Data Mining 2 Homework 3

Jennifer Probst

21 April 2021

Exercise 1: Data Imputation via SVD

a) The result of my imputation can be seen below. In Figure 1 we can see the original image, the image with 60 percent of the data missing and the imputed image. We observe that the structure is recovered quite well, all the shapes are visible but the coloring is a little different.

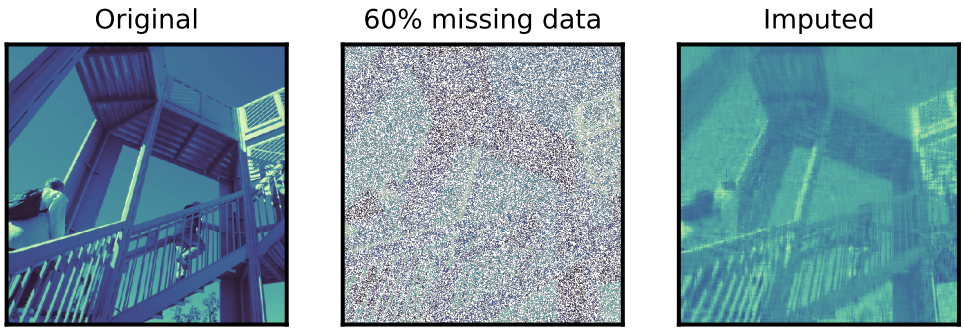

Figure 1: Original image, image with 6p percent data missing and imputed image.

b) Below in Figure 2 we can see a plot for all tested ranks r with the corresponding mean squared errors and the optimal rank highlighted in red in the plot. The optimal rank r is 19 and its mean squared error is 787.04.

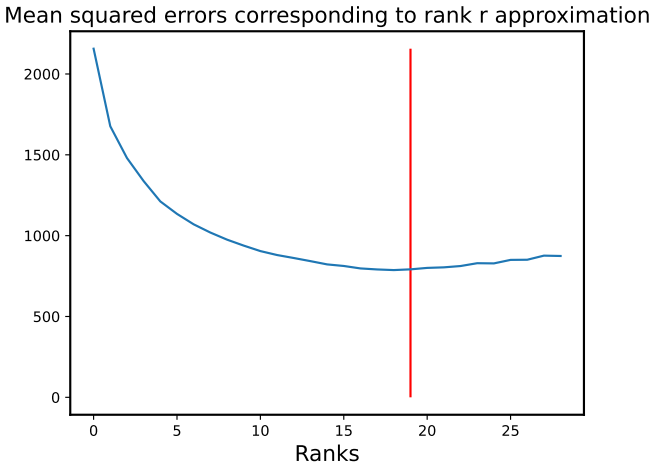

Figure 2: Plot of all tested ranks r with the corresponding mean squared errors, optimal r highlighted in red.

c) From Figure 2 we see that if we choose the value r too small or large the mean squared error increases again (imputed pixel values do resemble the test values less good). If we choose a too small value for r the picture will be too much approximated (we use too less dimensions and lose a lot of information), shapes will get less clear and colors different. If we choose a too large value for r the imputed values seem to get worse again. This could be if some correlation was described by the first few dimensions but are not that expressive anymore in combination with more dimensions (increasing r value).

Exercise 2: Kernel PCA

a) The results of the PCA reduction on the Moon data can be seen in Figure 3. The samples are colored according to their class labels. We observe that the plot looks really similar to the original image which also had 2 dimensions, so with 2 PCs we can recover the original data as it does not show any obvious linear trends.

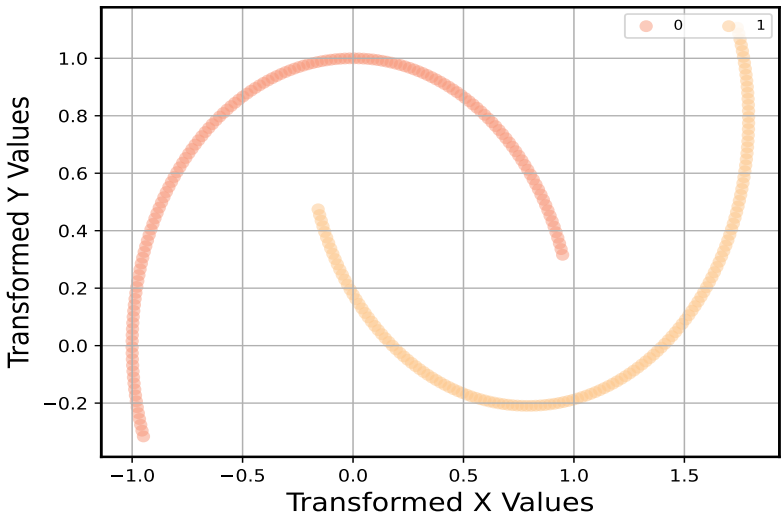

Figure 3: Plot of the transformed data in their first two PCs. Samples are colored according to their class labels.

b) Please see code.

c) The results of the application of RBFKernelPCA with values of [1; 5; 10; 20] of gamma can be seen in the following Figures 4 - 7. The figures show the first two principle components of the transformed data. The plots look quite different to the one by normal PCA. In the PCA plot it would be hard to separate the two classes by a simple classifier (e.g. SVM) but in the case of gamma = 10 or gamma = 20 we can just use a linear classifier. Gamma = 1 and gamma = 5 do not seem to be optimal as there is some class overlap making the separation harder. The optimal gamma can be found by letting a classifier classify the projected datapoints and subsequently calculate metrics of performance such as precision, recall and accuracy.

