Information about the course

Schedule: Tuesday and Thursday: from 8am to 10am

Room: 52 IC3

This course will cover in breadth many algorithms of machine learning/data mining. We will cover, in general terms, of the following machine learning "tasks": data transformation, classification, regression and clustering.

We will cover a variety of algorithms, providing some intuitions on how, why and when they work, but for only a few of them there will be a in depth study of the mathematical formulation and properties. We believe that this breadth approach will allow the student to solve the more common problems they may face in data mining.

The course will be based on the book *Inteligencia Artificial: uma abordagem de aprendizado de maquina* (in Portuguese) by Faceli, Lorena, Gama & Carvalho. There will be also some extra material that complements some of the topics covered in the book.

The course will be evaluated in two forms, to be selected by the student.
• based on 7 individual almost weekly exercises. Theses exercises aim at familiarizing the student with machine learning packages, and procedures.

The exercises must be completed using either R (with a set of packages for machine learning) or Python (with the scikit learn package and maybe others) - details below.

• the first 4 individual almost weekly exercises above, and a more scientific exploration of data mining for a particular data set/problem that you may be interested in. This research project can be done in groups of 1 to 3 students.

There will be two reports of 5 to 10 pages, to be turned in at the mid semester and at the end. The first (due date to be announced) will cover the bibliographic review of the problem you are attacking, as well as some results of the use of the "standard" machine learning algorithms on that dataset.

A final report by the end of the semester, with more detailed comparisons, with the general aim of producing a publishable paper on that problem.

Exercises
1. Exercise 1 in English, in Portuguese, due 18/3. My solution to exercise 1 is here
2. Exercise 2 in English, in Portuguese, due 25/3.

Computer resources

R:

- Available at http://www.r-project.org/
- R packages of interest: caret, e1071, cluster (already a part of base R), fpc

Python:

- Available at: python site. I find the Continuum Anaconda Python distribution a very useful and complete (for ML purposes) Python distribution
- Scikit Learn (already included in the Anaconda distribution including its dependencies). There are two other packages PyBrain and mlpy but I have never used them.

Discussion groups on ML:

- Meta-optimize (very deep low volume)
- Cross Validated (more on statistics but also on ML
- I believe there is something on ML on Quora but I have not check it yet.

Deadlines for the research projects

-
Course materials

- The curse of dimensionality: R program and figure generated