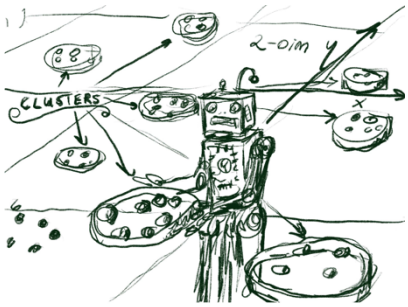
TÉCNICO LISBOA

LEIC-T 2024/2025
Aprendizagem - Machine Learning
Homework 4
Deadline 28/10/2024 21:00
Submit on Fenix as pdf



I) (7 pts) Clustering

Given the data

$\mathbf{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}-1\\-1\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}0.5\\0.55\end{pmatrix},$

$\pi_1 = 0.6, \pi_2 = 0.4$

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

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)

Download the jupyter notebook HM4_24_CL.ipynb.

a) (0.5 pts)
Load the build in data set “wine”. Perform k-Means and EM clustering with 3 clusters and indicate the silhouette as defined in the notebook. Which clustering, k_means or EM-Clustering is better?
b) (0.5 pts)
Perform PCA with two components on the build in data set “wine”.
Plot the scatter plot. Can the three classes be separated? (one sentence)
c) (1 pts)
Perform k-means and EM clustering with 3 clusters on the generated data set in question b) and indicate the silhouette as defined in the notebook. Plot the scatter plot of the clustering as in the notebook. Which clustering, k_means or EM-Clustering is better? Why are the silhouette values different to the values in a)? (one/two sentence/s)

d) (1 pts)
Load the build in data set “breast_cancer” perform 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.