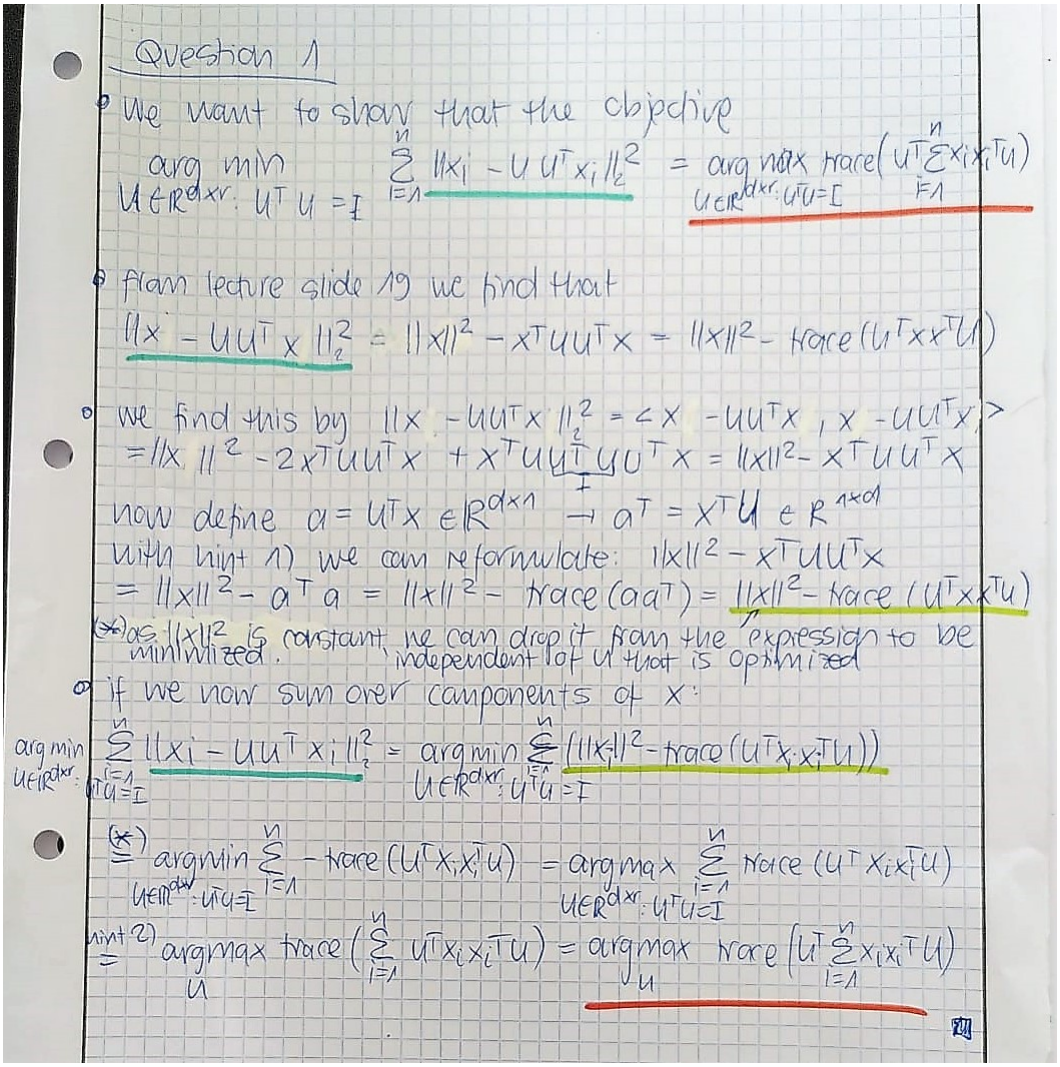
Data Mining 2 Homework 1

Jennifer Probst

16 March 2021

Exercise 1: Proof



Exercise 2

a) see code
b) see code
c) See the plot of the transformed data in a 2-dimensional subspace of their first two principal components in Figure 1. In the plot the classes are colored by their class label. We observe that the classes are not clearly separated of each other, especially the yellow class is totally mixed into the other ones. All the classes seem to have at least some overlap.
d) The plot of the cumulative variance explained by the principle components can be seen in Figure 3. One can observe that the first few PCs explain almost all the variance. Already 69 percent of the variance (more than 50 percent) is explained by the first PC, the first three principal components explain about 87 percent of the variance and the first five components more than 95 percent of the variance.
e) We perform a data processing step with mean centering of the data and standardization. Looking at the plot of the samples (Figure 2) plotted across their first two components, we find that the data is much better separable, there is only slight overlap between the yellow and green and the green and violet classes. But there are many more PCS needed to explain the variance in the data (Figure 4). We now need 2 PCs to explain 50 percent of the data, three to explain 80 percent of the variance and 17 PCs are needed to explain 95 percent of the variance.
Data normalisation might be a necessary step to properly separate the data. If we don't normalize the data we give features with large values big importance but relative to their large values the variance might not even be so big and the feature not that important.

