

LCPB 23-24 Exercise 2, data visualization and clustering

Exercise 4A

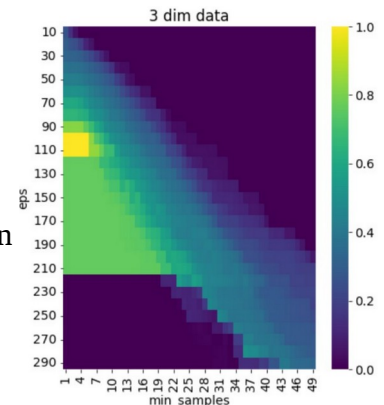
Visualize and clusterize the data in the file **x_12d.dat** (N=600 samples, L=12 dimensions), which also has labels for checking the performances (y_12d.dat).

1. “*eps*” (ϵ) and “*minPts*” (m_p) in DBSCAN algorithm for clustering

Refine the grid with more values of ϵ and m_p and plot a heat-map showing the normalized mutual information (NMI) between true and predicted clusters, similar to the one on the right.

Is the high NMI region showing a correlation between ϵ and m_p ?

Note: In the lesson we have looked at the typical distance between a point and its closest neighbor, but this does not say what the typical distance is from the 2nd, 3rd, ..., m_p -neighbor. The plots of ranked distances to the i -th neighbor might also help choose the ϵ for a given $i=m_p$.



2. Understanding the 12-dimensional data

Use the principal component analysis (PCA) to visualize the first components of the data. Does it help understand its structure?

3. Compare different clustering methods

- Perform a k-means clustering of the data, with $k=3$. Does it work better than DBSCAN? Why?
- Perform a hierarchical clustering of the data and [plot the corresponding dendrogram](#). Does it work better than DBSCAN?

4. OPTIONAL: Visualize the data with other [methods from the scikit package](#)