  $p\left(\mathbf{x}^{(5)} \mid C=3\right)=p\left(x_1^{(5)}=1 \mid C=3\right) p\left(x_2^{(5)}=1 \mid C=3\right) p\left(x_3^{(5)}=0 \mid C=3\right) p\left(x_4^{(5)}=0 \mid C=3\right)=$
  0.000054
- Probability: $p\left(\mathbf{x}^{(5)}\right)=0.4 \times 0.48+0.39 \times 0.0004896+0.21 \times 0.000054=$
  0.1922

So, we have that:

$$p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\right)=p\left(\mathbf{x}^{(1)}\right) p\left(\mathbf{x}^{(2)}\right) p\left(\mathbf{x}^{(3)}\right) p\left(\mathbf{x}^{(4)}\right) p\left(\mathbf{x}^{(5)}\right)=0.00018$$

As we can see, after the iteration, the data is more probable which suggests that the model captures the data better.

---

2) Consider the following training data without labels:

$$\left\{\mathbf{x}^{(1)}=\begin{pmatrix}4\end{pmatrix}, \mathbf{x}^{(2)}=\begin{pmatrix}0\end{pmatrix}, \mathbf{x}^{(3)}=\begin{pmatrix}1\end{pmatrix}\right\}$$

We want to model the data with a mixture of two normal distributions. Initialize the likelihoods as follows:

$$p(\mathbf{x} \mid C=1)=\mathcal{N}\left(\mu^1=0, \sigma^1=1\right)$$

$$p(\mathbf{x} \mid C=2)=\mathcal{N}\left(\mu^2=1, \sigma^2=1\right)$$

Also, initialize the priors as follows:

$$p\left(C=1\right)=0.5$$

$$p\left(C=2\right)=0.5$$

a) Perform one expectation maximization iteration.

Solution:
Each iteration has two steps. Let us do one:
E-Step: Assign each point to the cluster that yields higher posterior

- For $\mathbf{x}^{(1)}$:
  - For cluster $C=1$:
    - Prior: $p\left(C=1\right)=0.5$
    - Likelihood: $p\left(\mathbf{x}^{(1)}\mid C=1\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(1)}-\mu^1\right)^2}{(\sigma^1)^2}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(4-0)^2}{1^2}\right)=$ $\frac{1}{\sqrt{2\pi}}\exp\left(-8\right)=0.000134$
    - Joint Probability: $p\left(C=1,\mathbf{x}^{(1)}\right)=p\left(C=1\right)p\left(\mathbf{x}^{(1)}\mid C=1\right)=0.5\times 0.000134=0.000067$
  - For cluster $C=2$:
    - Prior: $p\left(C=2\right)=0.5$
    - Likelihood: $p\left(\mathbf{x}^{(1)}\mid C=2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(1)}-\mu^2\right)^2}{(\sigma^2)^2}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(4-1)^2}{1^2}\right)=$ $\frac{1}{\sqrt{2\pi}}\exp\left(-4.5\right)=0.0044$
    - Joint Probability: $p\left(C=2,\mathbf{x}^{(1)}\right)=p\left(C=2\right)p\left(\mathbf{x}^{(1)}\mid C=2\right)=0.5\times 0.0044=0.0022$
  - So, we can compute the normalized posteriors for each cluster:
    - $C=1$: $p\left(C=1\mid\mathbf{x}^{(1)}\right)=\frac{p\left(C=1,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)}=\frac{0.000067}{0.000067+0.0022}=$ 0.0293
    - $C=2$: $p\left(C=2\mid\mathbf{x}^{(1)}\right)=\frac{p\left(C=2,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)}=\frac{0.0022}{0.000067+0.0022}=$ 0.9707
- For $\mathbf{x}^{(2)}$:
  - For cluster $C=1$:
    - Prior: $p\left(C=1\right)=0.5$
    - Likelihood: $p\left(\mathbf{x}^{(2)}\mid C=1\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(2)}-\mu^1\right)^2}{(\sigma^1)^2}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(0-0)^2}{1^2}\right)=$ 0.399
    - Joint Probability: $p\left(C=1,\mathbf{x}^{(2)}\right)=p\left(C=1\right)p\left(\mathbf{x}^{(2)}\mid C=1\right)=0.5\times 0.3989=0.1995$
  - For cluster $C=2$:
    - Prior: $p\left(C=2\right)=0.4$

