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 it.
You are given following 1-D input vector $\mathbf{p} = \{1, 1, 1, 1, 1, 10, 10, 10, 10, 10\}$ that can be regarded as 10 separate 1-D data points (Note, we are using small bold letters to address vectors). We are trying to cluster these data points using SOM that has four nodes (some time regarded as neurons). The tricky part is that every node in this grid (SOM) is at the same time associated with grid coordinate and weight. In our case four nodes are represented by the grid coordinate vector $\mathbf{v} = \{1, 2, 3, 4\}$ and vector of weights $\mathbf{w} = \{2, 4, 6, 8\}$, such that for example node with grid coordinate 2 has weight 4, and node with grid coordinate 4, has weight 8.
Your task will be to simulate five iterations of SOM algorithm, which can be summarized as follows:

- Choose one data point $p \in \mathbf{p}$ at random
- Find the best matching unit (BMU) u for the chosen point p, it is the node with smallest Euclidean distance to the given point.
- Update each node weight $w[i], i \in \{1, 2, 3, 4\}$ using following formula:

  $$w[i] = w[i] + \theta(u, v[i]) \cdot \alpha \cdot (p - w[i]),$$

  where $\theta(u, v[i])$ is a neighborhood function, which is defined:

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

  and

  $$\alpha = 0.01$$

  where, u and $v[i]$ are the **grid node coordinates** not **weights** and α 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...

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!