Aprendizagem 2023

Lab 9-10: Clustering

Practical exercises

1. Consider the following training data without labels:

	y_1	y_2
$\mathbf{x}_1$	0	0
$\mathbf{x}_2$	1	0
$\mathbf{x}_3$	0	2
$\mathbf{x}_4$	2	2

and the initialization centroids: $\mu_1 = [2\ 0]^T$ and $\mu_2 = [2\ 1]^T$

a) Apply the *k*-means until convergence

Step 1: Assign points to clusters

Let us start with $\mathbf{x}_1$:

$$\|\mathbf{x}_1 - \boldsymbol{u}_1\|_2{}^2 = \left\|\binom{0}{0} - \binom{2}{0}\right\|_2{}^2 = \left\|\binom{-2}{0}\right\|_2{}^2 = 4$$

$$\|\mathbf{x}_1 - \boldsymbol{u}_2\|_2{}^2 = \left\|\binom{0}{0} - \binom{2}{1}\right\|_2{}^2 = \left\|\binom{-2}{-1}\right\|_2{}^2 = 5$$

$$\arg\min_{c\in\{1,2\}} \|\mathbf{x}_1 - \boldsymbol{u}_c\|_2{}^2 = \arg\min_{c\in\{1,2\}} \{4,5\} = c_1$$

Moving to remaining observations:

$$\arg\min_{c\in\{1,2\}} \|\mathbf{x}_2 - \boldsymbol{u}_c\|_2{}^2 = \arg\min_{c\in\{1,2\}} \{1,2\} = c_1$$

$$\arg\min_{c\in\{1,2\}} \|\mathbf{x}_3 - \boldsymbol{u}_c\|_2{}^2 = \arg\min_{c\in\{1,2\}} \{8,5\} = c_2$$

$$\arg\min_{c\in\{1,2\}} \|\mathbf{x}_4 - \boldsymbol{u}_c\|_2{}^2 = \arg\min_{c\in\{1,2\}} \{4,1\} = c_2$$

Step 2: Adjust centroids

$$\boldsymbol{u}_1 = \frac{\mathbf{x}_1 + \mathbf{x}_2}{2} = \binom{1/2}{0}, \qquad \boldsymbol{u}_2 = \binom{1}{2}$$

Step 3: Verify convergence

$$\|\boldsymbol{u}_1{}^{old} - \boldsymbol{u}_1\|_2{}^2 = \left\|\binom{2}{0} - \binom{1/2}{0}\right\|_2{}^2 = \frac{9}{4}, \qquad \|\boldsymbol{u}_2{}^{old} - \boldsymbol{u}_2\|_2{}^2 = 2$$

Centroids changed. We need to do *another iteration*.

Step 1: Assign points to clusters.

$$\arg\min_{c\in\{1,2\}} \|\mathbf{x}_1 - \boldsymbol{u}_c\|_2{}^2 = c_1, \arg\min_{c\in\{1,2\}} \|\mathbf{x}_2 - \boldsymbol{u}_c\|_2{}^2 = c_1, \mathbf{arg}\min_{c\in\{1,2\}} \|\mathbf{x}_3 - \boldsymbol{u}_c\|_2{}^2 = c_2, \mathbf{arg}\min_{c\in\{1,2\}} \|\mathbf{x}_3 - \boldsymbol{u}_c\|_2{}^2 = c_2$$

Step 2: Adjust centroids

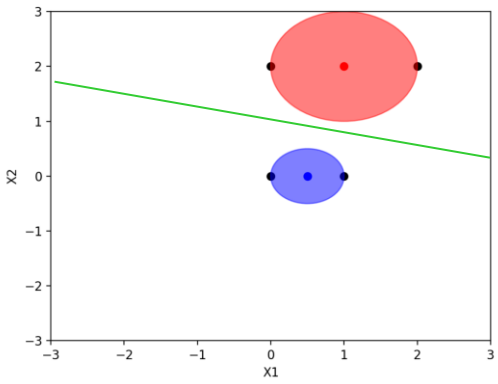
$$\boldsymbol{u}_1 = \binom{1/2}{0}, \qquad \boldsymbol{u}_2 = \binom{1}{2}$$

Step 3: Verify convergence

$$\|\boldsymbol{u}_1{}^{old} - \boldsymbol{u}_1\|_2{}^2 = 0, \qquad \|\boldsymbol{u}_2{}^{old} - \boldsymbol{u}_2\|_2{}^2 = 0$$

No centroid change, so the algorithm converged.

b) Plot the data points and draw the clusters



c) Compute the silhouette of observation $\mathbf{x}_1$, cluster c_1 and overall solution

