Data Visualization – Assignment 4

– Visualizing tabular data with many features –

P. Rüdiger, J. Lukasczyk · H. Leitte – winter term 2018/19

In the fourth assignment we will work on complex data that cannot be directly visualized. Hence, we will combine mathematical data analysis with visualization. We will concentrate on tabular data with many features and use principal component analysis (PCA) as mathematical tool.

In exercise 1, we will work on a cars dataset with 26 features and revise the methods we have learned in class. In exercise 2, we will try to characterize cancer cells using approx. 7000 features. We will see some new techniques and (hopefully) think about the critical aspects of PCA.

Please hand in your solutions in the OLAT system. Submit the jupyter notebook (including the analysis task as text cells with markup) and a hmtl copy of the notbook. Do not hand in individual files, no zip-file.

Due date: Thursday, 03. Jan. 2019, 22:00

Exercise 1 – Characterizing car types

14 points

The first dataset `93cars.dat.csv` contains 93 new car models for the 1993 year. 26 attributes are given for each car including manufacturer, prices, mpg, engine and body sizes. The dataset description can be found here `http://ww2.amstat.org/publications/jse/datasets/93cars.txt`. Missing values are imputed by median values.

Due to the number of variables, visualization approaches using direct mappings are not applicable. Your task in this exercise will be to describe characteristics of different car types using PCA analysis as performed in the lecture.

Task 1a) (2P): Load the data in the jupyter notebook and answer the following questions:

- (1P) Which of the variables are quantitative and can be used in a PCA?
- (1P) Which variables can be used to divide the cars into groups? How many groups do you obtain and what size do they have? Do you expect differences between the groups and if yes which? (Use at least two variables for grouping.)

Task 1b) (2P): Compute the PCA using the `scikit-learn` learn implementation as discussed in the lecture. Remember to only use the attributes you identified before and standardize the data.

Task 1c) (2P): Find out how many principal components to use. Use the provided plot to answer the following questions:

- (1P) How much variance explains the first and the first two principal component roughly?
- (1P) How many components do you need to explain 90% of variance in the data (roughly, use figure estimate)?

Task 1d) (8P): You are given a projection of the data onto the first two principal components in the jupyter notebook. You can also specify one of the input variables as label that is used to color the glyphs.

Your task is to explain the new coordinate system.

- (3P) Extend the visualization to a biplot including labeled vectors for the original axes.
- (3P) Explain the PC1 and PC2 axis. What is typical for cars on the left vs. right (x-axis)? What is typical for cars located towards the top vs. bottom of the chart (y-axis)?
- (2P) Use one of the groupings you discussed in Task 1a) and color the plot accordingly. What can you tell about the groups? Feal free to comment on additional findings here.

Exercise 2 – Analyzing tumor data

18 points

The tumor dataset provides data for 64 tumors (each from a different patient). For each tumor gene expressions were measured for 6830 genes. Hence, we have a 64×6830 dataset which is given as a transposed matrix ($\rightarrow 6830 \times 64$) as e.g. excel cannot handle that many columns.

The cells mostly come from known cancer types and if known the respective information is given in the data. The classes are breast, cns (central nervous system), colon, leukemia, melanoma, nsclc (non-small-cell lung cancer), ovarian, prostate, renal, K562A, K562B, MCF7A, MCF7D (those four are laboratory tumor cultures which we will not use) and unknown.

We will try to characterize each cancer type using the gene expression data. Idea: Different types of cells have different expression profiles. Many diseases, including cancer, fundamentally involve breakdowns in the regulation of gene expression. The expression profile of cancer cells becomes abnormal, and different kinds of cancers have different expression profiles. The goal is to find abnormal gene expression profiles to characterize different types of cancer (each tumor/patient belongs to one of the cancer classes).

The data originates from the following project:

`http://genome-www.stanford.edu/nci60/`

and is provided, for example, through the book Friedman et al. (2001). The elements of statistical learning. Springer.:

`http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`

Task 2a) (8P): First we want to judge if the task is feasible.

- (1P) Create 6 scatterplots for random gene combinations. Can you find clusters for the different cancer types, i.e. do the different colors form groups of points that are distinct from the other colors?
- (2P) Investigate the second chart given in the notebook and analyze the code. What does the chart show?
- (3P) What is the current runtime of the algorithm? How can you improve it? What is the improved runtime?
- (2P) In summary, what do you think? Is it possible to classify different types of cancer using gene expression data? Which cancer types are probably very hard to describe which are easy? Why?

Task 2b) (8P): Now we will compute the PCA of the tumor data and augment cancer classes.

- (4P) The notebook already contains a PCA plot. Fill in the code to compute the PCA and add a convex hull to the points of each class (see sample picture).
- (3P) Identify a tumor class that
  - forms a cluster in the projection.
  - does not form a compact cluster.
  - cannot be distinguished from others using our PCA analysis.

  For each answer: Say which, and explain your answer.

Task 2c) (2P): Take a look at the provided explained variance plot. How reliable is you analysis based on the first two components?