1 Introduction

The goal of this assignment is to practice using different clustering techniques. We will contrast K-means, Soft K-Means and DBSCAN, 3 techniques which you have seen in class.

1.1 Getting started

As in the previous practical, you will use the MLDemos toolkit that provides a collection of machine learning algorithms which you can apply on hand-made as well as real-world datasets.

I) Load your data The data you will be using is the same you collected during your previous assignment (Assignment I, Part I PCA). If your data set is clearly separable you should create a new one or edit your data set to create partial overlaps between the different classes.

II) Cluster your data After loading your data, open the Algorithms window and select the Clustering tab. Choose the method and parameters (e.g. metric power, stiffness, ϵ, etc.) and hit the Cluster button.

2 Assessing Clustering Methods

2.1 Qualitative assessment

For this part, ignore the optimize and test buttons in mldemos.

Assess qualitatively the performance of the 3 clustering methods you have at your disposal on your dataset. For each method, test different values of its hyperparameters (K the number of clusters for K-Means, K and β for Soft-K-Means, ϵ and minPoints for DBSCAN) and observe the different clustering results you obtain.

For K-means, try changing the type of metric (Euclidean, Manhattan, Infinite and Polynomial).
2.2 Quantitative assessment

For this part, ignore the test button and the associated Train Ratio list in mldemos.

You will perform this quantitative assessment only on K-Means and Soft-K-Means.

Look at the values of the RSS, BIC and AIC metrics for each choice of parameter in the blue box and report on major changes in values for RSS, AIC and BIC for those particular choices of hyperparameters. (See figure 2). The F1-Mesure in this box is always given for 100% of labelled points.

Using the optimization of the set of clusters for K-Means and Soft-K-Means clustering on your dataset, and using BIC, AIC, and RSS as metrics choice can also enables you to see the
values with respect to the chosen number of clusters. (See figure 2)

Then, using the optimization process, perform also a comparative analysis of the effect of providing part of the labels in semi-supervised clustering, using the F1-measure. After choosing F1 to optimize, select different ratios of labelled datapoints just below, typically going from very low ratios to high ratios, and optimize. Keep PrintScreens of the different graphs to be able to compare. Note that the values in these graphs are scaled so that each metric can be compared to all the others. Don’t pay attention to the F1-Measure given in the blue box which is always computed on 100% of labelled points but compare how the ratio of labelled points influences the F1-Measure for different numbers of clusters looking at the graph and so the optimized number of clusters which is found.

2.3 Special note on the F-Measure

The F-Measure is a means for evaluating the clustering ability in a setting where the data is only partially labeled. The equations you have seen during the lecture are re-iterated below:

\[
F(C, K) = \sum_{c_i \in C} \frac{|c_i|}{M} \max_k \{F_1(c_i, k)\} 
\]

\[
F_1(c_i, k) = \frac{2 \cdot R(c_i, k) P(c_i, k)}{R(c_i, k) + P(c_i, k)}
\]

\[
R(c_i, k) = \frac{N_{ik}}{N_{c_i}}
\]

\[
P(c_i, k) = \frac{N_{ik}}{N_k}
\]

Where \(M\) is the number of labeled data points, \(K\) is the number of clusters, \(C = \{c_i\}\) is the set of classes, \(N_{ik}\) is the number of members of class \(c_i\) in cluster \(k\), \(N_{c_i}\) is the number of members of class \(c_i\) and \(N_k\) is the number of members in cluster \(k\).

The \(F_1\)-Measure, is the harmonic mean between precision, \(P\), and recall, \(R\), and lies strictly in the interval \([0, 1]\).

\[
F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}
\]

The reason it is denoted by \(F_1\) in MLDemos is that \(\beta = 1\), which gives equal weighting to both the precision and recall in the computation of the measure. Saying that we just used the \(F_1\)-Measure is vague, the type of weighting should also be specified: \(F_1\) (in MLDemos only \(\beta = 1\) is available). MLDemos displays the negative of the \(F_1\)-Measure such that when it is maximal, \(F_1 = -1.0\), and when it is minimal, \(F_1 = 0.0\). The reason for this is that it stays comparable to the other measures which are optimal when they are at a minimum.