        - ∗ Likelihood: $p\left(\mathbf{x}^{(2)} \mid C=2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(2)}-\mu^2\right)^2}{(\sigma^2)^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(0-1)^2}{1^2}\right) = 0.242$
        - ∗ Joint Probability: $p\left(C=2, \mathbf{x}^{(2)}\right) = p\left(C=2\right) p\left(\mathbf{x}^{(2)} \mid C=2\right) = 0.5 \times 0.242 = 0.121$
    - So, we can compute the posteriors for each cluster:
        - ∗ $C=1$: $p\left(C=1 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=1,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)} = \frac{0.1995}{0.1995+0.121} = 0.62$
        - ∗ $C=2$: $p\left(C=2 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=2,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)} = \frac{0.121}{0.1995+0.121} = 0.38$
- – For $\mathbf{x}^{(3)}$:
    - For cluster $C=1$:
        - ∗ Prior: $p\left(C=1\right) = 0.5$
        - ∗ Likelihood: $p\left(\mathbf{x}^{(3)} \mid C=1\right) = \frac{1}{\sqrt{2\pi(\sigma^1)^2}} \exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(3)}-\mu^1\right)^2}{(\sigma^1)^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(1-0)^2}{1^2}\right) = 0.242$
        - ∗ Joint Probability: $p\left(C=1, \mathbf{x}^{(3)}\right) = p\left(C=1\right) p\left(\mathbf{x}^{(3)} \mid C=1\right) = 0.5 \times 0.242 = 0.121$
    - For cluster $C=2$:
        - ∗ Prior: $p\left(C=2\right) = 0.5$
        - ∗ Likelihood: $p\left(\mathbf{x}^{(3)} \mid C=2\right) = \frac{1}{\sqrt{2\pi(\sigma^2)^2}} \exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(3)}-\mu^2\right)^2}{(\sigma^2)^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(1-1)^2}{1^2}\right) = 0.399$
        - ∗ Joint Probability: $p\left(C=2, \mathbf{x}^{(3)}\right) = p\left(C=2\right) p\left(\mathbf{x}^{(3)} \mid C=2\right) = 0.5 \times 0.399 = 0.1995$
    - So, we can compute the posteriors for each cluster:
        - ∗ $C=1$: $p\left(C=1 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=1,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)} = \frac{0.121}{0.121+0.1995} = 0.38$
        - ∗ $C=2$: $p\left(C=2 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=2,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)} = \frac{0.1995}{0.121+0.1995} = 0.62$

M-Step: Re-estimate cluster parameters such that they fit their assigned elements

For each cluster we need to find the new prior and likelihood parameters. For each likelihood, we compute the mean and standard deviation using all points weighted by their posteriors:

$$\mu^c = \frac{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right) \mathbf{x}^{(n)}}{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right)}$$

And the covariance matrix as follows.

$$\sigma^c = \sqrt{\frac{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right) \left(\mathbf{x}^{(n)}-\mu^c\right)^2}{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right)}}$$

For the priors we perform a weighted mean of the posteriors:

$$p\left(C=c\right)=\frac{\sum_{n=1}^{N}p\left(C=c\mid \mathbf{x}^{(n)}\right)}{\sum_{l=1}^{k}\sum_{n=1}^{N}p\left(C=l\mid \mathbf{x}^{(n)}\right)}$$

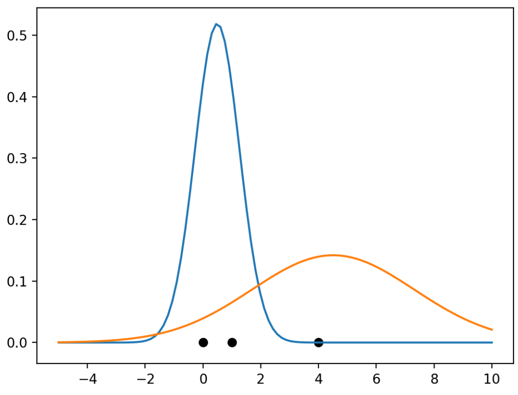
So, let us estimate the new parameters for each cluster.

