Data Mining
Homework Assignment #6

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

March 20, 2014

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

Task 1

Use the data from last week Task 4, and simulate K-means algorithm. Use initial centers of (2,6), (2,8), (5,8). Explain algorithm step by step. Then use the same data and simulate K-medoids, starting from cluster center points D, E, and H. Data is this:

	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 2

Can you find a different initial 3 starting "centers" for K-means of previous task that would produce a different final result? Use 2D plot of the data to assist you.

Task 3

Install and run mldemos (`http://mldemos.epfl.ch/`) and try out the clustering with K-means. Identify situations when K-means clearly does not cluster as expected as compared the true clustering expected by you. Make screenshots and discuss why it happens.

Task 4

When you have identified why such unpleasant situations arise - can you propose some remedy to it? Propose some heuristics how to overcome such issues.

Task 5

During the lecture we described the Self-organizing maps (SOM) clustering method and principle. Implement the SOM algorithm in the modification that has only 1-dimensional "grid". E.g. that has 30 or 100 or n grid elements. Take the new datapoint and assign it to the most similar grid point, then update that point and a nearby range of other points. Outline the exact algorithm in pseudocode.

Task 6 (2pt)

Implement your 1-D SOM algorithm yourself and apply it.

Task 7 (2pt)

Perform clustering analysis on student progress data set that you have seen on Thursday and Friday practice sessions. I have modified it a bit, added more information. Due to the privacy reasons I could not add to many columns. As this exercise is rather optional, don't limit yourself with just k-means or k-medoids, try different clustering approached that you have learned so far. Pose interesting questions e.g. can you distinguish between different groups, do you see the difference in grading style of TAs, or perhaps you can identify people that work together on homeoworks etc. Try to visualize your hypothesis, use statistics that we have studied previously. Be creative!