

Unsupervised Learning (MT7050) - Project 1

Instructions: This project consists of only 1 task with 5 parts that should be solved individually. Unless specified in the task, you are free to use any programming package in this project.

The solution should be submitted on 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 documentation to describe the essential parts.

TASK 1

This task allows you to understand more about the method of kernel PCA and perform validation for the results.

First you need to generate a 2-dim concentric circle dataset with 3 circles similar to the one discussed in the class, i.e., the dataset should look like (but not exactly the same as) Fig. 1.

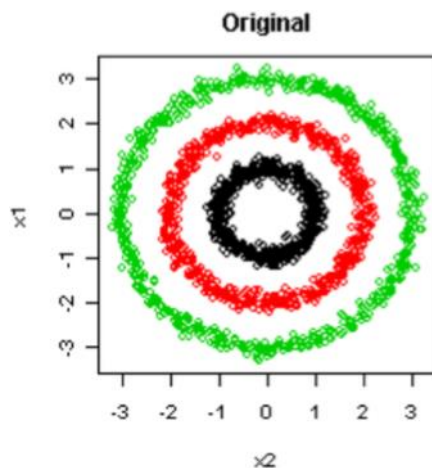


Fig.1. Concentric circle dataset

The radius of the 3 circles are 1, 2 and 3 units, respectively. The number of data points in each circle is 200, so there are 600 data points in total. You should also add some random noise to the data points so that the dataset looks like the one in Fig. 1. Each student should therefore have a different dataset. In plotting figures in your report, please use the same color scheme (black, red and green) as in Fig. 1 to label your data points.

A) Plot the generated dataset. (15p)

B) Perform kernel PCA on the dataset using a degree-2 polynomial kernel and then plot the data points in the plane of the first 2 kPCs. (15p)

C) Perform a validation of the results in part B) using a Shepard diagram for **the distance rank ordering**, i.e., the x-axis is the Euclidean distance rank in the original space, and the y-axis is the Euclidean distance rank in the 2-dim kPC space. From the Shepard diagram, discuss how the distance ranking is preserved in different ranges of ranks. (25p)

D) Now perform kernel PCA on the dataset using a Gaussian kernel with 4 different standard deviations $\sigma = 0.1, 1, 3, 10$, and for each, plot the data points in the plane of the first 2 kPCs. Discuss how σ affects the embedding results, especially in the ability to resolve the cluster structures in the projected kPC space. (20p)

E) Perform validation of the results from part D) using a Shepard diagram for **the distance rank ordering**. From the 4 Shepard diagrams associated with different σ , discuss how the distance ranking is preserved in different ranges of ranks. Finally, plot the Spearman rho (i.e. the Pearson correlation coefficient of the rank-based Shepard diagram) against σ and provide statistical explanations for the observed Spearman rho dependence on σ . (25p)