- For $C=1$:
  - For the likelihood:
    - $\mu^1=\frac{0.0293\left(4\right)+0.62\left(0\right)+0.38\left(1\right)}{0.0293+0.62+0.38}=0.495$
    - $\sigma^1=\sqrt{\frac{0.0293(4-0.495)^2+0.62(0-0.495)^2+0.38(1-0.495)^2}{0.0293+0.62+0.38}}=0.769$
    - So, the new likelihood is: $p\left(\mathbf{x}\mid C=1\right)=\mathcal{N}\left(\mu^1=0.495,\sigma^1=0.769\right)$
  - For the prior: $p\left(C=1\right)=\frac{p\left(C=1|\mathbf{x}^{(1)}\right)+p\left(C=1|\mathbf{x}^{(2)}\right)+p\left(C=1|\mathbf{x}^{(3)}\right)}{p\left(C=1|\mathbf{x}^{(1)}\right)+p\left(C=1|\mathbf{x}^{(2)}\right)+p\left(C=1|\mathbf{x}^{(3)}\right)+p\left(C=2|\mathbf{x}^{(1)}\right)+p\left(C=2|\mathbf{x}^{(2)}\right)+p\left(C=2|\mathbf{x}^{(3)}\right)}=$ $\frac{0.0293+0.6225+0.3775}{(0.0293+0.6225+0.3775)+(0.9707+0.3775+0.6225)}=0.3431$
- For $C=2$:
  - For the likelihood:
    - $\mu^2=\frac{0.9707\left(4\right)+0.3775\left(0\right)+0.6225\left(1\right)}{0.9707+0.3775+0.6225}=4.5$
    - $\sigma^2=\sqrt{\frac{0.9707(4-4.5)^2+0.3775(0-4.5)^2+0.6225(1-4.5)^2}{0.9707+0.3775+0.6225}}=2.81$
    - So, the new likelihood is: $p\left(\mathbf{x}\mid C=2\right)=\mathcal{N}\left(\mu^2=4.5,\sigma^2=2.81\right)$
  - For the prior: $p\left(C=2\right)=\frac{p\left(C=2|\mathbf{x}^{(1)}\right)+p\left(C=2|\mathbf{x}^{(2)}\right)+p\left(C=2|\mathbf{x}^{(3)}\right)}{p\left(C=1|\mathbf{x}^{(1)}\right)+p\left(C=1|\mathbf{x}^{(2)}\right)+p\left(C=1|\mathbf{x}^{(3)}\right)+p\left(C=2|\mathbf{x}^{(1)}\right)+p\left(C=2|\mathbf{x}^{(2)}\right)+p\left(C=2|\mathbf{x}^{(3)}\right)}=$ $\frac{0.9707+0.3775+0.6225}{(0.0293+0.6225+0.3775)+(0.9707+0.3775+0.6225)}=0.6569$

Having the new priors and likelihoods we could go for another iteration. However, the exercise only asks for one.

b) Plot the points and sketch the clusters.

Solution:



3) Consider the following training data without labels:

$$\left\{ \mathbf{x}^{(1)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$$

We want to model the data with a mixture of two multivariate normal distributions. Initialize the likelihoods as follows:

$$p\left(\mathbf{x} \mid C = 1\right) = \mathcal{N}\left(\mu^1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{\Sigma}^1 = \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix}\right)$$

$$p\left(\mathbf{x} \mid C = 2\right) = \mathcal{N}\left(\mu^2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma}^2 = \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix}\right)$$

Also, initialize the priors as follows:

$$p\left(C = 1\right) = 0.6$$

$$p\left(C = 2\right) = 0.4$$

a) Perform one expectation maximization iteration.

Solution:
Each iteration has two steps. Let us do one:
E-Step: Assign each point to the cluster that yields higher posterior

- For $\mathbf{x}^{(1)}$:
  - For cluster $C = 1$:
    - Prior: $p\left(C = 1\right) = 0.6$
    - Likelihood: $p\left(\mathbf{x}^{(1)} \mid C = 1\right) = \frac{1}{2\pi} \frac{1}{det(\mathit{\Sigma}^1)} \exp\left(-\frac{1}{2}\left(\mathbf{x}^{(1)} - \mu^1\right)^T \left(\mathit{\Sigma}^1\right)^{-1} \left(\mathbf{x}^{(1)} - \mu^1\right)\right) =$ $\frac{1}{2\pi}\frac{1}{1} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^T \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix} \left(\begin{pmatrix} 2 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix} 0 \\ 0 \end{pmatrix}^T \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) =$ $\frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix} 0\,0 \end{pmatrix} \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = \frac{1}{2\pi} \exp\left(0\right) = \frac{1}{2\pi}$
    - Joint Probability: $p\left(C = 1, \mathbf{x}^{(1)}\right) = p\left(C = 1\right) p\left(\mathbf{x}^{(1)} \mid C = 1\right) = 0.6 \times \frac{1}{2\pi} = 0.095$
  - For cluster $C = 2$:
    - Prior: $p\left(C = 2\right) = 0.4$
    - Likelihood: $p\left(\mathbf{x}^{(1)} \mid C = 2\right) = \frac{1}{2\pi} \frac{1}{det(\mathit{\Sigma}^2)} \exp\left(-\frac{1}{2}\left(\mathbf{x}^{(1)} - \mu^2\right)^T \left(\mathit{\Sigma}^2\right)^{-1} \left(\mathbf{x}^{(1)} - \mu^2\right)\right) =$ $\frac{1}{2\pi}\frac{1}{1} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)^T \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix} \left(\begin{pmatrix} 2 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix} 2 \\ 2 \end{pmatrix}^T \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right) =$ $\frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix} 2\,2 \end{pmatrix} \begin{pmatrix} 1\,0 \\ 0\,1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right) = \frac{1}{2\pi} \exp\left(-4\right) = 0.003$

