TKO3103 Exercise 1, Data understanding and visualization

Technicalities

Contact info: You can access me at office/by email/by moodle messages: Paavo Nevalainen ptneva@utu.fi . Office hours: Mon 10-11:00, ICT C5064 till Nov 23rd, 452B (Agora ICT) from Dec 7th.
Grading: 3+3 points, work can be done **alone** or **in pairs**. Reasonable answers to all the specific questions result in reasonable grades. The points count fully in your exam evaluation.
Personal submits: Both members of the pair have to submit the same document in Moodle.
Exam qualification: return both exercises first.
Penalty from delays: 1 points per each week after the deadline.
Exam disqualification: See details from the submission timetable in the end.
Deadline: Mon Nov 30th 2015 at 23:55 (ex. 1), Mon Dec 14th 2015 at 23:55 (ex. 2)
Next exercise available: Tue Nov 17tth 2015
Tools: python, Matlab or R (in the preferred order)
Goal: Familiarize yourself with some of the data understanding and visualization methods discussed in the course.
Python aides: gettingStartedWithPython.pdf, pythonInstallInstructions.pdf
Data set: http://archive.ics.uci.edu/ml/datasets/Wine+Quality

- Solo students:
  odd student number: choose **red wine**
  even student number: choose **white wine**
- **S**tudents in pairs: apply the above rule to the person first in alphabetical order of surenames

Tasks

1) Plot histograms of the attributes. To determine the number of bins use at least three different methods introduced in the lectures (Sturges' rule, Scott's rule, square-choice, Freedman-Diaconis rule). Compare the results.

2) Produce scatter plot of the data and the parallel coordinates representation. (If you don't find an implementation, it is easy to code oneself too.)

3) Principal component analysis (PCA) with and without normalization: project the data to the first two principal components. What can be observed?

4) Produce 2D MDS representation for the data. Compare the result to PCA.

5) Calculate the Spearson and Kendall's tau correlation tables for the attributes.

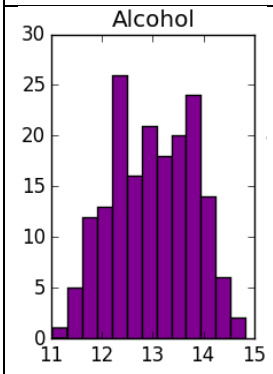
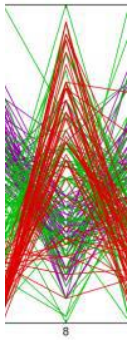
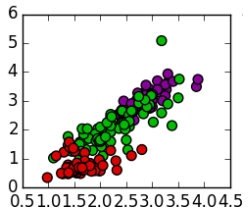
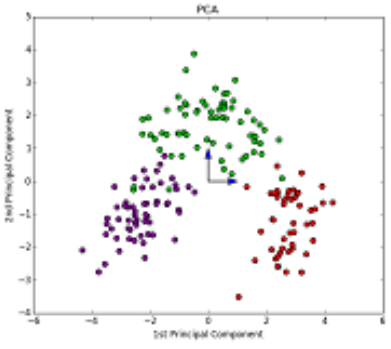
If some of the concepts sound unfamiliar, attend lectures and study the slides.

Report

Report that includes the results and conclusions made. Submit it via Moodle **in pdf format**.
Include your name (and the name of your partner). See example snapshots at the next page.

Visual examples

Remember to include a short explanation of what you did to get a visualization.

A histogram	A parallel coordinates representation	A scatterplot	A PCA plot
			

Points decline after the deadline...

Monday/Tue/Wed/Thu/Fri/Sat/Sun			
00:00 Tuesday...	exercise 1	exercise 2	...Monday 23:55
...	3	3	Nov 30th
Dec 1st	2	3	Dec 7th
Dec 8th	1	3	Dec 14th
Dec 15th	0	2	Dec 21st
Dec 22st	0	1	Dec 28th
Dec 29th	0	0	Jan 4th
Jan 5th	disqualified		...