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 1, 2020

Summer term 2020

Visual Data Analysis

Assignment Sheet 3

Solution has to be uploaded by May 10, 2020, 23:59
to https://uni-bonn.sciebo.de/s/sd7XmsLorbS39Kt 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 (Kernel Density Estimation, *10 Points*)

In the lecture, we introduced Kernel Density Estimation (KDE) for visualizing the density of a given dataset and talked about the importance of the right bandwidth choice. In this exercise, you will confirm it practically. You will also experiment with different kernel functions in KDE.

Hint: You do not need to implement KDE by yourself, implementations are available in several Python packages. However, they differ in how they interpret the bandwidth parameter. In particular, scikit learn uses it directly as the kernel bandwidth, while scipy interprets it as a factor that will be multiplied by the standard deviation of the dataset to obtain the kernel bandwidth. To obtain the expected results, please use the implementation from scikit learn for this assignment.

a) Please create a set of 2000 random data points taken from two normal distributions with the same standard deviation $\sigma = 2$, and different means. 25% of your points should use $\mu_1 = 0$, the remaining 75% $\mu_2 = 9$. (1P)
b) Plot a histogram of the data you created in a). Select a number of bins that clearly reveals the two Gaussians. (1P)
c) Apply KDE with "Gaussian" and "Linear" kernels, with bandwidths equal to 1. Plot the results. Describe which differences you observe, and briefly explain why they arise. (2P)
d) Use some established heuristic to estimate the optimal bandwidth for your data. Briefly state which heuristic you chose and what is the result. Apply KDE with a Gaussian kernel function for the result you obtained, and for bandwidths equal to 0.07 and 2.0. Describe the differences you observe. (4P)
e) Briefly summarize your insights about the importance of the bandwidth and kernel function parameters in KDE. (2P)

Exercise 2 (Multidimensional Data Filtering and Visualization, *26 Points*)

In practice, data filtering and visualization often go hand in hand. In this task, you will continue to use pandas and seaborn for these tasks. We will explore how differences in (sensory) wine quality relate to