        - Joint Probability: $p\left(C=2,\mathbf{x}^{(1)}\right) = p\left(C=2\right)p\left(\mathbf{x}^{(1)} \mid C=2\right) = 0.4\times 0.003 = 0.0012$
    - So, we can compute the normalized posteriors for each cluster:
        - $C=1$: $p\left(C=1 \mid \mathbf{x}^{(1)}\right) = \frac{p\left(C=1,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)} = \frac{0.095}{0.095+0.0012} = 0.9879$
        - $C=2$: $p\left(C=2 \mid \mathbf{x}^{(1)}\right) = \frac{p\left(C=2,\mathbf{x}^{(1)}\right)}{p\left(C=1,\mathbf{x}^{(1)}\right)+p\left(C=2,\mathbf{x}^{(1)}\right)} = \frac{0.0012}{0.095+0.0012} = 0.0121$
- For $\mathbf{x}^{(2)}$:
    - For cluster $C=1$:
        - Prior: $p\left(C=1\right)=0.6$
        - Likelihood: $p\left(\mathbf{x}^{(2)} \mid C=1\right) = \frac{1}{2\pi}\frac{1}{det(\varSigma^1)}\exp\left(-\frac{1}{2}\left(\mathbf{x}^{(2)}-\mu^1\right)^T\left(\varSigma^1\right)^{-1}\left(\mathbf{x}^{(2)}-\mu^1\right)\right) = \frac{1}{2\pi}\frac{1}{1}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0\\2\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\left(\begin{pmatrix}0\\2\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}-2\\0\end{pmatrix}^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}-2\\0\end{pmatrix}\right)$ $\frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}-2\;0\end{pmatrix}\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}-2\\0\end{pmatrix}\right) = \frac{1}{2\pi}\exp\left(-2\right) = 0.0215$
        - Joint Probability: $p\left(C=1,\mathbf{x}^{(2)}\right) = p\left(C=1\right)p\left(\mathbf{x}^{(2)} \mid C=1\right) = 0.6\times 0.0215 = 0.0129$
    - For cluster $C=2$:
        - Prior: $p\left(C=2\right)=0.4$
        - Likelihood: $p\left(\mathbf{x}^{(2)} \mid C=2\right) = \frac{1}{2\pi}\frac{1}{det(\varSigma^2)}\exp\left(-\frac{1}{2}\left(\mathbf{x}^{(2)}-\mu^2\right)^T\left(\varSigma^2\right)^{-1}\left(\mathbf{x}^{(2)}-\mu^2\right)\right) = \frac{1}{2\pi}\frac{1}{1}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0\\2\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\left(\begin{pmatrix}0\\2\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}0\\2\end{pmatrix}^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}0\\2\end{pmatrix}\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}0\;2\end{pmatrix}\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}0\\2\end{pmatrix}\right) = \frac{1}{2\pi}\exp\left(-2\right) = 0.0215$
        - Joint Probability: $p\left(C=2,\mathbf{x}^{(2)}\right) = p\left(C=2\right)p\left(\mathbf{x}^{(2)} \mid C=2\right) = 0.4\times 0.0215 = 0.0086$
    - So, we can compute the posteriors for each cluster:
        - $C=1$: $p\left(C=1 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=1,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)} = \frac{0.0129}{0.0129+0.0086} = 0.6$
        - $C=2$: $p\left(C=2 \mid \mathbf{x}^{(2)}\right) = \frac{p\left(C=2,\mathbf{x}^{(2)}\right)}{p\left(C=1,\mathbf{x}^{(2)}\right)+p\left(C=2,\mathbf{x}^{(2)}\right)} = \frac{0.0086}{0.0129+0.0086} = 0.4$
- For $\mathbf{x}^{(3)}$:
    - For cluster $C=1$:
        - Prior: $p\left(C=1\right)=0.6$
        - Likelihood: $p\left(\mathbf{x}^{(3)} \mid C=1\right) = \frac{1}{2\pi}\frac{1}{det(\varSigma^1)}\exp\left(-\frac{1}{2}\left(\mathbf{x}^{(3)}-\mu^1\right)^T\left(\varSigma^1\right)^{-1}\left(\mathbf{x}^{(3)}-\mu^1\right)\right) = \frac{1}{2\pi}\frac{1}{1}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\left(\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}-2\\-2\end{pmatrix}^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix}\right)$ $\frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}-2\;-2\end{pmatrix}\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix}\right) = \frac{1}{2\pi}\exp\left(-4\right) = 0.003$

