Data Mining
Homework Assignment #7

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

07-03-2014

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

Task 1

Explain one advantage and one disadvantage of DBSCAN over the K-means. Briefly discuss the situations when DBSCAN would fail. Draw a 2-D example of one such situation (e.g. on paper or mldemos or using any other visualization tool).

Task 2

Use already well-studied example dataset from two previous homeworks and apply DBSCAN using $eps = 2$, $MinPts = 2$.Declare, which points are noise, border and core points.

Task 3

Take a look at the following tool: `http://biit.cs.ut.ee/misc/imgvalidate/`. There is a file already uploaded for you. It contains 388 rows (from rev0 to rev387) of a picture. Values in each row correspond to "pixel color values" - rgb components of a pixel. In this file rows are shuffled. For example, if you insert the row ids as they are provided in the shuffled document (like here:), you will see a mess. Your task is to recover an original picture using 1-D Self-organizing maps (SOM), where a vector of grid coordinates $\mathbf{v} = \{0, 1, 2, 3.., 387\}$ and a vector of weights $\mathbf{w} = \{\boldsymbol{w_0}, \boldsymbol{w_1}, .., \boldsymbol{w_{387}}\}$, where $\boldsymbol{w_i}$ is each row of a picture(see HW 6.5 for the explanation of SOM notations). If you manage to cluster and/or serialize rows properly, then you will see an image of Mona Lisa.

Task 4

Compare results from task 3 with another clustering or seriation method (e.g. K-means, DBSCAN, two-way seriation etc.).

Task 5

Look at example of binary matrices from here (tarball here). Write a script that calculates some goodness measure for estimating how good the two-way seriation is. For example, give a reward for 1-s maximally surrounded by 1-s. Explain your measure and describe on toy examples how to maximize this measure by reordering. In the following example, the proceeding matrices would be more expensive than previous ones:

```
10101      10101     11100
01010 ->   10101 -> 11100
10101      01010     00011
```

Task 6 (2pt)

Write a program that performs re-ordering of rows and columns and tries to optimize for measure devised in the task 5. You can attempt some brute force, or evolutionary strategies (genetic programming, differential evolution, simulated annealing, etc). How big examples can you handle and how certain can you be that your code gives a good (or best) answer?