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.

Alright, this ain't rocket science either. For each point, find the closest cluster.

(1) For the first point, for example.

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

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

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

And for the rest:

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

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

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

(2)

Update each centroid as a the mean point of its cluster's points.

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

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

Verify convergence: did any centroid change? If not, convergence is reached. Otherwise, do another iteration.

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

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

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

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

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

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

No centroid change, so the alogrithm converged.

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

(From another exercise)

Solution:

For $k = 2$:

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

Just have to add up the distance of each point to its cluster's centroid.

In the case of k-means, because of the way hard clustering works, this value will be 1 for the centroid of the cluster the point in question belongs to, and 0 otherwise.

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.

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

So, for $k = 2$:

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

Since we're computing these values only for comparison purposes, this term is not relevant and can be excluded from the calculations.

For $k = 3$:

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

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

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:

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

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

Also, initialize the priors as follows:

$$p(C = 1) = 0.6$$

$$p(C = 2) = 0.4$$

a) Perform one expectation maximization iteration.

This takes a little more time.

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

Similar to k-means, but all points have a probability of belonging to each of the clusters. This is called soft clustering.

For each point, we compare the posterior yielded by each cluster.
Starting with the first point:

- For $\mathbf{x}^{(1)}$:
  - For cluster $C = 1$:
    - Prior: $p(C = 1) = 0.6$
    - Likelihood: $p(\mathbf{x}^{(1)} \mid C = 1) = \frac{1}{2\pi}\frac{1}{det(\Sigma^1)} \exp\left(-\frac{1}{2}(\mathbf{x}^{(1)} - \mu^1)^T (\Sigma^1)^{-1} (\mathbf{x}^{(1)} - \mu^1)\right) =$
      $\frac{1}{2\pi}\frac{1}{1} \exp\left(-\frac{1}{2}\left(\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}2\\2\end{pmatrix}\right)^T \begin{pmatrix}1&0\\0&1\end{pmatrix} \left(\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}2\\2\end{pmatrix}\right)\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix}0\\0\end{pmatrix}^T \begin{pmatrix}1&0\\0&1\end{pmatrix} \begin{pmatrix}0\\0\end{pmatrix}\right) =$
      $\frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix}0&0\end{pmatrix} \begin{pmatrix}1&0\\0&1\end{pmatrix} \begin{pmatrix}0\\0\end{pmatrix}\right) = \frac{1}{2\pi} \exp(0) = \frac{1}{2\pi}$
    - Joint Probability: $p(C = 1, \mathbf{x}^{(1)}) = p(C = 1)\, p(\mathbf{x}^{(1)} \mid C = 1) = 0.6 \times \frac{1}{2\pi} = 0.095$
  - For cluster $C = 2$:
    - Prior: $p(C = 2) = 0.4$
    - Likelihood: $p(\mathbf{x}^{(1)} \mid C = 2) = \frac{1}{2\pi}\frac{1}{det(\Sigma^2)} \exp\left(-\frac{1}{2}(\mathbf{x}^{(1)} - \mu^2)^T (\Sigma^2)^{-1} (\mathbf{x}^{(1)} - \mu^2)\right) =$
      $\frac{1}{2\pi}\frac{1}{1} \exp\left(-\frac{1}{2}\left(\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}0\\0\end{pmatrix}\right)^T \begin{pmatrix}1&0\\0&1\end{pmatrix} \left(\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}0\\0\end{pmatrix}\right)\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix}2\\2\end{pmatrix}^T \begin{pmatrix}1&0\\0&1\end{pmatrix} \begin{pmatrix}2\\2\end{pmatrix}\right) =$
      $\frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix}2&2\end{pmatrix} \begin{pmatrix}1&0\\0&1\end{pmatrix} \begin{pmatrix}2\\2\end{pmatrix}\right) = \frac{1}{2\pi} \exp(-4) = 0.003$
    - Joint Probability: $p(C = 2, \mathbf{x}^{(1)}) = p(C = 2)\, p(\mathbf{x}^{(1)} \mid C = 2) = 0.4 \times 0.003 = 0.0012$

- So, we can compute the normalized posteriors for each cluster:
  - $C = 1$: $p(C = 1 \mid \mathbf{x}^{(1)}) = \frac{p(C=1,\mathbf{x}^{(1)})}{p(C=1,\mathbf{x}^{(1)}) + p(C=2,\mathbf{x}^{(1)})} = \frac{0.095}{0.095+0.0012} = 0.9879$
  - $C = 2$: $p(C = 2 \mid \mathbf{x}^{(1)}) = \frac{p(C=2,\mathbf{x}^{(1)})}{p(C=1,\mathbf{x}^{(1)}) + p(C=2,\mathbf{x}^{(1)})} = \frac{0.0012}{0.095+0.0012} = 0.0121$

Same scheiße for x2 and x3.

