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

b) Plot the data points and draw the clusters

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

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.

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 $k = 3$

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

c) 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:

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

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

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

Plot the clusters after one iteration of the EM algorithm.

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 priors:

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

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

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

b) How much the fitting probability increased?

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

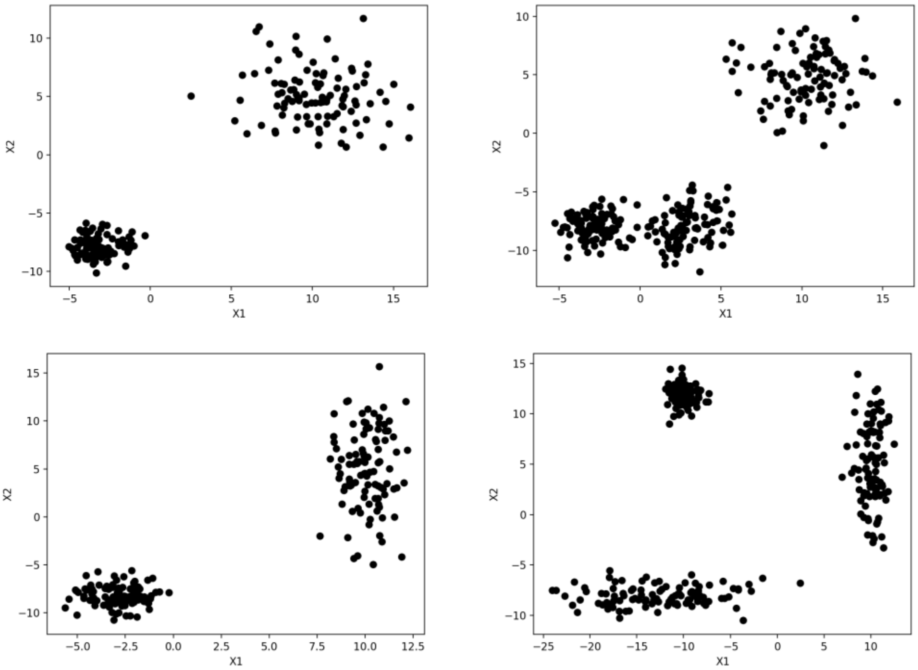
a) Assuming observations $\mathbf{x}_1$, $\mathbf{x}_4$ and $\mathbf{x}_7$ to be the initial seeds, identify the centroids after the first epoch using:

   i. k-means

   ii. k-median

b) When is median preferred over mean?

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



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

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.

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

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