Data Mining
Homework Assignment #4

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

March 13, 2014

You are free to use any programming language you are comfortable with. Hints in R are optional.

Task 1

Listen to the presentation by Tamara Munzner: Keynote on Visualization Principles - `http://vizbi.org/Videos/26205288`. Summarize the key take-home messages from her presentation.

Task 2

Take a look at three following charts:

- `http://www.billboard.com/biz/articles/news/digital-and-mobile/5827354/the-download-hits-middle-age-and-it-shows`
- `http://junkcharts.typepad.com/.a/6a00d8341e992c53ef019b021a098f970b-pi`, source
- `http://www.technologyreview.com/graphiti/520491/mobile-makeover/`

Choose one of them and answer following questions:

- What is the key message for the chart?
- Is it easy to grasp?
- Is it possible to improve it according to Tamara's suggestions? How?
- (Optional) Redraw it (either on paper or using some visualization tool) according to your suggestions.

Task 3

Read data from the file, data.txt, calculate mean and variance for every feature (column). Compute correlation between pairs of features x1 vs y1, x2 vs y2 etc. Compare results. Plot each of these pairs separately (such that x feature is on the x-axis and y feature on the y-axis). Interpret the results.

Task 4

Perform a "Single Link" clustering of 2-D data from the table below. Use Euclidean distance as a distance measure. Draw a dendrogram/tree with node height at the distance at where the clusters were merged. Hint: Draw the points first on 2D and then perform manual simulation. (Solutions on paper are ok).

	X	Y
A	2	4
B	7	3
C	3	5
D	5	3
E	7	4
F	6	8
G	6	5
H	8	4
I	2	5
J	3	7

Task 5

Think of university study information system - all data about every student, grades, and dates, teachers, feedback, etc. Propose some visual report/summary that might help programme managers, head of the institutes, deans, and rector to decide about student progress in a curriculum (or comparisons in between curricula and institutes, for example). What information would need to be fitted on such a visualisation? Make an illustration of your idea (can be done on paper). Be as creative as possible.

Task 6 (2pt)

Use MeV (`http://www.tm4.org/mev.html`) or R or any other tool offering hierarchical clustering and cluster hierarchically some data of interest to you or USArrests data set that we have provided you with on the homework page.