Preserving the Euclidean distance assumption, let us compute the silhouette of $\boldsymbol{c_1}$:

$$s(\mathbf{x_1}) = 1 - \frac{a(\mathbf{x_1})}{b(\mathbf{x_1})} = 1 - \frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2}{\frac{1}{2}(\|\mathbf{x}_1 - \mathbf{x}_3\|_2 + \|\mathbf{x}_1 - \mathbf{x}_4\|_2)} = 1 - \frac{1}{2.4} = 0.58(3)$$

$$s(\mathbf{x_2}) = 1 - \frac{\|\mathbf{x}_2 - \mathbf{x}_1\|_2}{\frac{1}{2}(\|\mathbf{x}_2 - \mathbf{x}_3\|_2 + \|\mathbf{x}_2 - \mathbf{x}_4\|_2)} = 1 - \frac{1}{\sqrt{5}} = 0.553$$

$$s(\mathbf{c_1}) = \frac{s(\mathbf{x_1}) + s(\mathbf{x_2})}{2} = \frac{0.58(3) + 0.553}{2} = 0.568$$

The silhouette of a solution is the average of cluster silhouettes.

$$s(\mathbf{x_3}) = 0.052, \qquad s(\mathbf{x_4}) = 0.225, \qquad s(\mathbf{c_2}) = 0.133$$

$$silhouette(C) = \frac{s(\mathbf{c_1}) + s(\mathbf{c_2})}{2} = \frac{0.568 + 0.133}{2} = 0.35$$

Silhouette level is moderate, thus there is some evidence for clusters to be cohesive and well-separated.

d) Knowing $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_4$ to be annotated as positive and $\mathbf{x}_3$ as negative (ground truth). Compute the purity of *k*-means against the given ground truth.

$$purity(C, L) = \frac{1}{n}\sum_{k=1}^{K} \max_j \left(|\boldsymbol{c}_k \cap \boldsymbol{l}_j|\right) = \frac{1}{4}(1 + 2) = \frac{3}{4}$$

2. Consider the following data:

	y_1	y_2	y_3
$\mathbf{x}_1$	1	0	0
$\mathbf{x}_2$	8	8	4
$\mathbf{x}_3$	3	3	0
$\mathbf{x}_4$	0	0	1
$\mathbf{x}_5$	0	1	0
$\mathbf{x}_6$	3	2	1

and let the initial *k* centroids be the first *k* data points

a) Apply *k*-means with *k*=2 and *k*=3

Starting with $K = 2$, let us advance with the first iteration.

$$\boldsymbol{u}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{u}_2 = \begin{pmatrix} 8 \\ 8 \\ 4 \end{pmatrix}$$

Step 1: Assign points to clusters

$$\mathbf{arg}\min_{c\in\{1,2\}}\|\mathbf{x}_1-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2\}}\{0,129\}=c_1,$$

$$\mathbf{arg}\min_{c\in\{1,2\}}\|\mathbf{x}_2-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2\}}\{129,0\}=c_2$$

$$\mathbf{arg}\min_{c\in\{1,2\}}\|\mathbf{x}_3-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2\}}\{13,66\}=c_1$$

$$\mathbf{arg}\min_{c\in\{1,2\}}\|\mathbf{x}_4-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2\}}\{2,137\}=c_1$$

$$\mathbf{arg}\min_{c\in\{1,2\}}\|\mathbf{x}_5-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2\}}\{2,129\}=c_1$$

$$\mathbf{arg}\min_{c\in\{1,2\}}\|\mathbf{x}_6-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2\}}\{9,70\}=c_1$$

Step 2: Compute new centroids by averaging the cluster elements

$$\boldsymbol{u}_1=\frac{1}{5}\left(\begin{pmatrix}1\\0\\0\end{pmatrix}+\begin{pmatrix}3\\3\\0\end{pmatrix}+\begin{pmatrix}0\\0\\1\end{pmatrix}+\begin{pmatrix}0\\1\\0\end{pmatrix}+\begin{pmatrix}3\\2\\1\end{pmatrix}\right)=\begin{pmatrix}7/5\\6/5\\2/5\end{pmatrix},\qquad \boldsymbol{u}_2=\begin{pmatrix}8\\8\\4\end{pmatrix},$$

Step 3: Verify convergence

$$\|\boldsymbol{u}_1{}^{old}-\boldsymbol{u}_1\|_2{}^2=\left\|\begin{pmatrix}1\\1\\0\end{pmatrix}-\begin{pmatrix}7/5\\6/5\\2/5\end{pmatrix}\right\|_2{}^2=1.33,\qquad \|\boldsymbol{u}_2{}^{old}-\boldsymbol{u}_2\|_2{}^2=\left\|\begin{pmatrix}8\\8\\4\end{pmatrix}-\begin{pmatrix}8\\8\\4\end{pmatrix}\right\|_2{}^2=0$$

Centroids changed. We need to do another iteration.

Step 1: Assign points to clusters: all observations assign to c_1 with the exception of $\mathbf{x}_2$

Step 2: Compute new centroids by averaging the cluster elements: $\boldsymbol{u}_1=\begin{pmatrix}7/5\\6/5\\2/5\end{pmatrix},\boldsymbol{u}_2=\begin{pmatrix}8\\8\\4\end{pmatrix}$

Step 3: Verify convergence, $\|\boldsymbol{u}_1{}^{old}-\boldsymbol{u}_1\|_2{}^2=\|\boldsymbol{u}_2{}^{old}-\boldsymbol{u}_2\|_2{}^2=0$

No centroid change, so the algorithm converged.