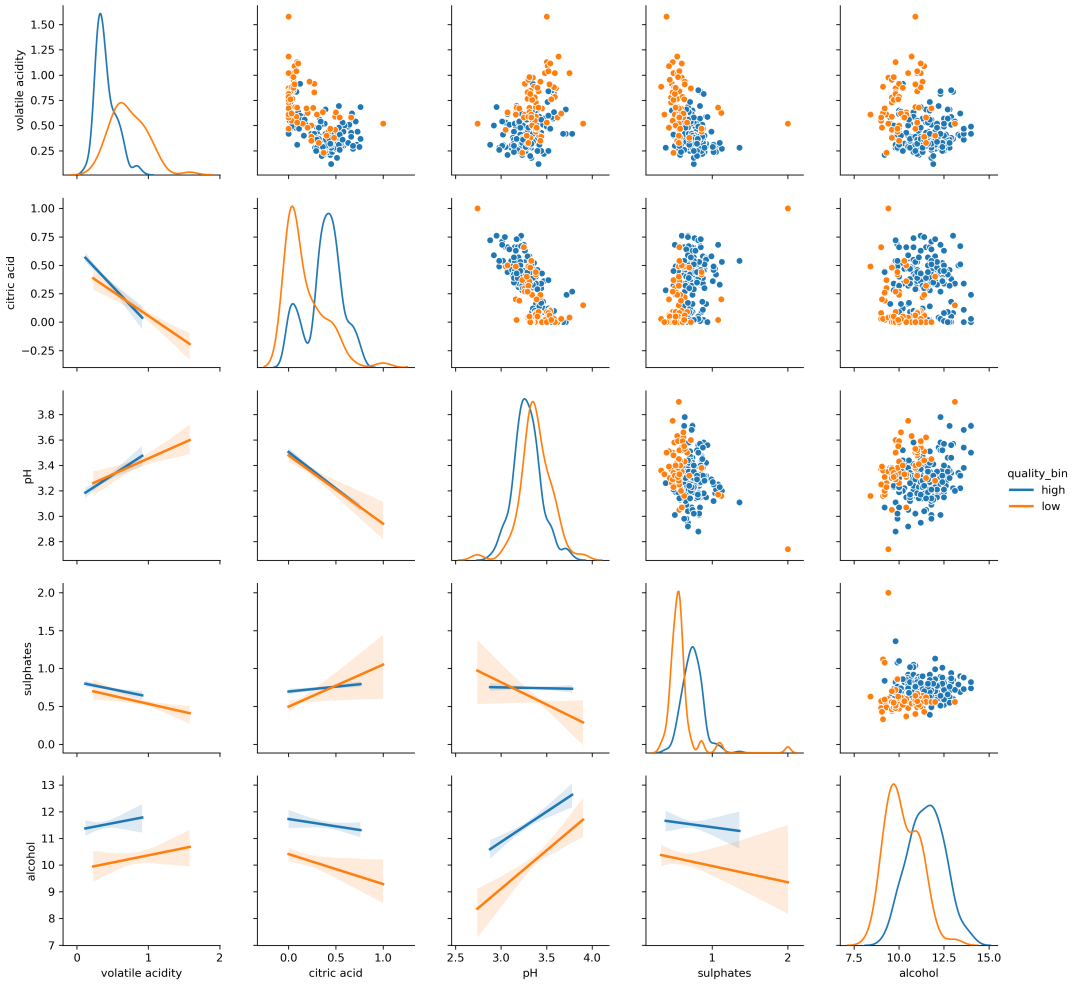

Figure 1: The visualization which you should obtain in step h) of this task.

physicochemical quantities, such as amount of residual sugar or acidity. You can find a more detailed description of this dataset at the UCI Machine Learning Repository

Hint: Most tasks require writing very little, and not very complex code. The main task will be to identify and make proper use of the required functionality in pandas, seaborn, and scikit learn.

a) Read the data given in `winequality-red.csv`, available from the lecture webpage, and print the first few rows. (1P)

b) A numerical rating of sensory wine quality is given in the column "quality". Display the distribution of these scores with a histogram. What is the range of this score in the data? (1P)

c) Derive a coarser classification of quality into "low", "medium", and "high", by grouping together the two lowest, the intermediate, and the two highest quality scores that occur in the dataset, respectively. Replace the original "quality" column with a new column "quality bin" that contains these labels. (3P)

d) We would like to investigate differences between high and low quality wines. Therefore, create a filtered data frame in which the medium-quality wines are omitted. (1P)

e) Visualize all numerical attributes in a scatterplot matrix. Color the two quality levels differently. (1P) *Hint:* Using seaborn, you can create this plot with a single (and simple) line of code.

f) Based on the visualization, name five attributes that appear to best distinguish between high and low quality. (1P)

g) Now, use an automated feature selection technique to identify five attributes that distinguish between high and low quality. More specifically, please use the F score from a one-way analysis of variance (ANOVA) to rank the attributes, as implemented in scikit learn's `f_classif`. What are the five best attributes according to this measure? Are they the same as those you identified visually? Create a filtered data frame that only contains the top five attributes, plus the "quality bin". (4P)

h) Create a matrix similar to the one in Fig. 1: It should compare the two quality bins with respect to the five top-ranking attributes, using density estimates (on the diagonal), scatterplots (in the upper triangular part), and plots of pairwise linear regression models (in the lower triangular part). *Hint:* It's convenient to use seaborn's `PairGrid` class for this. (4P)

i) Based on the visualization, which attributes appear to be strongly correlated regardless of quality? For which attributes does the amount of correlation appear to depend on the quality? Does any of the attributes appear to have a multimodal distribution? Point out one or multiple data points that appear to be outliers. (4P)

j) Compute the distance consistency of all scatter plots. Which pair of variables leads to the highest distance consistency? (6P) *Hint:* To our knowledge, distance consistency is not implemented in any widely used Python package. For this subtask, we expect that you will have to write some code yourself.

Exercise 3 (Evaluating PCP Variants, *14 Points*)

Parallel Coordinate Plots are a popular technique for visualizing multidimensional data. They represent a basic idea for which many variations have been proposed, even though the lecture only covered a tiny fraction of them. In this task, you will get to know some of these modifications. To this end, please read the paper `holten-pcp-evaluation-2010.pdf`, which is available from the lecture webpage.

Please answer the following questions in your own words. Remember that **we will not grant even partial credit for copy-pasted text.**

a) In addition to standard PCPs, the authors test eight variations. For each of them, briefly describe the proposed modification, and why the authors expected it to improve the visualization. Use 1-2 sentences for each of the eight cases. (8P)

b) Within the user study, what task did the subjects have to perform? Name another task for which Parallel Coordinates are frequently used in practice, but which was not included in the study. (2P)

c) To which extent did the results of the study match the authors' hypotheses? (2P)

d) Which of the explored modifications would you consider using when designing a visualization based on Parallel Coordinates? Briefly justify your answer. (2P)

Good Luck!