Prof. Dr. Thomas Schultz
Mohammad Khatami (khatami@cs.uni-bonn.de)
Ikram Jumakulyyev (ijumakulyyev@cs.uni-bonn.de)

University of Bonn
B-IT/Computer Science II
May 8, 2020

Summer term 2020

Visual Data Analysis

Assignment Sheet 4

Solution has to be uploaded by May 18, 2020, 8:00 a.m.
to https://uni-bonn.sciebo.de/s/0a1e9bFxuQhjkcC with password `vda.2020`

Please bundle the results (as PDF) and scripts (*.py/*.ipynb files) in a single ZIP file. Submit each solution only once, but include names and email addresses of all team members in the PDF and each script. Name the file `vda-2020-xx-names.zip`, where `xx` is the assignment sheet number, and `names` are your last names.

If you have questions concerning the exercises, please write to our mailing list:
vl-scivis@lists.iai.uni-bonn.de.

Exercise 1 (Principal Component Analysis, *25 Points*)

In this task, you will use Principal Component Analysis (PCA) to explore a Breast Cancer Dataset, which contains 699 samples, 458 of them indicating a benign tumor and 241 a malignant one. In the experiments, each measurement is graded between 1 to 10 at the time of sample collection. Nine measured characteristics such as clump thickness, uniformity of cell size, uniformity of cell shape etc. were found to differ the most between the benign and malignant samples. More details on the attributes and experiments can be found in the `breast-cancer-wisconsin.names` file.

a) Read the `breast-cancer-wisconsin.xlsx` file. Note that there are some instances with missing data, which have to be imputed before we can run PCA. Pandas offers convenient functions for this. Apply an imputation method that makes sense for this dataset, and briefly explain your decision. (3P)

b) Create a plot that, for any number n, shows what fraction of the overall variance in the data is contained in the first n principal components. Make sure that you only include the nine relevant numerical attributes in the PCA, not the sample codes or class IDs. How many components do we need to cover $\geq 90\%$ of the variance? (5P) *Hint:* You may use the implementation of PCA that is provided in the Python package scikit-learn.

c) Each sample is now characterized by a point in PCA space. Create a scatter plot matrix that shows the first five principal components. Each diagonal cell should contain two overlaid density plots, one for the benign and one for the malignant class. Use different colors to distinguish between the classes, and add a legend that clearly states which samples are benign or malignant. (3P)

d) Which PCA mode shows the strongest difference between the benign and the malignant samples? Name the original variables that have the highest and lowest weights in its definition, respectively. (3P)

e) Scatterplot matrices often reveal outliers in the data. Visually identify at least one sample that is far away from the others, remove it from the dataset, and re-generate the scatterplot matrix without it. (3P)

f) In the breast cancer dataset, all variables x_i have a similar range, $x_i \in [1, 10]$. If the variables of a dataset have very different ranges, for example one variable $x_1 \in [1000, 2000]$ and another one $x_2 \in [1, 5]$, how would this affect the PCA? Could it make sense to pre-process the data in such cases? Why and how? (3P)

g) Explain why, on this dataset, we cannot use Linear Discriminant Analysis (LDA) to create an alternative 5D embedding in which the classes are more clearly separated. Use the LDA implementation in scikit-learn to perform a 1D LDA embedding and plot it (in a scatterplot) against the first principal component. Do they show a clear correlation? Is this true in general, or specific to the dataset? (5P)

Exercise 2 (Dimensionality Reduction, *14 Points*)

Please answer the following questions in your own words.

a) Assume a PCA projection $\mathbf{y} = U_k^T (\mathbf{x} - \bar{\mathbf{x}})$ as defined in Chapter 4, Slide 17. Given a point $\mathbf{y}_j \in \mathbb{R}^k$ in k-dimensional PCA space, how can we recover a corresponding point $\mathbf{x}_j \in \mathbb{R}^p$ in the original p-dimensional space? Can we find such a $\mathbf{x}_j$ for any $\mathbf{y}_j$? Is $\mathbf{x}_j$ unique? Briefly explain why. (4P)

b) In Chapter 4, Slide 34, we specified the following equation for the projection of point $\mathbf{x}$ in Kernel PCA:

$$\mathbf{y} = \mathrm{D}_k^{-1}\mathrm{U}_k^{\mathrm{T}}\mathrm{H}\left(\mathbf{k} - \frac{1}{n}\mathrm{K}\mathbf{1}\right) \quad \text{with} \quad k_i = K(\mathbf{x}_i, \mathbf{x}) \tag{1}$$

However, we did not justify it in full detail. Please provide a step-by-step derivation starting from the equation on Slide 28,

$$y_a = \sum_{i=1}^{n} \alpha_i^a \mathbf{z}_i^{\mathrm{T}} (\mathbf{x} - \bar{\mathbf{x}}) \tag{2}$$

Hint: Write down Eq. (2) in feature space (involving $\Phi(\mathbf{x})$), replace dot products in feature space with evaluations of the kernel function K, then consider the relationship between the α_i^a in Eq. (2) and the matrices in Eq. (1). (5P)

c) If the given dissimilarities correspond to a valid metric, Multidimensional Scaling (MDS) can be solved as an eigenvector problem, in analogy to kernel PCA. What problem can arise if we attempt to use this approach even though the metric assumption is violated? (2P)

d) The ISOMAP algorithm involves construction of a neighborhood graph. Briefly explain two different approaches for this step and describe a situation in which you would prefer one of them over the other. (3P)

Exercise 3 (Comparing RadViz and Star Coordinates, *11 Points*)

In the lecture, we introduced star coordinates and RadViz as alternative radial data visualization techniques. The paper `rubio-sanchez-radviz-sc-2016.pdf`, which is available from the lecture webpage, provides a closer investigation of their relationship. Please read it and answer the following questions in your own words. Remember that **we will not grant even partial credit for copy-pasted text.**

a) In general, the authors of this paper prefer star coordinates over RadViz. Briefly explain two reasons for their preference. (2P)

b) Despite this, the authors mention a specific use case where they consider RadViz to be superior to star coordinates. Briefly explain what this use case is and why they prefer RadViz in this case. (2P)

c) Neither RadViz nor star coordinates provide a one-to-one mapping between a data point's high-dimensional location and its two-dimensional projection. Interactively modifying the anchor points or axes, respectively, can reduce the resulting ambiguities. However, the paper mentions two examples in which, unlike star coordinates, RadViz introduces ambiguities that cannot be resolved with any re-arrangement of the anchor points. Point out these two examples. Could the extended RadViz that we learned about in the lecture resolve them? (3P)

d) Several methods have been proposed to optimize the locations of anchor points in RadViz to obtain an improved separation of classes in the resulting projection. How did two such algorithms compare to supervised linear dimensionality reduction techniques in the experiments reported in this paper, in terms of (i) computational effort and (ii) achieved class separation? (2P)

e) The authors propose that star coordinates could be combined with Linear Discriminant Analysis in order to perform a manual feature selection. Briefly explain how that approach works. (2P)

Good Luck!