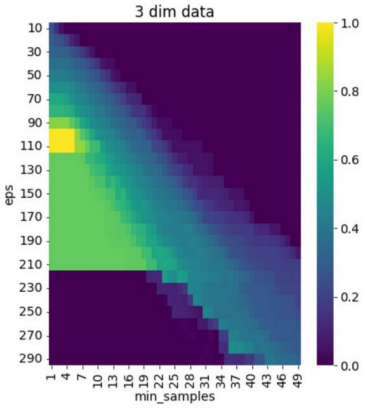
LCPB 23-25 Exercise 3, data visualization and clustering

Exercise 4

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?

   The plots of ranked distances to the i-th neighbor might help choose the ε for a given i=m_P. How does the optimal value for ε given m_P relate to the ranked distances to the m_P-th neighbor?

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 with k-means**
   Perform a k-means clustering of the data, with k=3. Does it work better than DBSCAN? Why?

4. **OPTIONAL: Compare with hierarchical clustering**
   Perform a hierarchical clustering of the data and plot the corresponding dendrogram. Does it work better than DBSCAN? Which measure did you use for distance (distance of closest points between clusters, distance between cluster centers, …)? Does it affect the result?

5. **OPTIONAL: Visualize the data** with other methods from the scikit package