

Data Mining: Practical Assignment #5

Due on Thu & Fri, June 01-02 2017,

Task 1

Today, we investigate the Self Organizing Map (SOM), which is a clustering algorithm partitioning data in an unsupervised fashion. In the tutorial zip-file you find several files with 2D data (`triangle.dat`, `box.dat`, `clusters.dat`, `capitals.dat`) and 3D data (`box.dat`, `sphere.dat`). Also available is the program `data_create.py` we used to create the data.

Explore how the given SOM represents these data, i.e. which clustering is found by the algorithm. To understand SOM functionality better try different network topologies (1D and 2D) and sizes. 1D means here that the SOM consists of a chain of neurons, while 2D refers to a grid structure. In particular, perform the following experiments using the `SOM.py`:

1. Observe and explain how a 1D network clusters the data `triangle.dat`.
2. Observe and explain how a 2D network folds itself into the data `box.dat`.
3. Another cluster algorithm you know from the lecture is implemented in `KMEANS.py`. Compare its behaviour to a SOM on the data `clusters.dat`. Set $k = 4$ for both algorithms and explain the outcome.

Task 2

Now solve the TSP with the Genetic Algorithm (GA) given in `GA.py`. Look into the code to recapitulate how this method works.

1. Describe the genotype (representation of a problem solution).
2. Describe the following steps of the algorithm:
 - (a) Cost function.
 - (b) Parent selection.
 - (c) Crossover operator.
 - (d) Mutation operator.
 - (e) Survivor selection.
3. In which of these steps and for which purpose is randomness being used?
4. Vary the parameters in the program (specifically mutation rate and crossover) and describe your observations.
5. From all the parameters involved in the algorithm, which configuration optimized your results (i.e. produced the best result)?
6. Can the costs increase during learning (can we escape from local minima if not allowing increasing costs) and what is/would be the reasons for that?

@home Task

To prepare the next tutorial, your homework will be to learn about the following topics:

- Fuzzy Logic Operations
- Modelling with Fuzzy Logic: What could be interesting tasks to model with Fuzzy Logic?

Recommended literature:

1. Han J. & Kamber, M. Data mining: Concepts and techniques. Elsevier/Morgan Kaufmann, Amsterdam, 2006.
2. Kantardzic, M.: Data mining : concepts, models, methods, and algorithms. Wiley, NY, 2011.