Data Mining CM

Course title - Intitulé du cours	Data Mining CM
Level / Semester - Niveau /semestre	M2 / S1
School - Composante	Ecole d'Economie de Toulouse
Teacher - Enseignant responsable	François BACHOC
Other teacher(s) - Autre(s) enseignant(s)	
Other teacher(s) - Autre(s) enseignant(s)	
Other teacher(s) - Autre(s) enseignant(s)	
Other teacher(s) - Autre(s) enseignant(s)	
Other teacher(s) - Autre(s) enseignant(s)	
Lecture Hours - Volume Horaire CM	20
TA Hours - Volume horaire TD	
TP Hours - Volume horaire TP	14
Course Language - Langue du cours	Anglais
TA and/or TP Language - Langue des TD et/ou TP	Anglais

Teaching staff contacts - Coordonnées de l'équipe pédagogique :

François Bachoc: francois.bachoc@math.univ-toulouse.fr Université Paul Sabatier, batiment IR1, bureau 104. First contact by email or after classes. Meetings can be arranged by email.

Course's Objectives - Objectifs du cours :

The objective of this module is to introduce statistical methods in order to explore data structures. The extraction of information is central to the new challenges introduced by access to ever larger databases and this set of methods finds its interest in studies where potentially large scale data intervenes. In the first part of the course, we introduce the classic optimisation methods that are PCA and its variants (CA, MCA, MDA,...). Our approach is intentionally general in order to illustrate the mathematical principles common to all these tools which could easily be incorporated into a variety of settings. The second part of the course is an overview of methods commonly used in the industry to explore and process data. We talk about decision trees (CART for example), neural network (perceptrons, Kohonen networks), moving averages and ascending hierarchical classification. This class also includes sessions on machines to implement all of these methods through the use of the R software. In addition, during these practicals, we will have the opportunity to introduce some more advanced points such as model selection, cross-validation, bootstrap, ...

Prerequisites - Pré requis :

Common knowledge in applied mathematics.

Practical information about the sessions - Modalités pratiques de gestion du cours :

Computers are accepted. Students should be active during lab sessions. Students should arrive on time, though, exceptionally, late students might be accepted in class.

Grading system - Modalités d'évaluation :

The grade will be entirely constituted by the final exam.

Bibliography/references - Bibliographie/références :

The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Hastie, Tibshirani and Friedman.

Session planning - Planification des séances :

There will be 8 lectures, together with lab sessions (10 hours).

Distance learning – Enseignement à distance :

In case of impossibility of onsite teaching, various online options are possible for the course. In particular, online classes will be held and electronic documents will be made accessible to the students.