TÉCNICO LISBOA

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework 4
Deadline 30/10/2024 20:00
Submit on Fenix as pdf

I) (7 pts) Clustering

Given the data

$$\mathbf{x}_1 = \begin{pmatrix}2.5\\2.5\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}2\\2\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}0.5\\0.55\end{pmatrix},$$

$$\pi_1 = 0.6, \pi_2 = 0.4$$

$$C_1\left(\mathbf{u}_1 = \begin{pmatrix}2\\2\end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix}1 & 1\\0 & 1\end{pmatrix}\right), \quad C_2\left(\mathbf{u}_2 = \begin{pmatrix}0.5\\0.5\end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}\right).$$

i) (6 pts)

Perform one iteration of EM clustering algorithm step by step and determine the new parameters. Indicate all the calculations step by step. (To make the calculation easier for each step you can use a computer, however you should be able to do it by hand)

ii) (1 pts)

Performing a hard assignment of observations to clusters identify the silhouette of the larger cluster

II Software Experiments (3pts)

a) (2 pts)
Download the jupyter notebook HM4_CL.ipynb. Load the build in data set “wine” preform k-means and EM clustering with 2, 3, 4 cluster and indicate the silhouette as defined in the notebook for each experiment. Which k-value give the ideal value.

Perform PCA with two components with 2, 3, 4 cluster and indicate the silhouette as defined in the notebook for each experiment. Which k-value give the ideal value? Is the ideal k value the same with PCA and without?

b) (1 pts)

Load the build in data set “breast_cancer” preform k-means and EM clustering with 2 cluster and indicate the silhouette as defined in the notebook. Which one is better?

Perform PCA with two components with 2, cluster and indicate the silhouette as defined in the notebook. Plot the scatter plot. When you compare the plots and the silhouette values and look at the scatter plot of the PCA mapped data, what is your conclusion. Short, one sentence pls.