Starting with $K=3$, let us advance with the first iteration.

$$\boldsymbol{u}_1=\begin{pmatrix}1\\0\\0\end{pmatrix},\qquad \boldsymbol{u}_2=\begin{pmatrix}8\\8\\4\end{pmatrix},\qquad \boldsymbol{u}_3=\begin{pmatrix}3\\3\\0\end{pmatrix}$$

Step 1: Assign points to clusters

$$\mathbf{arg}\min_{c\in\{1,2,3\}}\|\mathbf{x}_1-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2,3\}}\{0,129,13\}=c_1$$

$$\mathbf{arg}\min_{c\in\{1,2,3\}}\|\mathbf{x}_2-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2,3\}}\{129,0,66\}=c_2$$

$$\mathbf{arg}\min_{c\in\{1,2,3\}}\|\mathbf{x}_3-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2,3\}}\{13,66,0\}=c_3$$

$$\mathbf{arg}\min_{c\in\{1,2,3\}}\|\mathbf{x}_4-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2,3\}}\{2,137,19\}=c_1$$

$$\mathbf{arg}\min_{c\in\{1,2,3\}}\|\mathbf{x}_5-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2,3\}}\{2,129,13\}=c_1$$

$$\mathbf{arg}\min_{c\in\{1,2,3\}}\|\mathbf{x}_6-\boldsymbol{u}_c\|_2{}^2=\mathbf{arg}\min_{c\in\{1,2,3\}}\{9,70,2\}=c_3$$

Step 2: Compute new centroids: $\boldsymbol{u}_1=\begin{pmatrix}1/3\\1/3\\1/3\end{pmatrix},\ \boldsymbol{u}_2=\begin{pmatrix}8\\8\\4\end{pmatrix},\ \boldsymbol{u}_3=\begin{pmatrix}3\\5/2\\1/2\end{pmatrix}$

Step 3: Verify convergence

$$\|\boldsymbol{u}_1{}^{old}-\boldsymbol{u}_1\|_2{}^2=0.67,\ \|\boldsymbol{u}_2{}^{old}-\boldsymbol{u}_2\|_2{}^2=0,\|\boldsymbol{u}_3{}^{old}-\boldsymbol{u}_3\|_2{}^2=0.5$$

Centroids changed. We need to do another iteration.

Step 1: Assign points to clusters: $\mathbf{x}_1, \mathbf{x}_4$ *and* $\mathbf{x}_5$ assign to c_1, $\mathbf{x}_2$ assign to c_2, $\mathbf{x}_3$and $\mathbf{x}_6$ assign to c_3

Step 2: Compute new centroids, $\boldsymbol{u}_1 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$, $\boldsymbol{u}_2 = \begin{pmatrix} 8 \\ 8 \\ 4 \end{pmatrix}$, $\boldsymbol{u}_3 = \begin{pmatrix} 3 \\ 5/2 \\ 1/2 \end{pmatrix}$

Step 3: Verify convergence, $\|\boldsymbol{u}_1{}^{old} - \boldsymbol{u}_1\|_2{}^2 = \|\boldsymbol{u}_2{}^{old} - \boldsymbol{u}_2\|_2{}^2 = \|\boldsymbol{u}_3{}^{old} - \boldsymbol{u}_3\|_2{}^2 = 0$

No centroid change, so the algorithm converged.

b) Which *k* provides a better clustering in terms of cohesion (sum of intra-cluster distance)?

For $K = 2$:

$$D_{intra} = \sum_{k=1}^{2} \sum_{\mathbf{x}_i \in c_k} \|\mathbf{x}_i - \boldsymbol{u}_k\|_2{}^2$$

$$= \|\mathbf{x}_1 - \boldsymbol{u}_1\|_2{}^2 + \|\mathbf{x}_2 - \boldsymbol{u}_2\|_2{}^2 + \|\mathbf{x}_3 - \boldsymbol{u}_1\|_2{}^2 + \|\mathbf{x}_4 - \boldsymbol{u}_1\|_2{}^2 + \|\mathbf{x}_5 - \boldsymbol{u}_1\|_2{}^2 + \|\mathbf{x}_i - \boldsymbol{u}_1\|_2{}^2$$

$$= 1.76 + 0 + 5.96 + 3.76 + 2.16 + 3.56 = 17.2$$

For $K = 3$:

$$D_{intra} = \sum_{k=1}^{3} \sum_{\mathbf{x}_i \in c_k} \|\mathbf{x}_i - \boldsymbol{u}_k\|_2{}^2 = 3.0$$

$K = 3$ has more tightly packed clusters, which is expected for an increasing K.

c) Which *k* provides a better clustering in terms of separation (mean inter-cluster centroid distance)?