  - Joint Probability: $p\left(C=1,\mathbf{x}^{(3)}\right) = p\left(C=1\right)p\left(\mathbf{x}^{(3)} \mid C=1\right) = 0.6\times 0.003 = 0.0017$
- For cluster $C=2$:
  - Prior: $p\left(C=2\right) = 0.4$
  - Likelihood: $p\left(\mathbf{x}^{(3)} \mid C=2\right) = \frac{1}{2\pi}\frac{1}{det(\Sigma^2)}\exp\left(-\frac{1}{2}\left(\mathbf{x}^{(3)}-\mu^2\right)^T\left(\Sigma^2\right)^{-1}\left(\mathbf{x}^{(3)}-\mu^2\right)\right) = \frac{1}{2\pi}\frac{1}{1}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\left(\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}0\\0\end{pmatrix}^T\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(0\;0\right)\begin{pmatrix}1\,0\\0\,1\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}\right) = \frac{1}{2\pi}\exp\left(0\right) = \frac{1}{2\pi}$
  - Joint Probability: $p\left(C=2,\mathbf{x}^{(3)}\right) = p\left(C=2\right)p\left(\mathbf{x}^{(3)} \mid C=2\right) = 0.4\times \frac{1}{2\pi} = 0.0637$
- So, we can compute the posteriors for each cluster:
  - $C=1$: $p\left(C=1 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=1,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)} = \frac{0.0017}{0.0017+0.0637} = 0.0267$
  - $C=2$: $p\left(C=2 \mid \mathbf{x}^{(3)}\right) = \frac{p\left(C=2,\mathbf{x}^{(3)}\right)}{p\left(C=1,\mathbf{x}^{(3)}\right)+p\left(C=2,\mathbf{x}^{(3)}\right)} = \frac{0.0637}{0.0017+0.0637} = 0.9733$

M-Step: Re-estimate cluster parameters such that they fit their assigned elements

For each cluster we need to find the new prior and likelihood parameters. For each likelihood, we compute the mean and covariances using all points weighted by their posteriors:

$$\mu^c = \frac{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right)}$$

And the covariance matrix as follows.

$$\Sigma_{ij}^c = \frac{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right)\left(\mathbf{x}_i^{(n)}-\mu_i^c\right)\left(\mathbf{x}_j^{(n)}-\mu_j^c\right)}{\sum_{n=1}^{3} p\left(C=c \mid \mathbf{x}^{(n)}\right)}$$

For the priors we perform a weighted mean of the posteriors:

$$p\left(C=c\right) = \frac{\sum_{n=1}^{N} p\left(C=c \mid \mathbf{x}^{(n)}\right)}{\sum_{l=1}^{k}\sum_{n=1}^{N} p\left(C=l \mid \mathbf{x}^{(n)}\right)}$$

So, let us estimate the new parameters for each cluster.

– For $C=1$:
- For the likelihood:
  - $\mu^1 = \frac{0.9879\begin{pmatrix}2\\2\end{pmatrix}+0.6\begin{pmatrix}0\\2\end{pmatrix}+0.0267\begin{pmatrix}0\\0\end{pmatrix}}{0.9879+0.6+0.0267} = \frac{\begin{pmatrix}1.9759\\3.1759\end{pmatrix}}{1.6147} = \begin{pmatrix}1.2237\\1.9669\end{pmatrix}$
  - $\Sigma_{11}^1 = \frac{0.9879(2-1.2237)(2-1.2237)+0.6(0-1.2237)(0-1.2237)+0.0267(0-1.2237)(0-1.2237)}{0.9879+0.6+0.0267} = 0.94996$

        - $\varSigma^1_{12} = \varSigma^1_{21} = \frac{0.9879(2-1.2237)(2-1.9669)+0.6(0-1.2237)(2-1.9669)+0.0267(0-1.2237)(0-1.9669)}{0.9879+0.6+0.0267} = 0.0405$
        - $\varSigma^1_{22} = \frac{0.9879(2-1.9669)(2-1.9669)+0.6(2-1.9669)(2-1.9669)+0.0267(0-1.9669)(0-1.9669)}{0.9879+0.6+0.0267} = 0.0651$
        - $\varSigma^1 = \begin{pmatrix} 0.94996 & 0.0405 \\ 0.0405 & 0.0651 \end{pmatrix}$
        - So, the new likelihood is: $p\left(\mathbf{x} \mid C=1\right) = \mathcal{N}\left(\mu^1 = \begin{pmatrix} 1.2237 \\ 1.9669 \end{pmatrix}, \boldsymbol{\Sigma}^1 = \begin{pmatrix} 0.94996 & 0.0405 \\ 0.0405 & 0.0651 \end{pmatrix}\right)$

    - For the prior: $p\left(C=1\right) = \frac{p\left(C=1|\mathbf{x}^{(1)}\right)+p\left(C=1|\mathbf{x}^{(2)}\right)+p\left(C=1|\mathbf{x}^{(3)}\right)}{p\left(C=1|\mathbf{x}^{(1)}\right)+p\left(C=1|\mathbf{x}^{(2)}\right)+p\left(C=1|\mathbf{x}^{(3)}\right)+p\left(C=2|\mathbf{x}^{(1)}\right)+p\left(C=2|\mathbf{x}^{(2)}\right)+p\left(C=2|\mathbf{x}^{(3)}\right)} = \frac{0.9879+0.6+0.0267}{(0.9879+0.6+0.0267)+(0.0121+0.4+0.9733)} = 0.5382$

- For $C=2$:

    - For the likelihood:

