T-61.6060 Special Course in Computer and Information Science VI P

Data analysis and environmental informatics 5 cr (Spring 2006)
Hollmen, Sulkava, Heikinheimo

Data analysis exercise

The purpose of the exercise is to solve some data analysis problems in the scope of the course *Data analysis and environmental informatics* using real-life data. You will think of the solutions to the given problems, implement the solutions, present and discuss the results in a report formatted like a conference paper. In each of the tasks, it is important that you give grounds on the selected methodology and discuss the findings, especially the validated accuracy and the relevance to the original task to be solved.

The deadline of the exercise is 31st of May. Return your report to the course assistant.

Prediction of the age of Abalon

Use the Abalone database from the UCI Machine Learning Repository, available at `http://www.ics.uci.edu/~mlearn/MLRepository.html` to predict the age of Abalon (ear shell living in the warm seas) from a set of eight physical measurements.

Segmentation of environmental time series data

Use a segmentation algorithm of your choice to divide two time series into segments. The first time series is the Senegal river annual discharge data for the years 1903–1986 (P. Hubert, `http://www.cig.ensmp.fr/~hubert/segment.htm`). The second time series is the reconstructed mean annual Northern Hemisphere temperature data for the years 1400–1980 (M. E. Mann, R. S. Bradley, and M. K. Hughes `http://www1.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/mann1998/`). These data can be loaded into Matlab from `http://www.cis.hut.fi/Opinnot/T-61.6060/timeseriesdata.mat`.

How did you choose the number and type (e.g. constant or linear) of segments? Is this kind of analysis reasonable for these data? Why or why not? Can you think of a better way to do time series segmentation for these data? If you wanted to study the possible causes of the changes seen in the two time series data, what kind of auxiliary data would you use and why?

Climatology database, population database

Combine the two sources of data in the nordic countries region and try to predict the population density with the help of climatological data. Is this possible, and if yes, how accurately? Either way, justify your answer. In addition, try to find distinct climatological regions by means of clustering or segmentation. Find geographical regions in the data that resembles Espoo as much as possible.

The data can be loaded from `http://www.cis.hut.fi/Opinnot/T-61.6060/climate_pop_data.mat`. The file contains the variables `longitude`, `latitude`, `population`, `temperature`, `temperature_range`, `elevation`, `precipitation` and `precipitation_seasonality`. The data is presented in a 162 times 108 grid. To plot the variables in Mathlab you can use for instance the commands:

```
x1 = reshape(longitude,[162 108]);
x2 = reshape(latitude,[162 108]);
x3 = reshape(temperature, [162 108]);

figure;pcolor(x1,x2,x3);title('Temperature');
```