For $K = 2$:

$$D_{inter} = \frac{1}{K^2} \sum_{c_i} \sum_{c_j} \|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2{}^2 = \|\boldsymbol{u}_1 - \boldsymbol{u}_1\|_2{}^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2{}^2 + \|\boldsymbol{u}_2 - \boldsymbol{u}_1\|_2{}^2 + \|\boldsymbol{u}_2 - \boldsymbol{u}_2\|_2{}^2$$

$$= \frac{1}{4}(0 + 102.76 + 0 + 102.76) = 51.39$$

For $K = 3$:

$$D_{inter} = \frac{1}{K^2} \sum_{c_i} \sum_{c_j} \|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2{}^2 = 46.67$$

3. Considering the following data points:

$$\{ x_1 = (4), x_2 = (0), x_3 = (1) \}$$

and a mixture of two normal distributions with the following initialization of likelihoods:

$$P(x \mid k = 1) = N(u_1 = 1, \sigma_1 = 1)$$

$$P(x \mid k = 2) = N(u_2 = 0, \sigma_2 = 1)$$

and priors: $p\ (k = 1) = 0.5$ and $p\ (k = 2) = 0.5$

Plot the clusters after one iteration of the EM algorithm.

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

For x_1:

- for cluster $c = 1$:
  - ∗prior: $p(c = 1) = 0.5$
  - ∗likelihood: $p(x_1 \mid c = 1) = \frac{1}{\sqrt{2\pi\sigma_1{}^2}} exp\left(-\frac{1}{2}\frac{(x_1 - \mu_1)^2}{\sigma_1{}^2}\right) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{(4-1)^2}{1}\right) = 0.0044$
  - ∗joint probability: $p(c = 1, x_1) = p(x_1 \mid c = 1)p(c = 1) = 0.5 \times 0.0044 = 0.0022$

• for cluster $c = 2$:

∗ prior: $p(c = 2) = 0.5$, likelihood: $p(x_1 | c = 2) = 0.000134$, joint prob: $p(c = 2, x_1) = 0.000067$

• Normalized posteriors: $p(c = 1 | x_1) = 0.9707, p(c = 2 | x_1) = 0.0293$

For x_2: normalized posteriors: $p(c = 1 | x_2) = 0.38, p(c = 2 | x_2) = 0.62$

For x_3: normalized posteriors: $p(c = 1 | x_3) = 0.62, p(c = 2 | x_3) = 0.38$

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

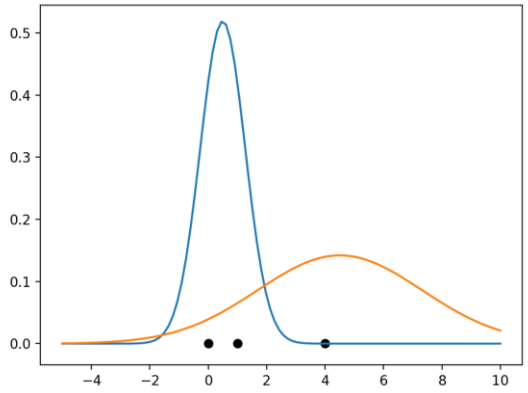
$$\mu_c = \frac{\sum_{i=1}^{3} p(c \mid x_i)\, x_i}{\sum_{i=1}^{3} p(c \mid x_i)}, \qquad \sigma_c = \sqrt{\frac{\sum_{i=1}^{3} p(c \mid x_i)\, (x_i - \mu_c)}{\sum_{i=1}^{3} p(c \mid x_i)}}$$

For the priors we perform a weighted mean of the posteriors, $p(c = k) = \frac{\sum_{i=1}^{3} p(c=k \mid x_i)}{\sum_{j=1}^{2} \sum_{i=1}^{3} p(c=j \mid x_i)}$

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

– for $C = 2$: $\mu_2 = \frac{4 \times 0.0293 + 0 \times 0.62 + 1 \times 0.38}{0.0293 + 0.62 + 0.38} = 0.495$, $\sigma_2 = \sqrt{\frac{(4-0.495)^2 \times 0.0293 + (0-0.495)^2 \times 0.62 + (1-0.495)^2 \times 0.38}{0.0293 + 0.62 + 0.38}} = 0.769$

– for $C = 1$: $\mu_1 = 4.5$, $\sigma_1 = 2.81$



4. Consider the following Boolean data:

	y_1	y_2	y_3	y_4
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	0	1	1	1
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	0	0	1	1
$\mathbf{x}_5$	1	1	0	0

Assuming the presence of 3 clusters, variables to be conditionally independent, and the following likelihoods:

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

a) Perform one expectation maximization iteration.

The question tells us that all features are conditionally independent given the cluster. So, we can write the likelihoods as follows: $p(\mathbf{x} \mid c = 1) = p(x_1 \mid c = 1)\, p(x_2 \mid c = 1)\, p(x_3 \mid c = 1)\, p(x_4 \mid c = 1)$

Furthermore, the question tells us that all distributions are initialized uniformly.