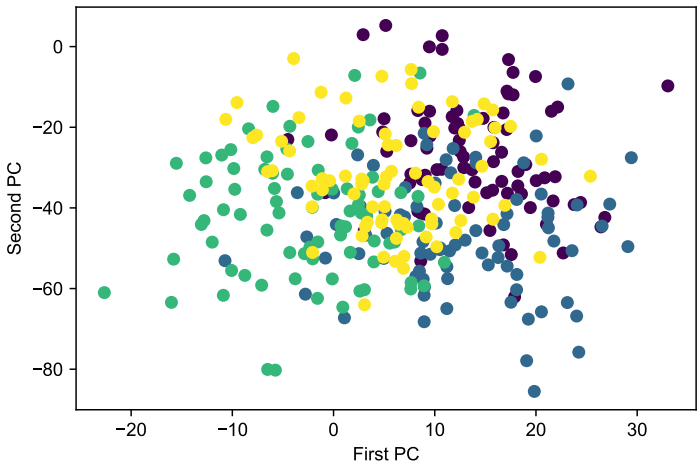


Figure 1: Plot of the transformed data in their first two PCs. Samples are colored according to their class labels.

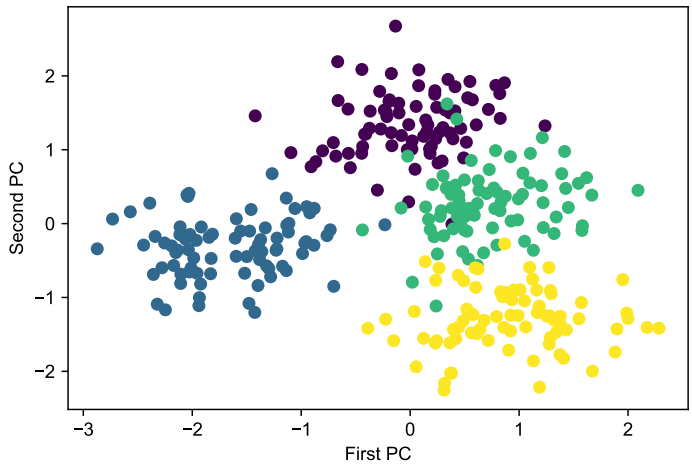


Figure 2: Plot of the normalized and transformed data in their first two PCs. Samples are colored according to their class labels.

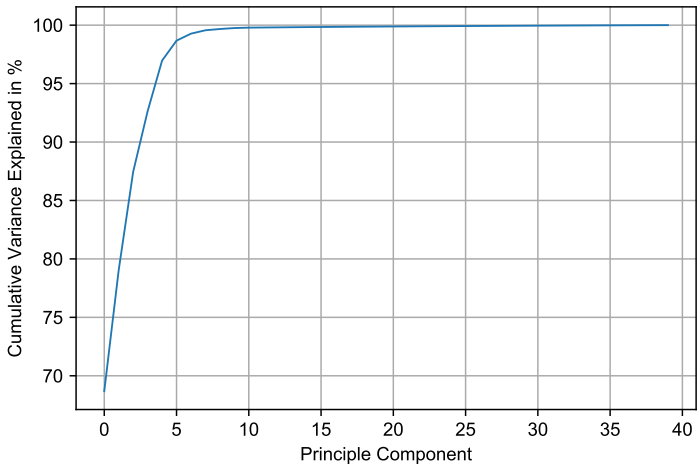


Figure 3: Cumulative variance explained plot for all your principal components for analysis on non-normalized data.

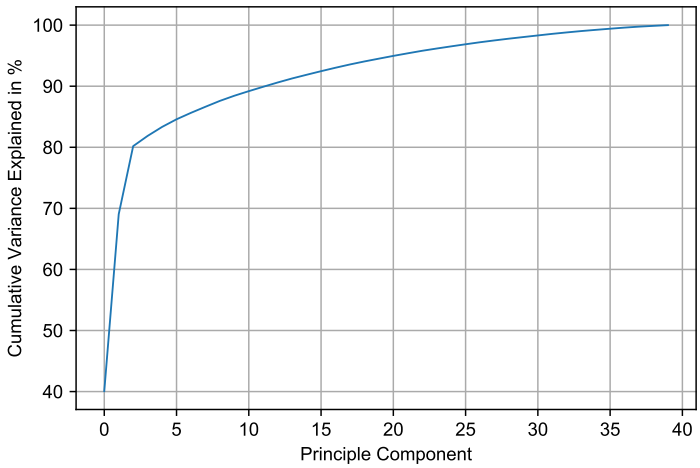


Figure 4: Cumulative variance explained plot for all your principal components for analysis on normalized data.