        - $\mu^2 = \frac{0.0121\begin{pmatrix} 2 \\ 2 \end{pmatrix}+0.4\begin{pmatrix} 0 \\ 2 \end{pmatrix}+0.9733\begin{pmatrix} 0 \\ 0 \end{pmatrix}}{0.0121+0.4+0.9733} = \begin{pmatrix} 0.0174 \\ 0.5949 \end{pmatrix}$
        - $\varSigma^2_{11} = \frac{0.0121(2-0.0174)(2-0.0174)+0.4(0-0.0174)(0-0.0174)+0.9733(0-0.0174)(0-0.0174)}{0.0121+0.4+0.9733} = 0.0345$
        - $\varSigma^2_{12} = \varSigma^2_{21} = \frac{0.0121(2-0.0174)(2-0.5949)+0.4(0-0.0174)(2-0.5949)+0.9733(0-0.0174)(0-0.5949)}{0.0121+0.4+0.9733} = 0.0245$
        - $\varSigma^1_{22} = \frac{0.0121(2-0.5949)(2-0.5949)+0.4(2-0.5949)(2-0.5949)+0.9733(0-0.5949)(0-0.5949)}{0.0121+0.4+0.9733} = 0.8359$
        - $\varSigma^2 = \begin{pmatrix} 0.0345 & 0.0245 \\ 0.0245 & 0.8359 \end{pmatrix}$
        - So, the new likelihood is: $p\left(\mathbf{x} \mid C=2\right) = \mathcal{N}\left(\mu^2 = \begin{pmatrix} 0.0174 \\ 0.5949 \end{pmatrix}, \boldsymbol{\Sigma}^2 = \begin{pmatrix} 0.0345 & 0.0245 \\ 0.0245 & 0.8359 \end{pmatrix}\right)$

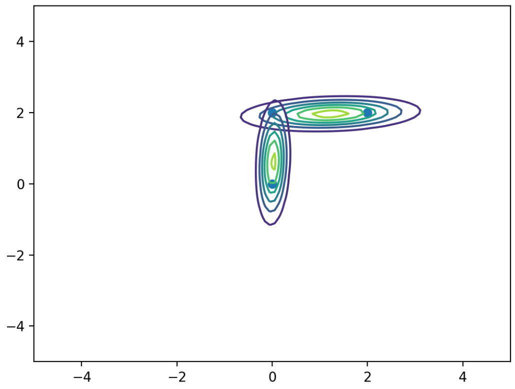
    - For the prior: $p\left(C=2\right) = \frac{p\left(C=2|\mathbf{x}^{(1)}\right)+p\left(C=2|\mathbf{x}^{(2)}\right)+p\left(C=2|\mathbf{x}^{(3)}\right)}{p\left(C=1|\mathbf{x}^{(1)}\right)+p\left(C=1|\mathbf{x}^{(2)}\right)+p\left(C=1|\mathbf{x}^{(3)}\right)+p\left(C=2|\mathbf{x}^{(1)}\right)+p\left(C=2|\mathbf{x}^{(2)}\right)+p\left(C=2|\mathbf{x}^{(3)}\right)} = \frac{0.0121+0.4+0.9733}{(0.9879+0.6+0.0267)+(0.0121+0.4+0.9733)} = 0.4618$

Having the new priors and likelihoods we could go for another iteration. However, the exercise only asks for one.

---

b) Plot the points and sketch the clusters.

Solution:



c) Verify that after one iteration the probability of the data increased.

Solution:

The probability of the observed data $p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right)$ can be decomposed into the product of the probability of each point assuming indepent, identically distributed samples:

$$p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right) = p\left(\mathbf{x}^{(1)}\right) p\left(\mathbf{x}^{(2)}\right) p\left(\mathbf{x}^{(3)}\right)$$

So, we need to compute the probability of each point before and after the EM update.

By the law of total probability we have that:

$$p\left(\mathbf{x}^{(n)}\right) = p\left(\mathbf{x}^{(n)}, C=1\right) + p\left(\mathbf{x}^{(n)}, C=2\right) = p\left(C=1\right) p\left(\mathbf{x}^{(n)} \mid C=1\right) + p\left(C=2\right) p\left(\mathbf{x}^{(n)} \mid C=2\right)$$

From a) we have that before the iteration:

- For $\mathbf{x}^{(1)}$:
  - $p\left(\mathbf{x}^{(1)}, C=1\right) = 0.095$
  - $p\left(\mathbf{x}^{(1)}, C=2\right) = 0.0012$
  - $p\left(\mathbf{x}^{(1)}\right) = 0.095 + 0.0012 = 0.0962$
- For $\mathbf{x}^{(2)}$:
  - $p\left(\mathbf{x}^{(2)}, C=1\right) = 0.0129$
  - $p\left(\mathbf{x}^{(2)}, C=2\right) = 0.0086$
  - $p\left(\mathbf{x}^{(2)}\right) = 0.0129 + 0.0086 = 0.0215$
- For $\mathbf{x}^{(3)}$:
  - $p\left(\mathbf{x}^{(3)}, C=1\right) = 0.0017$
  - $p\left(\mathbf{x}^{(3)}, C=2\right) = 0.0637$

