Practical session:

Density-based clustering (DBSCAN)

Biomedical Data Science

Biomedical Data Science Lab (BDSLab)

ITACA, UPV

Contents

1. Objective

To cluster a biomedical dataset in a 2D representation using DBSCAN.

2. Material

- Seminar: Density-based learning for medical problems.
- The provided data.csv for biological scRNA-seq dataset reduced in 2D with tSNE.

3. Evaluation

This practical session will be evaluated taken into account the code files, which must be submitted in the Poliformat task, along with the practical session report. This report must include:

- Page 1: cover page, title, authors and professors.
- Page 2: contents.
- Page 3 and following pages: answers to the questions and exercises raised in each block.
- Last page: references.

4. Tasks

4.1. Block I. Data loading and exploration

4.1.1. Objective

To prepare and understand your working data.

4.1.2. Questions & exercises

1. Load your data using the proper column delimiter and decimal number indicator.
2. Explore your data with a scatter plot. Do you think your data is suitable for a density-based clustering approach? Justify your answer.
3. Calculate the pairwise Euclidean distances among points.
4. Explore the calculated distances:
   4.1. Plot a histogram, increasing the number of default bins to avoid a too smoothed representation.
   4.2. Sort the distances in ascending order and plot them with a line plot.
   4.3. Make a brief comment about the figures obtained in 4.1 and 4.2.
5. Calculate the k-graph of distances among points, evaluating different values for the *k* hyperparameter.
6. Explore the calculated k-graph distances:
   6.1. Plot a histogram, increasing the number of default bins to avoid a too smoothed representation.

6.2. Sort the distances in ascending order and plot them with a line plot.

6.3. Make a brief comment about the figures obtained in 6.1 and 6.2.

4.2. Block II. Patients clustering with DBSCAN

4.2.1. Objective

To understand the DBSCAN clustering algorithm, learning how to set its hyperparameters and interpreting its outcomes while applying it to tackle a real biomedical problem.

4.2.2. Questions & exercises

1. Set the minimum number of points to determine a cluster, as well as the epsilon parameter. You will have to study how to propose good hyperparameter combinations. Consider the results from the previous block, especially those graphs obtained in exercise 6. In addition, you can use rules of thumb, grid, random or surrogate search validated with complementary clustering assessment metrics, etc. You do not need to explore all the possibilities. Just choose/propose a method, but justifying why you have selected/proposed it.
2. Once you have selected your *optimal* hyperparameter configuration, perform DBSCAN and extract the clusters obtained, calculating the number of clusters afterwards. Explore your clusters with a scatter plot, using different colors to identify different clusters. Do not forget to include in this scatter plot the outlier points, also with a different color. Make a brief comment about your results.
3. Extract the type of each point, that is, if it is a core point, a border point or an outlier point. Finally, explore your point types with a scatter plot, using a different color for each point type. Write a short discussion about your findings.
4. Why I have not asked you to split your data into a training and test set, or into a pure training and validation set? Justify your answer.
5. Do you expect your selected/proposed method to set DBSCAN hyperparameters (question 1 of this block) to work properly with other data of the same dimensionality? Do you think it should perform well in higher dimensions beyond 2 and 3? Justify your answer.