So, we will have priors: $p(c = 1) = \frac{1}{3}$, $p(c = 2) = \frac{1}{3}$, $p(c = 3) = \frac{1}{3}$

We also know:

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

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: $p(c = 1) = \frac{1}{3}$
  - likelihood: $p(\mathbf{x_1} \mid c = 1) = p(a_{11} = 1 \mid c = 1)p(a_{12} = 0 \mid c = 1)p(a_{13} = 0 \mid c = 1)p(a_{14} = 0 \mid c = 1) = 0.8 \times 0.5 \times 0.9 \times 0.9 = 0.324$
  - joint probability: $p(c = 1, \mathbf{x_1}) = p(\mathbf{x_1} \mid c = 1)p(c = 1) = \frac{1}{3}0.324 = 0.108$
- For cluster $c = 2$:
  - prior: $p(c = 2) = \frac{1}{3}$, likelihood: $p(\mathbf{x_1} \mid c = 2) = 0.006$, joint probability: $p(c = 2, \mathbf{x_1}) = 0.002$
- For cluster $c = 3$:
  - prior: $p(c = 3) = \frac{1}{3}$, likelihood: $p(\mathbf{x_1} \mid c = 3) = 0.0072$, joint probability: $p(c = 3, \mathbf{x_1}) = 0.0024$
- Normalized posteriors: $p(\mathbf{x_1} \mid c = 1) = 0.961, p(\mathbf{x_1} \mid c = 2) = 0.018,\ p(\mathbf{x_1} \mid c = 3) = 0.021$

For $\mathbf{x_2}$: normalized posteriors: $p(\mathbf{x_2} \mid c = 1) = 0.006, p(\mathbf{x_2} \mid c = 3) = 0.894, p(\mathbf{x_2} \mid c = 2) = 0.100$
For $\mathbf{x_3}$: normalized posteriors: $p(\mathbf{x_3} \mid c = 1) = 0.0397, p(\mathbf{x_3} \mid c = 3) = 0.9524, p(\mathbf{x_3} \mid c = 2) = 0.0079$
For $\mathbf{x_4}$: normalized posteriors: $p(\mathbf{x_4} \mid c = 1) = 0.0143, p(\mathbf{x_4} \mid c = 3) = 0.0573, p(\mathbf{x_4} \mid c = 2) = 0.9284$
For $\mathbf{x_5}$: normalized posteriors: $p(\mathbf{x_5} \mid c = 1) = 0.9795, p(\mathbf{x_5} \mid c = 3) = 0.0181, p(\mathbf{x_5} \mid c = 2) = 0.0024$

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:

$$p(\mathbf{x}_i \mid c) = \frac{\sum_{k=1}^{3} p(c = k \mid X)\, \mathbb{I}[\mathbf{x}_i]}{\sum_{k=1}^{3} p(c = k \mid X)}$$

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

$$p(c = k) = \frac{\sum_{i=1}^{5} p(c = k \mid \mathbf{x}_i)}{\sum_{j=1}^{3}\sum_{i=1}^{5} p(c = j \mid \mathbf{x}_i)}$$

For $c = 1$:

- likelihood:

$$p(x_1 = 1 \mid c = 1) = \frac{0.961 \times 1 + 0.006 \times 0 + 0.0397 \times 0 + 0.0143 \times 0 + 0.9795 \times 1}{0.961 + 0.006 + 0.0397 + 0.0143 + 0.9795} = 0.97$$

$$p(x_2 = 1 \mid c = 1) = 0.51, p(x_3 = 1 \mid c = 1) = 0.01, p(x_4 = 1 \mid c = 1) = 0.02$$

- prior: $p(c = 1) = \frac{0.961+0.006+0.0397+0.0143+0.9795}{(0.961+0.006+0.0397+0.0143+0.9795)+(0.018+0.894+0.9524+0.0573+0.0181)+(0.021+0.1+0.0079+0.9284+0.0024)} = 0.4$

For $c = 2$:

- likelihood:

$p(x_1 = 1 \mid c = 2) = 0.0186, p(x_2 = 1 \mid c = 2) = 0.9612,\ p(x_3 = 1 \mid c = 2) = 0.4904,\ \ p(x_4 = 1 \mid c = 2) = 0.9519$

- prior: $p(c = 2) = 0.39$

For $c = 3$:

- likelihood:

$p(x_1 = 1 \mid c = 3) = 0.0221, p(x_2 = 1 \mid c = 3) = 0.1041, p(x_3 = 1 \mid c = 3) = 0.9705, p(x_4 = 1 \mid c = 3) = 0.1018$

- prior: $p(c = 3) = 0.21$

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

Assuming independent, identically distributed observations:

$$p(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}) = p(\mathbf{x_1})p(\mathbf{x_2})p(\mathbf{x_3})p(\mathbf{x_4})p(\mathbf{x_5})$$

where

$$p(\mathbf{x_i}) = p(\mathbf{x_i}, c=1) + p(\mathbf{x_i}, c=2) + p(\mathbf{x_i}, c=3) = p(c=1)p(\mathbf{x_i} \mid c=1) + p(c=2)p(\mathbf{x_i} \mid c=2) + p(c=3)p(\mathbf{x_i} \mid c=3)$$

For $\mathbf{x}_1$:

$$p(\mathbf{x}_1|c=1) = p(a_{11}=1|c=1)p(a_{12}=0|c=1)p(a_{13}=0|c=1)p(a_{14}=0|c=1) = 0.324$$

$$p(\mathbf{x}_1|c=2) = 0.006, p(\mathbf{x}_1|c=3) = 0.072$$

$$p(\mathbf{x}_1) = \frac{1}{3}0.324 + \frac{1}{3}0.006 + \frac{1}{3}0.072 = 0.1124$$