- $p\left(\mathbf{x}^{(3)}\right) = 0.0017 + 0.0637 = 0.0654$

$$p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right) = p\left(\mathbf{x}^{(1)}\right) p\left(\mathbf{x}^{(2)}\right) p\left(\mathbf{x}^{(3)}\right) = 0.00014$$

For after the iteration:

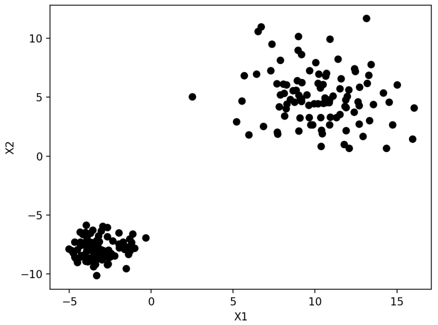
- For $\mathbf{x}^{(1)}$:
  - $p\left(\mathbf{x}^{(1)}, C=1\right) = 0.5382\mathcal{N}\left(\mathbf{x}^{(1)}; \mu^1 = \begin{pmatrix} 1.2237 \\ 1.9669 \end{pmatrix}, \mathbf{\Sigma}^1 = \begin{pmatrix} 0.94996 & 0.0405 \\ 0.0405 & 0.0651 \end{pmatrix}\right) = 0.2542$
  - $p\left(\mathbf{x}^{(1)}, C=2\right) = 0.4618\mathcal{N}\left(\mathbf{x}^{(1)}; \mu^2 = \begin{pmatrix} 0.0174 \\ 0.5949 \end{pmatrix}, \mathbf{\Sigma}^2 = \begin{pmatrix} 0.0345 & 0.0245 \\ 0.0245 & 0.8359 \end{pmatrix}\right) = 7.953 \times 10^{-26}$
  - $p\left(\mathbf{x}^{(1)}\right) = 0.2542 + 7.953 \times 10^{-26} = 0.2542$
- For $\mathbf{x}^{(2)}$:
  - $p\left(\mathbf{x}^{(2)}, C=1\right) = 0.5382\mathcal{N}\left(\mathbf{x}^{(2)}; \mu^1 = \begin{pmatrix} 1.2237 \\ 1.9669 \end{pmatrix}, \mathbf{\Sigma}^1 = \begin{pmatrix} 0.94996 & 0.0405 \\ 0.0405 & 0.0651 \end{pmatrix}\right) = 0.1499$
  - $p\left(\mathbf{x}^{(2)}, C=2\right) = 0.4618\mathcal{N}\left(\mathbf{x}^{(2)}; \mu^2 = \begin{pmatrix} 0.0174 \\ 0.5949 \end{pmatrix}, \mathbf{\Sigma}^2 = \begin{pmatrix} 0.0345 & 0.0245 \\ 0.0245 & 0.8359 \end{pmatrix}\right) = 0.128$
  - $p\left(\mathbf{x}^{(2)}\right) = 0.1499 + 0.128 = 0.2779$
- For $\mathbf{x}^{(3)}$:
  - $p\left(\mathbf{x}^{(3)}, C=1\right) = 0.5382\mathcal{N}\left(\mathbf{x}^{(3)}; \mu^1 = \begin{pmatrix} 1.2237 \\ 1.9669 \end{pmatrix}, \mathbf{\Sigma}^1 = \begin{pmatrix} 0.94996 & 0.0405 \\ 0.0405 & 0.0651 \end{pmatrix}\right) = 4.351 \times 10^{-14}$
  - $p\left(\mathbf{x}^{(3)}, C=2\right) = 0.4618\mathcal{N}\left(\mathbf{x}^{(3)}; \mu^2 = \begin{pmatrix} 0.0174 \\ 0.5949 \end{pmatrix}, \mathbf{\Sigma}^2 = \begin{pmatrix} 0.0345 & 0.0245 \\ 0.0245 & 0.8359 \end{pmatrix}\right) = 0.354$
  - $p\left(\mathbf{x}^{(3)}\right) = 4.351 \times 10^{-14} + 0.354 = 0.354$

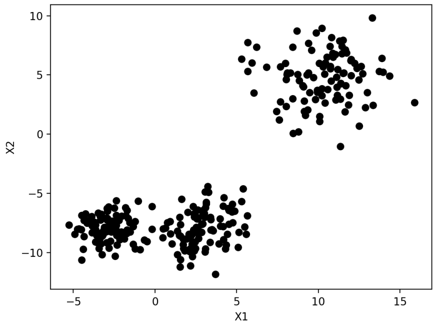
$$p\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right) = p\left(\mathbf{x}^{(1)}\right) p\left(\mathbf{x}^{(2)}\right) p\left(\mathbf{x}^{(3)}\right) = 0.025$$

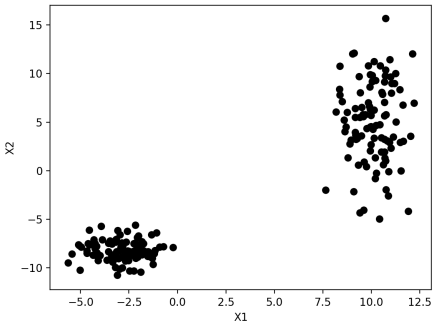
As we can see, after the iteration, the data is more probable which suggests that the model captures the data better.