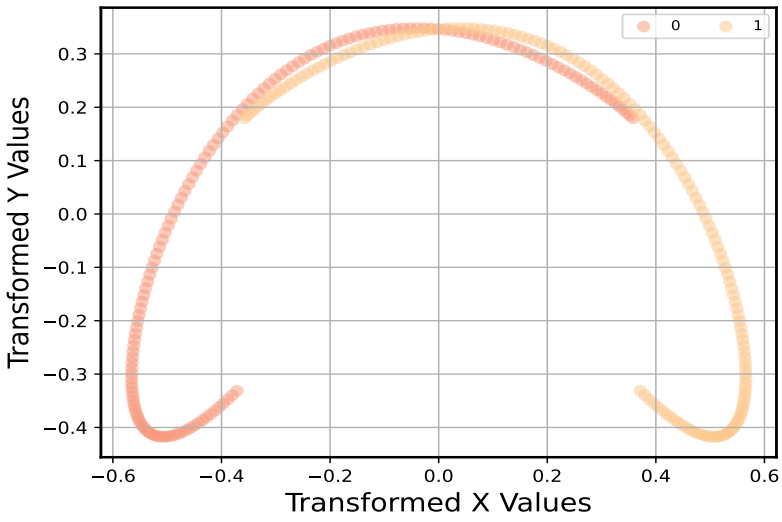

Figure 4: Plot of the transformed data using the first two PCs of Kernel PCA with gamma = 1.

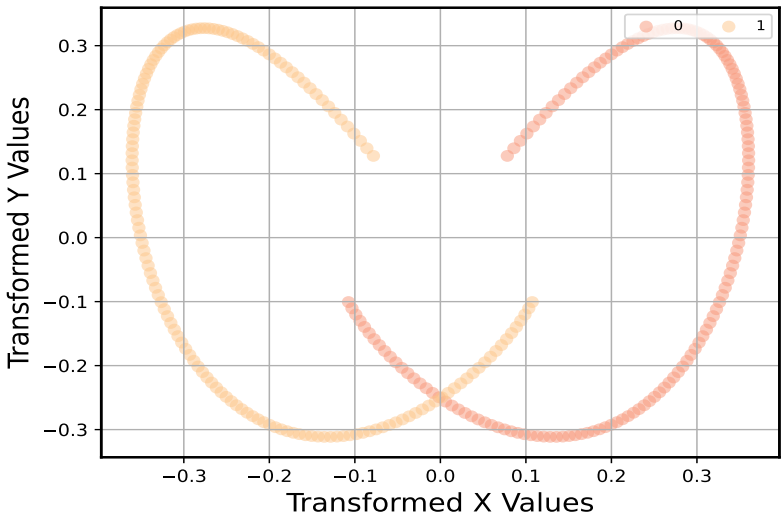

Figure 5: Plot of the transformed data using the first two PCs of Kernel PCA with gamma = 5.

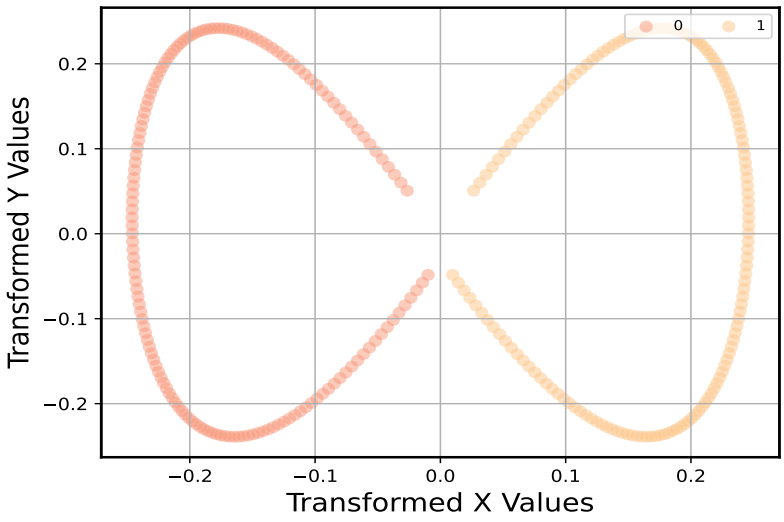

Figure 6: Plot of the transformed data using the first two PCs of Kernel PCA with gamma = 10.

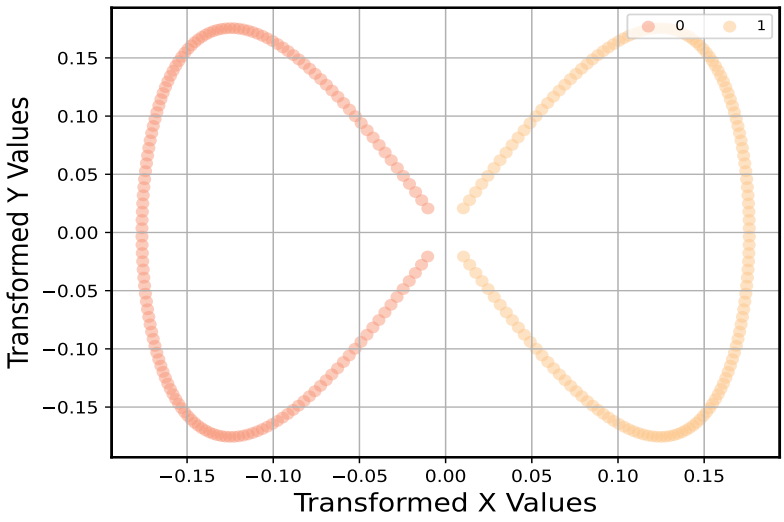

Figure 7: Plot of the transformed data using the first two PCs of Kernel PCA with gamma = 20.