Autumn 1999, Kaski & Lagus

Tik-61.184, Instructions for the project work

Goals

The purpose of the project work is to get hands-on experience with applying the SOM to data analysis, as well as on reporting your work to peers. Two kinds of studies are possible:

1. A (small) scientific study in which the goal is to make scientific experiments and end up in conclusions on, for example, some aspect of the SOM, on some feature extraction method, or on some application. The only constraint is that the study must involve the SOM.
2. A practical case-study where the goal is to use the SOM to explore some interesting data set and make conclusions on the data itself (and on the application that generated the data).

Difficulty

Please note that the project work is quite demanding like both scientific studies and practical projects usually are. Therefore it is essential to begin the project by defining the (preliminary) goals and methods in order to keep the project manageable.

Co-operation

The work as well as reporting can be carried out in groups of two. However, this should entail actually solving the arising problems together and learning from each other. If you simply divide the problem into two disjoint parts there is, of course, no chance of learning the other part.

Deadline

The deadline for the project work is **31.1.2000**.

Programs

Any SOM implementation may be used, but we recommend the programs SOM_PAK and SOM Toolbox, developed in the Neural Networks Research Centre and the Laboratory of Computer and Information Science and downloadable at `http://www.cis.hut.fi/cis/research/`.

Reporting

Describe your studies in a brief, clear report of about 5-10 pages. The report should contain

- an abstract
- a clear introduction that describes the problem setting and goals
- brief description of the methods and parameters used in the analysis
- description of the results
- conclusions
- references

Voluntary seminar

If you wish, you can present your results and get feedback in a seminar that will be held November 30. Please enroll at November 16 at the latest.

Possible topics

Any topic related to SOMs is possible in principle, but discuss your choice with the course assistant before starting the project. Projects that are connected to your own work or research interests are especially encouraged (but please do not report the results of any old study).

The book "Self-Organizing Maps" as well as the papers in the book "Kohonen Maps" provide examples of suitable studies—the project work is intended to have a much smaller scale, of course.

If you do not have any suitable data set to analyze, there are some suggestions below. The data sets are available from the course assistant.

Final note: It is easy to underestimate the amount of work in a research project. There usually are problems that cannot be anticipated. However, if you do underestimate even after this warning, it is acceptable to write a report on an unfinished project as long as you analyze clearly what went wrong and what you would do to solve the problem if you had the time.

Small research projects

These projects are intended to resemble real scientific studies in that the definition of the problem is left somewhat open. Creative imagination can and should be used in finding relevant and interesting problem settings that can be solved in the given amount of time. Some hints are given in each case and the lecturer can be consulted in case of problems.

1) Exploration of statistical data sets: Structures of welfare and poverty in the world. A data set consisting of 39 indicators describing different aspects of welfare and poverty of the countries can be used for organizing the countries on a SOM display. The purpose of this assignment is to study how the different choices made in preprocessing and teaching affect the resuls.

It is possible to study the effects of, for example, the following factors.

- selection of the input variables
- scaling of the variables
- size and shape of the map
- learning parameters (neighborhood function, learning rate)
- treatment of missing values in the input data

(Several of these should be studied.)

The following methods can be considered when evaluating the results: subjective evaluation, quantization error, cost function of the map, distances of pairs of samples on different maps. Other innovative solutions are encouraged.

Dataset: `poverty.dat`
Description of the abbreviations and list of the indicators: `poverty.des`

Readings:

S. Kaski and T. Kohonen (1996) Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world. In: Neural Networks in Financial Engineering. Singapore: World Scientific, pp. 498-507.

S. Kaski and T. Kohonen (1995) Structures of welfare and poverty in the world discovered by the self-organizing map. Technical Report A24. Helsinki University of Technology, Laboratory of Computer and Information Science.

2) Phonotopic map Digitized speech data can be used for forming so-called phonotopic maps where each location represents one type of a phoneme. The aim of this study is to form such maps starting from the raw speech waveforms and to visualize changes in the speech signal as trajectories on the map.

In this study it is important to apply suitable preprocessing methods to obtain input vectors that describe essential characteristics of the input data. It is recommended to use frequency analysis: relative signal power at different frequency bands has been used successfully in speech recognition. Such features can be computed with FFT from short-time windows, whereafter the absolute value of the spectra is logarithmized, and the power on suitable frequency bands is averaged.

Dataset: `speech.asc.gz`
Description of the data is given in `speech.des`

Readings:

T. Kohonen (1988) The "Neural" Phonetic Typewriter. IEEE Computer, March, pp. 11-22. (Note: The SOM theory in the paper is somewhat outdated. The new book is recommended instead.)

Two tutorial articles in Proc. IEEE.

3) Exploration of textual data Textual documents can be organized on a self-organizing map quite like any other data items if the documents are first encoded suitably as numerical vectors. The simplest methods for encoding them is to identify each word with one dimension in the data space and to use the frequency of occurrence of the word as the value of the dimension. The frequency of occurrence can be additionally weighted according to the importance of the word, measured in some manner.

Relevant questions that should be addressed in this study are: how to scale (weight) the dimensions (words) and how to reduce the dimensionality of the resulting vectors. Alternative weighting schemes include inverse document frequency and entropy over different topic areas, and of course discarding irrelevant words manually. Dimensionality can be reduced for example by principal component analysis.

The quality of the results can be measured, for example, by the overlap of different topic areas on the final map.

Dataset: `docterm.asc.gz` (`terms.asc`, `docs.asc`, `docs.tar.gz`)
Description of the data: `docterm.des`

Readings:

Kohonen et al. (1996) Very large two-level SOM for the browsing of newsgroups. Proc. ICANN'96, pp. 269-274. (Somewhat outdated but contains a description of entropy-based weighting.)

4) Visualizing the organization of a SOM The lattice of a Self-Organizing Map can be considered as an "elastic network" that is fitted into the data space. The network is non-linear but locally the non-linear surface can be approximated by a linear hyperplane. This property could be utilized to visualize the direction of the network at any location on the map, by fitting a surface to the model vectors of the neighbors of a node and visualizing the direction of that surface. There exist also other possible methods.

One possible **dataset:** `animals.dat`
descriptions of the components: `animals.des`

Readings: S. Kaski et al (1998) Methods for interpreting a self-organized map in data analysis. Proc. ESANN'98. To appear.