Unsupervised Learning (MT7050) - Project 3

Instructions: *This project consists of 2 tasks that should be solved individually. Unless it is specified in the task, you are free to use any programming package in this project.*

The solution should be submitted at the course webpage in a single .pdf file ***with your source code attached as appendices*** *(i.e., don't put the source code in the main text). Your source code should include clear comments and documentations to describe what are evaluating.*

TASK 1 (ISOMAP or LLE)

In this task, the same dataset Swiss_Roll.txt from project 2 is considered. You can use either ISOMAP OR LLE to embed the dataset in 2 dimensions. You will compare the performance of your chosen method with the CTD embed- ding in unfolding the Swiss roll. Note again that the Swiss_Roll.txt dataset contains 2000 points. If needed, you can randomly sample a subset of points, e.g. 700 points, from the dataset to perform this task.

A) Reason with justifications how to choose the number of neighbors, k, to construct a k NN-graph before performing ISOMAP or LLE. (5p)

B) Embed the Swiss roll dataset in 2 dimensions using your chosen method with the value of k chosen in the previous exercise. (20p)

C) Discuss what you see in the 2-dimensional embedding produced by your chosen method. How does it compare to the CTD embedding? If the method does not produce a rectangular representation as expected, explain why this might be the case. (10p)

D) Explore how the underlying graph affects the results produced by your chosen method. Specifically, change the number of nearest neighbors, k, and discuss how the embedding changes and why. (15p)

TASK 2 (Density-based Clustering)

In this task, you can use either DBSCAN OR the density-based method by Rodriguez and Laio to analyze the dataset Arbitrary_Shape.txt. The 1st and 2nd columns of the data file are the x- and y-coordinate, respectively.

A) Explain the choice of parameters and reason why they are suitable. (10p)

B) Perform the clustering with the chosen method and plot the clustering result by labelling the identified clusters and "noise" with different colors. (15p)

C) Perform validation of the clustering result using the Silhouette plot and the Silhouette index. Do these validation techniques provide a reasonable evaluation? (15p)

D) Experiment with different values of the input parameters. If you are per- forming DBSCAN fix the parameter *minPts* to the value chosen in part A and investigate how changing the ϵ parameter affects the clustering result. If you are using the method by Rodriguez and Laio change the distance cutoff parameter d_c. Remember to explain why the changes you make have the observed effect. (10p)