

Syllabus

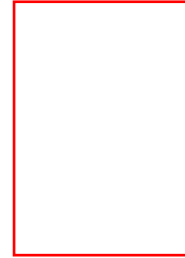
“Geospatial Data Mining” (E-Learning)

English e-learning course, corresponding to the contents of a lecture held at the Institute of Statistics and Information Management, Universidade Nova de Lisboa.

Teacher

Prof. Doutor Fernando Lucas Bação

<http://www.isegi.unl.pt/ensino/docentes/fbacao/index.html>



Goals

The goal is that students completing this course, should be able to:

- Ø Define Data Mining.
- Ø Explain the characteristic features of Data Mining.
- Ø Explain why Data Mining can be a valuable addition in the context of GIScience.
- Ø Analyse the implications of the geo prefix in Geographic Data Mining.
- Ø Understand the basic data preparation and pre-processing tasks.
- Ø Describe the general workings of the Self-Organizing Map.
- Ø Use Self-Organizing Maps in unsupervised classification tasks.
- Ø Describe the general workings of the Multi-layer Perceptron with backpropagation training.
- Ø Describe the general workings of Classification Trees.
- Ø Use Classification Trees and Multi-Layer Perceptron Neural Networks in supervised classification tasks.

Content

Part 1:

The idea of the 1st module of the course is to provide the basic concepts of data mining and knowledge discovery. Emphasis is added to geospatial (or geographical) data mining and to what the geo prefix implies. The student is introduced to the different perspectives from which data mining can be viewed.

Readings List:

1. [Definition of data mining](#), “Knowledge Discovery and Data Mining: towards a unifying framework”, Fayyad, Shapiro and Smyth.

2. [Typical tasks in data mining](#), "Data Mining: Exploiting the Hidden Trends in Your Data", by Herb Edelstein.
3. Geospatial data mining, "Geospatial Data Mining", by Fernando Bação.
4. [Is spatial data special?](#), "On the Particular Characteristics of Spatial Data and its Similarities to Secondary Data Used in Data Mining", Bação, F., Lobo, V., Painho, M., GIS PLANET 2005.
5. [Data Mining: Statistics and More?](#), by David J. Hand
6. [Is inductive machine learning just another wild goose \(or might it lay the golden egg\)?](#) – the perspective from Mark Gahegan

SELF-TEST

Part 2:

In the 2nd part we will deal with the concepts of unsupervised classification. The student starts with a reading about unsupervised classification followed by a small document on the fundamentals of data preparation and pre-processing. Next, two different tools are presented: the k-means algorithm and the self-organizing map. This point is closed with a discussion on profiling and the tools available to explore the unsupervised classification results.

Readings List:

1. Introduction to Unsupervised Classification, from Fernando Bação
2. The fundamentals of clustering, from Statsoft Electronic Book
3. Fundamentals of data preparation and pre-processing, from Fernando Bação
4. The k-means algorithm, from Statsoft Electronic Book
5. The self-organizing map (SOM), by Fernando Bação and Victor Lobo
6. Samuel Kaski and Teuvo Kohonen, "[Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world.](#)" In Apostolos-Paul N. Refenes, Yaser Abu-Mostafa, John Moody, and Andreas Weigend, editors, *Neural Networks in Financial Engineering*, pages 498--507. World Scientific, Singapore, 1996.
7. Result interpretation and profiling, by Fernando Bação

Demos and Applets:

- [SOFM](#) – My favourite demo on the workings of a SOM. Every time I need to explain the SOM, this always seems the easiest way to do it.
- Interactive Self-Organizing Map demonstrations – two applets from the

HUT people in Finland.

- Our own demos, developed by [Roberto Henriques](#), one of our whiz programmers. This is only a movie, in the future we will develop an applet for the internet. The software has been developed for geographic applications, nevertheless it seems to be a great tool to understand the basics of the SOM.