For $\mathbf{x}_2$: $p(\mathbf{x_2}|\, c=1) = 0.001, p(\mathbf{x_2}|\, c=2) = 0.144, p(\mathbf{x_2}|\, c=3) = 0.0162, p(\mathbf{x_2}) = 0.0537$

For $\mathbf{x}_3$: $p(\mathbf{x_3}|\, c=1) = 0.009, p(\mathbf{x_3}|\, c=2) = 0.216, p(\mathbf{x_3}|\, c=3) = 0.0018, p(\mathbf{x_3}) = 0.0756$

For $\mathbf{x}_4$: $p(\mathbf{x_4}|\, c=1) = 0.009, p(\mathbf{x_4}|\, c=2) = 0.036, p(\mathbf{x_4}|\, c=3) = 0.5832, p(\mathbf{x_4}) = 0.2094$

For $\mathbf{x}_5$: $p(\mathbf{x_5}|\, c=1) = 0.324, p(\mathbf{x_5}|\, c=2) = 0.006, p(\mathbf{x_5}|\, c=3) = 0.0008, p(\mathbf{x_5}) = 0.1103$

As a result: $p(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}) = p(\mathbf{x_1})p(\mathbf{x_2})p(\mathbf{x_3})p(\mathbf{x_4})p(\mathbf{x_5}) = 1.054 \times 10^{-5}$

From (a) we have

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

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

For $\mathbf{x}_1$: $p(\mathbf{x_1}|\, c=1) = 0.46, p(\mathbf{x_1}|\, c=2) = 0.00002, p(\mathbf{x_1}|\, c=3) = 0.000486, p(\mathbf{x_1}) = 0.1846$

For $\mathbf{x}_2$: $p(\mathbf{x_2}|\, c=1) = 0.000003, p(\mathbf{x_2}|\, c=2) = 0.438, p(\mathbf{x_2}|\, c=3) = 0.0095, p(\mathbf{x_2}) = 0.1728$

For $\mathbf{x}_3$: $p(\mathbf{x_3}|\, c=1) = 0.0003, p(\mathbf{x_3}|\, c=2) = 0.456, p(\mathbf{x_3}|\, c=3) = 0.000294, p(\mathbf{x_3}) = 0.178$

For $\mathbf{x}_4$: $p(\mathbf{x_4}|\, c=1) = 0.000144, p(\mathbf{x_4}|\, c=2) = 0.00096, p(\mathbf{x_4}|\, c=3) = 0.77, p(\mathbf{x_4}) = 0.1621$

For $\mathbf{x}_5$: $p(\mathbf{x_5}|\, c=1) = 0.48, p(\mathbf{x_5}|\, c=2) = 0.00049, p(\mathbf{x_5}|\, c=3) = 0.000054, p(\mathbf{x_5}) = 0.1922$

As a result: $p(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}) = p(\mathbf{x_1})p(\mathbf{x_2})p(\mathbf{x_3})p(\mathbf{x_4})p(\mathbf{x_5}) = 0.00018$

As we can see, after the iteration, the data is more likely to be generated by the mixture, suggesting that the mixture model is better at describing the observations.

5. Consider the following data points:

	y_1	y_2
$\mathbf{x}_1$	2	2
$\mathbf{x}_2$	0	2
$\mathbf{x}_3$	0	0

and a mixture of two multivariate normal distributions with the following likelihoods' initialization:

$$P(\mathbf{x} \mid k=1) = N(u_1 = \begin{bmatrix}2\\2\end{bmatrix}, \sigma_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix})$$

$$P(\mathbf{x} \mid k=2) = N(u_2 = \begin{bmatrix}0\\0\end{bmatrix}, \sigma_2 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix})$$

and priors: $p\ (k=1) = 0.6$ and $p\ (k=2) = 0.4$

a) Perform one expectation maximization iteration

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

For $\mathbf{x}_1$:

- for cluster $c=1$: likelihood: $p(\mathbf{x_1}|\, c=1) = \frac{1}{2\pi}$, joint prob.: $p(c=1, \mathbf{x_1}) = p(\mathbf{x_1}|\, c=1)p(c=1) = 0.095$
- for cluster $c=2$: likelihood: $p(\mathbf{x_1}|\, c=2) = 0.003$, joint prob.: $p(c=2, \mathbf{x_1}) = p(\mathbf{x_1}|\, c=2)p(c=2) = 0.0012$
- normalized posteriors: $p(c=1 \mid \mathbf{x_1}) = 0.9879, p(c=2 \mid \mathbf{x_1}) = 0.0121$

For $\mathbf{x_2}$: normalized posteriors: $p(c=1 \mid \mathbf{x_2}) = 0.6, p(c=2 \mid \mathbf{x_2}) = 0.4$

For $\mathbf{x_3}$: normalized posteriors: $p(c=1 \mid \mathbf{x_3}) = 0.0267, p(c=2 \mid \mathbf{x_3}) = 0.9733$

