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

B-IT
Bonn-Aachen International
Center for Information Technology
April 9, 2018

Summer term 2018

Visual Computing in the Life Sciences

Assignment Sheet 1

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 16, 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-1-smith.pdf.*

Exercise 1 (Read, Write, and Filter Data, *25 Points*)

In the first project, you will work with a Breast Cancer Dataset, which contains 699 sample information, 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 and etc. were found to differ the most between the benign and malignant samples. In this project we will use visualization tools to analyse this 9 dimensional data set. Note that there are some instances with missing data. For more details on attributes and the experiments read the `breast-cancer-wisconsin.names` file.

Please download the `breast-cancer-wisconsin.xlsx` file and proceed in the following steps and submit your final script. You will also need its results for the next assignments. Make sure to submit the output of your code as well as the source code.

a) Read the dataset and print the number of instances and columns, as well as the column names, to the terminal (3P).

b) Interpolate the missing values in a sensible way. Briefly explain your decision (3P).

c) Extract Benign and Malignant subgroups and print the number of instances for each subgroup (4P).

d) The F score is one way to determine how well a given variable distinguishes between two groups. F is large if the differences of the two groups means $\bar{x}_1$ and $\bar{x}_2$ to the grand mean $\bar{x}$ of all data points is large relative to the variances within the groups. Given groups of size n_1 and n_2 with items $x_{1,i}$ and $x_{2,i}$, respectively, F can be defined as

$$F = \frac{(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2}{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2 + \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2,i} - \bar{x}_2)^2}$$

Define a function that calculates F for any given attribute and for the benign and malignant class labels (8P). Use it to identify the five attributes that best separate the classes benign vs. malignant (3P). What do you expect the value of F will be for the "class" attribute itself? (1P)

e) Write a reduced dataset to disk. It should contain only the five most relevant attributes from d) for the 699 samples. There must be no missing values in the reduced dataset using part b) (3P).

Hint: You can use pandas, a powerful Python data analysis toolkit, for this assignment. It provides fast, flexible, and expressive data structures for working with relational or labeled data. To become familar with it, you can refer to `http://pandas.pydata.org/pandas-docs/stable/10min.html`.

Exercise 2 (Face-based Luminance Matching, *10 Points*)

Working in visualization requires the ability to efficiently extract relevant information from research papers. To practice this, please download the paper `kindlmann-luminance-2002.pdf` from the lecture webpage and answer the following questions (using 1-2 sentences for each). It is acceptable if you provide answers without having read the full paper in detail (even though it's a good paper and worth your time :). However, your answers have to use your own words. **We will not grant even partial credit for copy-pasted text.**

a) What is the Helmholtz-Kohlrausch effect? (2P)

b) Why are the authors proposing to use images of faces? (2P)

c) To what alternative method do the authors compare their newly proposed one in the user study? (2P)

d) Based on the result of the user study, what is the advantage of the newly proposed method? (2P)

e) Why do the authors have to know the monitor gamma while creating a colormap based on the result of the user study? (2P)

Good Luck!