Prof. Dr. Thomas Schultz
Shekoufeh Gorgi Zadeh (gorgi@cs.uni-bonn.de)

B-IT
Bonn-Aachen International
Center for Information Technology
April 16, 2018

Summer term 2018

Visual Computing in the Life Sciences

Assignment Sheet 2

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

This exercise can be submitted in **small groups** of 2-3 students. Please submit each solution only once, but clearly indicate who contributed to it and remember that all team members have to be able to explain it.

Solutions have to be sent until ***April 30, 2018, 10:30am*** *to gorgi@cs.uni-bonn.de. Please bundle the results as PDF and scripts in a single ZIP file. Include your names explicitly in the PDF. Use the following format for naming the files "vclsi-x-lastName.pdf" and "vclsi-x-lastName.ipynb", where x is the ID of assignment sheet. For example, for Alice Smith, the file name must be vclsi-2-smith.pdf.*

Exercise 1 (Producing a Scatterplot Matrix, *25 Points*)

In the previous assignment, you wrote a reduced dataset to disk that is limited to the benign and malignant classes and five variables that most strongly distinguish between benign and malignant samples. This week, you will create and interpret a basic visualization of that data.

In this assignment your final visualization should be a 5×5 matrix whose rows and columns are the measurements of the variables you selected last week. Diagonal cells visualize how the variables are distributed; off-diagonal cells visualize the relationship between the values of pairs of variables.

Please proceed in the following steps and submit your final script, the final image, and answers to the questions:

a) Each diagonal cell should contain two overlaid histograms, one for the benign and one for the malignant class. In the histogram, variable values should be on the x axis, the frequency of observing that value in each class should be on the y axis. Use different colors to distinguish between the classes, and add a legend. Your visual design should make it easy to answer the following questions (5P for implementation, 1P for justifying choice of colors, 3P for answering questions):
   - For which variable(s) you could find a range of values for which the class of the sample is certain? Write down the ranges.
   - Which variable(s) has(have) almost a uniform distribution for the malignant samples?

b) In each non-diagonal cell, display a scatter plot that visualizes the values of the corresponding pair of variables. Use different colors and opacities so that it is simple to relate these scatter plots to the density plots on the diagonal, and the size of the marker should reflect the number of overlapping points. (5P for implementation, 2P for answering questions):
   - Point out a pair of variables whose values have a positive correlation overall.

- Can you identify a pair of variables for which the values are highly correlated in one group of subjects (e.g. malignant), but less so in the other group?

c) Compute the distance consistency of all scatter plots. Which pair of variables leads to the highest distance consistency? (6P)

d) Imagine that, given only the values of two variables, you will be asked to decide whether they are from a benign sample, or a malignant one. Which pair of variables would you choose to make that decision? Why? (3P)

Hint: You can use the Python toolkit matplotlib to create plots. More information on it is available from http://matplotlib.org/.

Exercise 2 (Principal Component Analysis, *25 Points*)

It is difficult to fully visualize a very high-dimensional space. In the first assignment sheet and the previous exercise, we therefore focused on a few variables that we found to be particularly discriminative. In this exercise, we will instead employ dimensionality reduction on the values of all variables.

a) Perform a Principal Component Analysis (PCA) on the values.

Write a program to read the `breast-cancer-wisconsin.xlsx` file again. Interpolate missing values as before, but keep all variables this time. Make 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. 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.

b) Each sample is now characterized by a point in PCA space. Create a scatter plot matrix (in the same manner as in the previous sheet) that shows the first five principal components. This time, instead of histograms, each diagonal cell should contain two overlaid density plots, one for the benign and one for the malignant class. In the density plot, variable values should be on the x axis, the frequency of observing that value in each class should be on the y axis. Use different colors to distinguish between the classes, and add a legend. (5P)

c) In which PCA modes do you see a clear difference between the benign samples and the malignant samples, in which modes the difference is less? (3P)

d) Sometimes outliers (points that are quite far away from the rest of the data) could affect the data analysis. Provide the sample-Code or row index of the furthest point of malignant samples in the fourth PCA mode. Then remove that sample using its row index. (5P)

e) See what happens when we re-weight the variables to emphasize those that discriminate well between the benign and malignant classes. To do so, compute F scores (cf. sheet 2, task 1 d)) and multiply each data value by its corresponding F score. Create two scatter plots to compare PCA results with and without the re-weighting. (5P)

f) In the breast cancer data-set, all the variables have a similar range of values $v \in [1, 10]$. If the variables of a data-set have varying ranges, for example one variable have values around 1000 to 2000 and another around 1 to 5, how could this affect the PCA performance. Explain how would you solve this problem? (2P)

Good Luck!