Information Theory and Data Mining

-

Final Project

-

Mauro De Sanctis

General Recommendations

The project should be carried out using a mathematical programming language (i.e. Matlab or Octave). Good programming discipline should be followed when writing the Matlab-code. This means that the variable names should be logical, the code must be commented and it should be written in such a way that it is easy to follow and understand. Figures should have appropriate titles and axis labels and the use of the following commands is recommended: title, xlabel, ylabel, axis. The software implementation should find a solution which minimizes the processing time over a set of equivalent methods. See the Matlab commands: tic, toc, cputime.

Exercise

- Load the dataset "Breast Cancer Wisconsin Original" and/or the dataset "Mammographic Mass" into Matlab/Octave. You can import the dataset from: https://archive.ics.uci.edu/ml/datasets
- Pre-process the dataset converting strings or letters into integer numbers (if needed). Remove instances with missing values.
- Apply measures from information theory to extract any useful information from the dataset.