Additional reading

1. Mitchell, T., (1997) Machine Learning, McGraw Hill.
2. Hand, D. J., Mannila, H., Smyth, P. (2001) Principles of Data Mining (Adaptive Computation and Machine Learning), MIT Press.

SELF-TEST

Project 1:

Project 1 will deal with the application of the concepts related with unsupervised classification presented in Part 2. The project consists of two exercises in which the student uses the tools addressed in Part 2 (k-means algorithm and the self-organizing map) to classify data from satellite images. The first exercise is organized in a tutorial fashion and the student just has to follow the steps to achieve the desired result. The second exercise has no instructions and is meant to evaluate the level of understanding and autonomy of the student.

1. Datasets – Exercise 1, Exercise 2
2. Software - [SOM_PAK](#), a very good software package, very efficient and capable of processing very large datasets. It doesn't have a graphical interface, and all interaction is done through DOS command line. This may be frightening for some of you, but let me assure you that after the first shock (and some experiments) it is relatively easy to use. The manual is available [here](#), you should read it in order to be able to complete the proposed exercises.
3. Instructions - Step by step tutorial

Part 3:

In the Part 3 the objective is to study supervised classification methods. The student will

be introduced to typical process involved in supervised classification tasks and to two different types of tools: classification trees and feed-forward neural networks with backpropagation.

Readings List:

1. Introduction to Supervised Classification, from Fernando Bação
2. [Classification Trees](#), from Statsoft Electronic Book
3. [Neural Networks](#), from Statsoft Electronic Book
4. Additional topics on the use of Classification Trees, from Fernando Bação
5. Additional topics on the use of Neural Networks, from Fernando Bação
6. Results assessment, by Fernando Bação

Additional reading

1. Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford: University Press.
2. Fausett, L. (1994). *Fundamentals of Neural Networks*. New York: Prentice Hall.

SELF-TEST

Project 2

Project 2 will deal with the application of the concepts related with supervised classification presented in Part 3 of this course. The project consists of three exercises in which the student uses the tools addressed in Part 3 (classification trees and feed-forward neural networks with backpropagation) to classify data. The first two exercises are organized in a tutorial fashion and the student just has to follow the steps to achieve the desired result. The third exercise has no instructions and is meant to evaluate the level of understanding and autonomy of the student.

1. Datasets – Exercise 1, Exercise 2, Exercise 3
2. Software:
 - [CTree.xls](#) – an Excel based program, from Angshuman Saha, which builds decision trees. This software package was developed essentially as a learning aid and the "*performance is not too bad*". It is easy to use and it only requires the user to have Excel.

- [NNClass.xls](#) – another Excel based program, from Angshuman Saha, which builds neural networks. In his own words “(it) is a very basic implementation of FeedForward - BackPropagation Neural Network, used for prediction and classification problems.”

3. Instructions - Step by step tutorial

Style

- Problem-oriented approach with active knowledge acquisition
- Theory and practical project
- Asynchronous part: self study based on online materials, self-tests at the end of each unit, homework done in groups of 2 students, project.
- Synchronous part: discussion of problems and tasks in 2 synchronous sessions
- Access to teacher via E-Mail
- Students' interaction via forum
- One exam at the end of the course
- Student workload: 90 h (2x30 h), equivalent to 3 credit points (ECTS)

Participants

- Students at
- Knowledge in geoinformatics and statistics is highly recommended

Organization

- Exam January 15, 5:00 p.m.
- Start and end: January 10 - March 31, 2007
- Synchronous sessions:
 1. January 20, 6:00 p.m. GMT (part II)
 2. February 18, 6:00 p.m. GMT (part III)
- Max. number of participants: 20
- Student online activity will be tracked by the platform, completion of self-tests and online questions will be used to assess the progress of the course

Successful participation

- Do homework and project
- Send in your tasks (date to be announced)
- Send in the project (date to be announced)
- Attend synchronous sessions
- Pass exam