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

- For $\mathbf{x}^{(3)}$:
  - For cluster $C = 1$:
    - Prior: $p(C = 1) = 0.6$
    - Likelihood: $p\left(\mathbf{x}^{(3)} \mid C = 1\right) = \frac{1}{2\pi}\frac{1}{det(\Sigma^1)}\exp\left(-\frac{1}{2}\left(\mathbf{x}^{(3)} - \mu^1\right)^T\left(\Sigma^1\right)^{-1}\left(\mathbf{x}^{(3)} - \mu^1\right)\right) =$
    $\frac{1}{2\pi}\frac{1}{1}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}2\\2\end{pmatrix}\right)^T\begin{pmatrix}1&0\\0&1\end{pmatrix}\left(\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}2\\2\end{pmatrix}\right)\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}-2\\-2\end{pmatrix}^T\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix}\right)$
    $\frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}-2&-2\end{pmatrix}\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix}\right) = \frac{1}{2\pi}\exp(-4) = 0.003$
    - Joint Probability: $p\left(C = 1, \mathbf{x}^{(3)}\right) = p(C = 1)\,p\left(\mathbf{x}^{(3)} \mid C = 1\right) = 0.6\times 0.003 = 0.0017$
  - For cluster $C = 2$:
    - Prior: $p(C = 2) = 0.4$
    - Likelihood: $p\left(\mathbf{x}^{(3)} \mid C = 2\right) = \frac{1}{2\pi}\frac{1}{det(\Sigma^2)}\exp\left(-\frac{1}{2}\left(\mathbf{x}^{(3)} - \mu^2\right)^T\left(\Sigma^2\right)^{-1}\left(\mathbf{x}^{(3)} - \mu^2\right)\right) =$
    $\frac{1}{2\pi}\frac{1}{1}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}0\\0\end{pmatrix}\right)^T\begin{pmatrix}1&0\\0&1\end{pmatrix}\left(\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}0\\0\end{pmatrix}\right)\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}0\\0\end{pmatrix}^T\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}\right) =$
    $\frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{pmatrix}0&0\end{pmatrix}\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}\right) = \frac{1}{2\pi}\exp(0) = \frac{1}{2\pi}$
    - Joint Probability: $p\left(C = 2, \mathbf{x}^{(3)}\right) = p(C = 2)\,p\left(\mathbf{x}^{(3)} \mid C = 2\right) = 0.4\times \frac{1}{2\pi} = 0.0637$
  - 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

Now we need to update the clusters.

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:

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

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

Literally a weighted average with the posteriors.

- $\Sigma_{11}^1 = \frac{0.9879(2-1.2237)(2-1.2237)+0.6(0-1.2237)(0-1.2237)+0.0267(0-1.2237)(0-1.2237)}{0.9879+0.6+0.0267} =$ 0.94996

And each cell of the new covariance matrix.

- $\Sigma_{12}^1 = \Sigma_{21}^1 = \frac{0.9879(2-1.2237)(2-1.9669)+0.6(0-1.2237)(2-1.9669)+0.0267(0-1.2237)(0-1.9669)}{0.9879+0.6+0.0267} =$ 0.0405
- $\Sigma_{22}^1 = \frac{0.9879(2-1.9669)(2-1.9669)+0.6(2-1.9669)(2-1.9669)+0.0267(0-1.9669)(0-1.9669)}{0.9879+0.6+0.0267} =$ 0.0651
- $\Sigma^1 = \begin{pmatrix}0.94996 & 0.0405\\0.0405 & 0.0651\end{pmatrix}$

- So, the new likelihood is: $p\left(\mathbf{x} \mid C = 1\right) = \mathcal{N}\left(\mu^1 = \begin{pmatrix}1.2237\\1.9669\end{pmatrix}, \mathbf{\Sigma}^1 = \begin{pmatrix}0.94996 & 0.0405\\0.0405 & 0.0651\end{pmatrix}\right)$

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

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

Basically compare posteriors from before and after the iteration.