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

$$\mu_c = \frac{\sum_{i=1}^{3} p(c \mid \mathbf{x}_i)\, \mathbf{x}_i}{\sum_{i=1}^{3} p(c \mid \mathbf{x}_i)}, \qquad \Sigma_c{}^{(i,j)} = \frac{\sum_{k=1}^{3} p(c \mid \mathbf{x}_k)\,(a_{ki} - \mu_{ci})\,(a_{kj} - \mu_{cj})}{\sum_{k=1}^{3} p(c \mid \mathbf{x}_k)}$$

For the priors we perform a weighted mean of the posteriors: $p(c=k) = \frac{\sum_{i=1}^{3} p(c=k \mid \mathrm{x}_i)}{\sum_{j=1}^{2}\sum_{i=1}^{3} p(c=j \mid \mathrm{x}_i)}$

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

– for c = 1: $p(\mathbf{x} \mid \mathrm{c}=1) = N\left(\mu_1 = \begin{pmatrix}1.2237\\1.9669\end{pmatrix}, \Sigma_1 = \begin{pmatrix}0.94996 & 0.0405\\0.0405 & 0.0651\end{pmatrix}\right)$

– for c = 2: $p(\mathbf{x} \mid \mathrm{c}=2) = N\left(\mu_2 = \begin{pmatrix}0.0174\\0.5949\end{pmatrix}, \Sigma_2 = \begin{pmatrix}0.0345 & 0.0245\\0.0245 & 0.8359\end{pmatrix}\right)$

Priors: $p(c=1) = \frac{0.9879+0.6+0.0267}{(0.9879+0.6+0.0267)+(0.0121+0.4+0.9733)} = 0.5382, p(c=2) = 0.4618$

b) How much the fitting probability increased?

Before:

$$p(\mathbf{x_1}) = p(\mathbf{x_1}, c=1) + p(\mathbf{x_1}, c=2) = 0.095 + 0.0012 = 0.0962$$
$$p(\mathbf{x_2}) = p(\mathbf{x_2}, c=1) + p(\mathbf{x_3}, c=2) = 0.0129 + 0.0086 = 0.0215$$
$$p(\mathbf{x_3}) = p(\mathbf{x_3}, c=1) + p(\mathbf{x_3}, c=2) = 0.0017 + 0.0637 = 0.0654$$

As a result: $p(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}) = p(\mathbf{x_1})p(\mathbf{x_2})p(\mathbf{x_3}) = 0.00014$

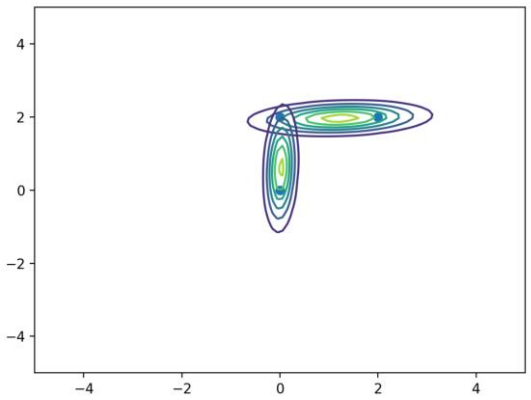
After:

$$p(\mathbf{x_1}, c=1) = 0.5382 \times N\left(\mu_1 = \begin{pmatrix}1.2237\\1.9669\end{pmatrix}, \Sigma_1 = \begin{pmatrix}0.94996 & 0.0405\\0.0405 & 0.0651\end{pmatrix}\right) = 0.2542, \;\; p(\mathbf{x_1}, c=2) = 7.953 \times 10^{-26}$$

$$p(\mathbf{x_1}) = 0.2542, \;\; p(\mathbf{x_2}) = 0.2779, \;\; p(\mathbf{x_3}) = 0.354$$

As a result: $p(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}) = p(\mathbf{x_1})p(\mathbf{x_2})p(\mathbf{x_3}) = 0.025$

c) Sketch the points and clusters



6. Consider the following dataset (Euclidean distance space):

	y_1	y_2
$\mathbf{x}_1$	2	10
$\mathbf{x}_2$	2	5
$\mathbf{x}_3$	8	4
$\mathbf{x}_4$	5	8
$\mathbf{x}_5$	7	5
$\mathbf{x}_6$	6	4
$\mathbf{x}_7$	1	2
$\mathbf{x}_8$	4	9

a) Assuming observations $\mathbf{x}_1$, $\mathbf{x}_4$ and $\mathbf{x}_7$ to be the initial seeds, identify the centroids after the first epoch using:

i. *k*-means

ii. *k*-median

Given $\bar{c}_1 = \mathbf{x}_1$, $\bar{c}_2 = \mathbf{x}_4$ and $\bar{c}_3 = \mathbf{x}_7$

Step 1: Assign points to clusters:

$$\mathbf{x_1}: argmin\{0.00,3.61,8.06\} = c_1, \quad \mathbf{x_2}: argmin\{5.00,4.24,3.16\} = c_3$$

$$\mathbf{x_3}: argmin\{8.49,5.00,7.28\} = c_2, \quad \mathbf{x_4}: argmin\{3.61,0.00,7.21\} = c_2$$

$$\mathbf{x_5}: argmin\{7.07,3.61,6.71\} = c_2, \quad \mathbf{x_6}: argmin\{7.21,4.12,5.39\} = c_2$$

$$\mathbf{x_7}: argmin\{8.06,7.21,0.00\} = c_3, \quad \mathbf{x_8}: argmin\{2.24,1.41,7.62\} = c_2$$

