



## Lab 9: Clustering

### Practical exercises

1. Consider the following training data without labels:

	y1	y2
$\mathbf{x}_1$	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  $\mu_2 = [2 \ 1]^T$

- Apply the  $k$ -means until convergence
- Plot the data points and draw the clusters
- Compute the silhouette of observation  $\mathbf{x}_1$ , cluster  $\mathbf{c}_1$  and overall solution
- 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 error classification rate (ECR) of  $k$ -means against the given ground truth.

2. Consider the following data:

	y1	y2	y3
$\mathbf{x}_1$	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

- Apply  $k$ -means with  $k = 2$  and  $k = 3$
  - Which  $k$  provides a better clustering in terms of cohesion (sum of intra-cluster distance)?
  - Which  $k$  provides a better clustering in terms of separation (mean inter-cluster centroid distance)?
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:

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

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

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

Plot the clusters after one iteration of the EM algorithm.

4. Consider the following Boolean data:

	y <sub>1</sub>	y <sub>2</sub>	y <sub>3</sub>	y <sub>4</sub>
<b>x</b> <sub>1</sub>	1	0	0	0
<b>x</b> <sub>2</sub>	0	1	1	1
<b>x</b> <sub>3</sub>	0	1	0	1
<b>x</b> <sub>4</sub>	0	0	1	1
<b>x</b> <sub>5</sub>	1	1	0	0

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

	$p(\mathbf{x}_1=1 \mid c=k)$	$p(\mathbf{x}_2=1 \mid c=k)$	$p(\mathbf{x}_3=1 \mid c=k)$	$p(\mathbf{x}_4=1 \mid c=k)$
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

- Perform one expectation maximization iteration.
- Verify that after one iteration the probability of the data increased.

5. Consider the following data points:

	y <sub>1</sub>	y <sub>2</sub>
<b>x</b> <sub>1</sub>	2	2
<b>x</b> <sub>2</sub>	0	2
<b>x</b> <sub>3</sub>	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  $p(k = 2) = 0.4$

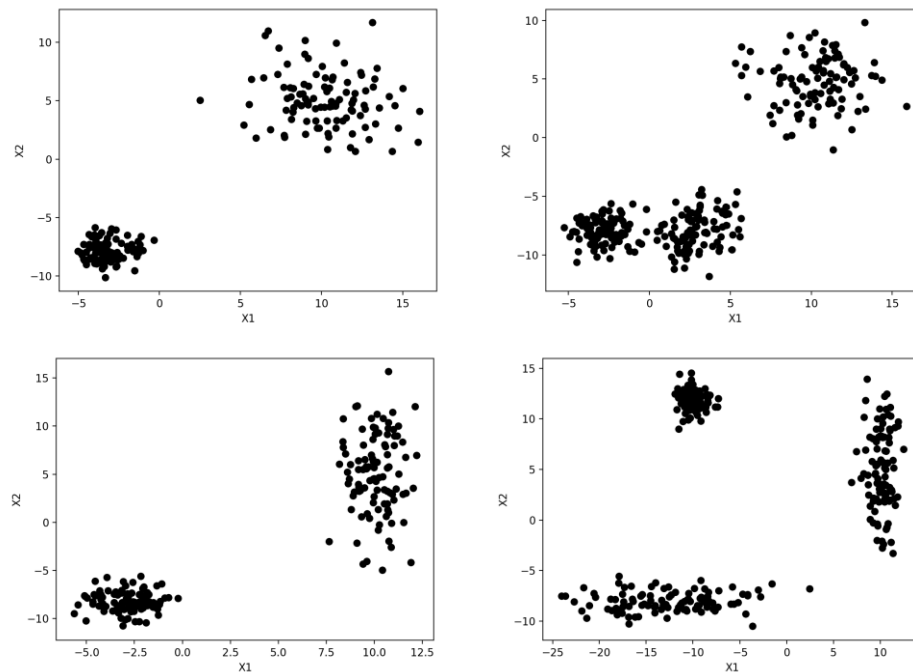
- Perform one expectation maximization iteration
- How much the fitting probability increased?
- Sketch the points and clusters

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

- Assuming observations **x**<sub>1</sub>, **x**<sub>4</sub> and **x**<sub>7</sub> to be the initial seeds, identify the centroids after the first epoch using:
  - k-means
  - k-median
- When is median preferred over mean?

	y <sub>1</sub>	y <sub>2</sub>
<b>x</b> <sub>1</sub>	2	10
<b>x</b> <sub>2</sub>	2	5
<b>x</b> <sub>3</sub>	8	4
<b>x</b> <sub>4</sub>	5	8
<b>x</b> <sub>5</sub>	7	5
<b>x</b> <sub>6</sub>	6	4
<b>x</b> <sub>7</sub>	1	2
<b>x</b> <sub>8</sub>	4	9

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



- For each scenario, justify whether  $k$ -means is suitable
- 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.
- When moving from numeric to ordinal data spaces, is Hamming distance proper to handle ordinal data with high cardinality?

## Programming quest

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

- 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  $k$

Resource: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

- After selecting two informative features OR extracting two principal components  $\text{PCA}(n\_components=2).fit(X)$ , visualize the produced clusters

Resource: <https://scikit-learn.org/stable/modules/clustering.html>