Solution:

The probability of the observed data $p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)})$ 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:

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

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

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

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

Another possibility:

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

This will change how we get the priors and likelihoods.

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:

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

We will use this instead of the gaussian distribution for the likelihoods.

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

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

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

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

And for the likelihoods:

	$p(x_1 = 1 \mid C = c)$	$p(x_2 = 1 \mid C = c)$	$p(x_3 = 1 \mid C = c)$	$p(x_4 = 1 \mid C = c)$
$c = 1$	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

Which means:

	$p(x_1 = 0 \mid C = c)$	$p(x_2 = 0 \mid C = c)$	$p(x_3 = 0 \mid C = c)$	$p(x_4 = 0 \mid C = c)$
$c = 1$	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

Example for x4

- For $\mathbf{x}^{(4)}$:
  - For cluster $C = 1$:
    - Prior: $p(C = 1) = \frac{1}{3}$
    - Likelihood:
      $p(\mathbf{x}^{(4)} \mid C = 1) =$
      $= p\left(x_1^{(4)} = 0 \mid C = 1\right) p\left(x_2^{(4)} = 0 \mid C = 1\right) p\left(x_3^{(4)} = 1 \mid C = 1\right) p\left(x_4^{(4)} = 0 \mid C = 1\right) =$
      $0.2 \times 0.5 \times 0.1 \times 0.9 = 0.009$
    - 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:
    - $C = 1$: $p\left(C = 1 \mid \mathbf{x}^{(4)}\right) = \frac{p\left(C=1, \mathbf{x}^{(4)}\right)}{p\left(C=1, \mathbf{x}^{(4)}\right) + p\left(C=2, \mathbf{x}^{(4)}\right) + p\left(C=3, \mathbf{x}^{(4)}\right)} = \frac{0.003}{0.003+0.012+0.1944} = 0.0143$
    - $C = 2$: $p\left(C = 2 \mid \mathbf{x}^{(4)}\right) = \frac{p\left(C=2, \mathbf{x}^{(4)}\right)}{p\left(C=1, \mathbf{x}^{(4)}\right) + p\left(C=2, \mathbf{x}^{(4)}\right) + p\left(C=3, \mathbf{x}^{(4)}\right)} = \frac{0.012}{0.003+0.012+0.1944} = 0.0573$
    - $C = 3$: $p\left(C = 3 \mid \mathbf{x}^{(4)}\right) = \frac{p\left(C=3, \mathbf{x}^{(4)}\right)}{p\left(C=1, \mathbf{x}^{(4)}\right) + p\left(C=2, \mathbf{x}^{(4)}\right) + p\left(C=3, \mathbf{x}^{(4)}\right)} = \frac{0.1944}{0.003+0.012+0.1944} = 0.9284$

Then, update likelihoods and prior.

- For $C = 1$:
  - For the likelihood:
    - $p(x_1 = 1 \mid C = 1) = \frac{0.961 \times 1 + 0.006 \times 0 + 0.0397 \times 0 + 0.0143 \times 0 + 0.9795 \times 1}{0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795} = 0.97$
    - $p(x_2 = 1 \mid C = 1) = \frac{0.961 \times 0 + 0.006 \times 1 + 0.0397 \times 1 + 0.0143 \times 0 + 0.9795 \times 1}{0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795} = 0.51$
    - $p(x_3 = 1 \mid C = 1) = \frac{0.961 \times 0 + 0.006 \times 1 + 0.0397 \times 0 + 0.0143 \times 1 + 0.9795 \times 0}{0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795} = 0.01$
    - $p(x_4 = 1 \mid C = 1) = \frac{0.961 \times 0 + 0.006 \times 1 + 0.0397 \times 1 + 0.0143 \times 0 + 0.9795 \times 0}{0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795} = 0.02$
  - For the prior:
    $p(C = 1) =$
    $= \frac{0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795}{(0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795) + (0.018 + 0.894 + 0.9524 + 0.0573 + 0.0181) + (0.021 + 0.1 + 0.0079 + 0.9284 + 0.0024)} =$
    0.4

Example for C1.