Step 2: Recompute centroids:

Using *k*-means:

$\bar{c}_1 = \mathbf{x}_1 = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$, $\bar{c}_2 = \frac{1}{5}(\mathbf{x_3} + \mathbf{x_4} + \mathbf{x_5} + \mathbf{x_6} + \mathbf{x_8}) = \frac{1}{5}\begin{pmatrix} 30 \\ 30 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$ and $\bar{c}_3 = \frac{1}{2}(\mathbf{x_2} + \mathbf{x_7}) = \begin{pmatrix} 1.5 \\ 3.5 \end{pmatrix}$

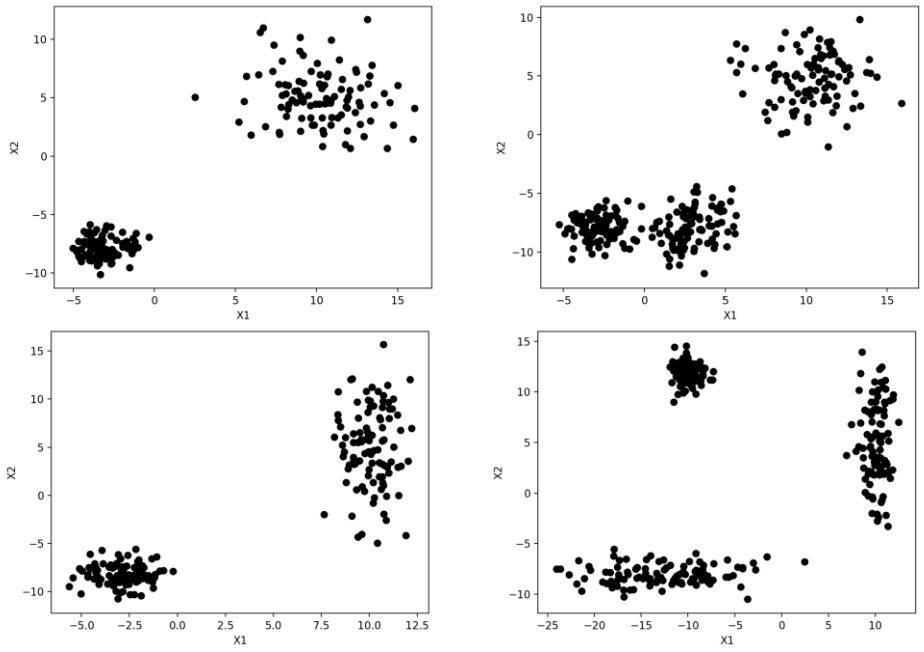
Using *k*-medians:

$\bar{c}_1 = \mathbf{x}_1 = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$, $\bar{c}_2 = \begin{pmatrix} \text{median}(8,5,7,6,4) \\ \text{median}(4,8,5,4,9) \end{pmatrix} = \begin{pmatrix} 6 \\ 5 \end{pmatrix}$ and, using midpoint rule, $\bar{c}_3 = \begin{pmatrix} 1.5 \\ 3.5 \end{pmatrix}$

b) When is median preferred over mean?

To handle outlier observations and, depending on the targeted ends, skewed variable distributions.

7. Consider the following four scenarios of plotted data sets:



a) For each scenario, justify whether *k*-means is suitable

In *k*-means all clusters are assumed to have a circle-like (globular) shape. So, we can assess whether this 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 y2 coordinate.

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

iv. two sets of points have an oval 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.

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

i. assuming k = 2, the mean of leftmost cluster would be around [-3 8] 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 [10 5] 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.

iii. assuming k = 2 we have one tightly packed circle and a vertically spread ellipsis.

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 [-12 1].
For the horizontally spread ellipsis we have mean [-13 8] and covariance $\Sigma = \begin{pmatrix} f & 0 \\ 0 & k \end{pmatrix}$ where f>k
For the vertically spread ellipsis we have mean [10 5] and a covariance $\Sigma = \begin{pmatrix} f & 0 \\ 0 & k \end{pmatrix}$ where k>f

c) When moving from numeric to ordinal data spaces, is Hamming distance proper to handle ordinal data with high cardinality?

Hamming distance between two (multivariate) observations is suggested for measuring distances in nominal data spaces, corresponding to the number of variables with different values. In ordinal data spaces, the Hamming distance is not sensitive to the degree of separation between two ordinal values (e.g. 'high' is more distant to 'low' than 'moderate'). This effect is heightened when the ordinal variables show high cardinality. In this context, Hamming is only appropriate when the rank differences per ordinal variable can be neglected.

Programming quest

Resources: *Clustering* notebook available at the course's webpage

8. Using the *iris* dataset (without the class variable)

a) Apply k-means with $k \in \{2,3,4,5,6,7,8,9,10\}$.
Choose the best k using the elbow method by plotting the SSE (inertia) per k

b) After selecting two informative features OR extracting two principal components `PCA(n_components=2).fit(X)`, visualize the produced clusters