Data Mining
Homework Assignment #6

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

March 21, 2014

You are free to use any programming language you are comfortable with.

Task 1

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

	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 3 initial “centers” for K-means that are different from the centers in the previous task and 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/`). Try out the clustering with K-means. Identify situations when K-means clearly does not cluster as compared to the true clustering (i.e. the clusters expected by you). Make screenshots and discuss, why it happens.

Task 4

Once you have identified the unexpcted situations, propose some remedy for it. In other words, propose some heuristics how to overcome these issues.

Task 5

In the lecture we have discussed the principle of self-organizing maps (SOM), which is very powerful and widely used clustering method, now we shall practice using them. Consider the following one dimensional data set D that contains ten points:

Point index	Points
1	1
2	1
3	1
4	1
5	1
6	10
7	10
8	10
9	10
10	10

We are trying to cluster above data set using one dimensional SOM that has four nodes with the following indexes:

Node index	Nodes
1	2
2	4
3	6
4	8

Your task will be to simulate five iterations of SOM algorithm using following parameters:

$$\theta(u,v) = \begin{cases} 1 & \text{if } u = v \\ 0.5 & \text{if } |u-v| = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha(s) = 0.01$$

where $\theta(u, u_{winner})$ is a neighborhood function, and $\alpha(s)$ is a learning restraint of the algorithm.

Here we will show how to perform first iteration, your task is to perform the rest four iterations and report. Make sure you understand how it works. Write the pseudocode of the SOM algorithm with your comments.

Task 6 (2pt)

Implement your 2-D SOM algorithm yourself and apply it on the data of xxx.

Task 7 (2pt)

Perform clustering analysis on the data of students' progress from this course. Here (link to the data) is a slightly modified version with some personal information added. Due to the privacy reasons we could not add too many columns. As this exercise is a bonus, don't limit yourself with k-means or k-medoids, try different clustering approaches 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!