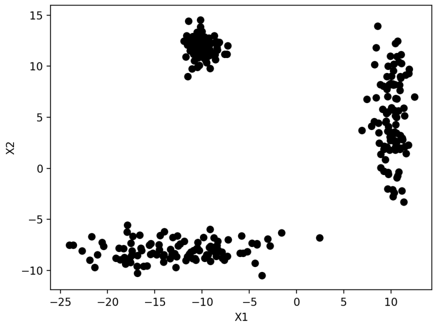
---

4) Consider the following four scenarios of plotted data sets:









a) For each scenario justify wether or not k-means would be suitable.

Solution:

In k-means all clusters are assumed to have a circle like shape. So, we can analyze these scenarios and see if that property holds.

i) $k = 2$ would provide two circle like clusters.

ii) $k = 3$ would provide three circle like clusters. However, if $k = 2$ we would get a bad fit on the points with negative x_2 coordinate.

iii) The right most cluster of points is clearly and ellipsis so k-means would not be able to capture this shape.

iv) We have two sets of points where the shape is oval so k-means would be a relatively bad fit to this data.

---

b) Assuming you apply EM clustering to model all scenarios what would the means and covariances look like? For simplicity, assume all covariance matrices are diagonal.

Solution:

After EM clustering, the means stay at the cluster centroids and the covariances describe the shape of the cluster.

i)

For the leftmost cluster the mean would be around $\begin{pmatrix} -3 \\ -8 \end{pmatrix}$ and the covariance must capture a tightly packed circle so the identity matrix could do it.

For the right most cluster the mean would be around $\begin{pmatrix} 10 \\ 5 \end{pmatrix}$ and the covariance must capture a spread circle so we could use a multiple of the identity matrix like $\begin{pmatrix} f & 0 \\ 0 & f \end{pmatrix}$ with $f > 1$.

ii)

Assuming $k = 3$, we have three circles with increasing levels of spread. Following the same logic from i), the values could be:

$\mu^1 = \begin{pmatrix} -3 \\ -8 \end{pmatrix}$ and $\Sigma^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$\mu^2 = \begin{pmatrix} 3 \\ -8 \end{pmatrix}$ and $\Sigma^2 = \begin{pmatrix} f & 0 \\ 0 & f \end{pmatrix}$ with $f > 1$.

$\mu^3 = \begin{pmatrix} 10 \\ 5 \end{pmatrix}$ and $\Sigma^3 = \begin{pmatrix} l & 0 \\ 0 & l \end{pmatrix}$ with $l > f$.

iii)

Assuming $k = 2$ we have one tightly packed circle and a vertically spread ellipsis.

Like before, the circle could be $\mu^1 = \begin{pmatrix} -3 \\ -8 \end{pmatrix}$ and $\Sigma^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

The ellipsis is centered at around $\mu^2 = \begin{pmatrix} 10 \\ 5 \end{pmatrix}$ and dimension x_1 as a smaller variance then x_2, so the covariance matrix could be $\Sigma^2 = \begin{pmatrix} f & 0 \\ 0 & k \end{pmatrix}$ with $k > f$.

iv)

Assuming $k = 3$ we have a tightly packed circle and two ellipsis.

For the circle again we can use the identiy as covariance and the mean is $\begin{pmatrix} -10 \\ 12 \end{pmatrix}$.

For the horizontally spread ellipsis we have mean $\begin{pmatrix} -13 \\ -8 \end{pmatrix}$ and a covariance $\Sigma^2 = \begin{pmatrix} f & 0 \\ 0 & k \end{pmatrix}$ where $f > k$.

For the vertically spread ellipsis we have mean $\begin{pmatrix} 10 \\ 5 \end{pmatrix}$ and a covariance $\Sigma^2 = \begin{pmatrix} f & 0 \\ 0 & k \end{pmatrix}$ where $k > f$.

3 Thinking Questions

a) Think about what measures would be desirable in a clustering. Think about distance between centroids and intra-cluster distances.

b) Think about initialization mechanisms and how they affect the final clustering.

c) Why do we need all those clustering indices?

d) Why do we need the covariance in EM? What is the difference between k-means and EM in terms of cluster shapes we can capture.

e) Is k-means really a kind of EM clustering?