**COURSE : Supervised Learning**

INSTRUCTOR : Tommaso Proietti

CONTACTS: Dipartimento di Economia e Finanza, Via Columbia 2. Tel 06 7259 5941

email : tommaso.proietti@uniroma2.it

**COURSE BACKGROUND**

The course provides an introduction to Statistical Learning and Data Mining.

The advances in information technology have made available very rich information data sets, often generated automatically as a by-product of the main institutional activity of a firm or business unit. Most organizations today produce an electronic record of essentially every transaction in which they are involved. Firms collect terabytes data over operating periods (transactions data, e.g. credit cards).

Data Mining deals with inferring and validating patterns, structures and relationships in data, as a tool to support decisions in the business environment.

The course offers an insight into the main statistical methodologies for the visualisation and the analysis of business and market data, providing the information requirements for specific tasks such as credit scoring, prediction and classification. Emphasis will be given to empirical applications using modern software tools.

**LEARNING OBJECTIVES**

The course has the following intended learning outcomes:

* to provide a thorough knowledge of data mining methods and statistical learning techniques;
* to provide the expertise to manage complexity in information and to be able to distill the stylized facts that are relevant for interpretation;
* to be able to predict business outcomes;
* to be able to select a predictive method among those available;
* to be able to communicate the statistical findings to a non expert audience;
* to be able to perform sophisticated statistical analyses with the appropriate software.
* to critically appraise the potential and the limitation of the available methodologies.

Particular attention is dedicated to the ability to communicate the statistical evidence in a systematic and synthetic way, using graphs and summaries, to a non-specialist target audience.

**METHODOLOGY**

The course covers the modern statistical methodologies for the visualization and the analysis of business and market data, which are relevant for making decisions in a complex and rapidly changing business environment.

The fundamental theme is supervised statistical learning, which deals with the prediction of quantitative and qualitative outcomes using a potentially large set of inputs. The two problems, regression and classification, constitute the core of the course.

Emphasis is given to the problem of variable and model selection and on the generalizability of a prediction method outside the training sample, via the optimization of the trade-off between model complexity and the in-sample goodness of fit.

Students develop their learning skills by comparing the teaching material provided by the instructor and exposed in the lectures with the readings suggested with weekly periodicity. The software tutorials and the analysis of cases studies in the assignments will help build their applied skills and their autonomous progress towards the intended learning outcomes.

**EXAM**

The assessment consists of

30% group assignment

70% final written exam

The group assignments aims at assessing the capabilities of processing and analyzing statistical modeling, as well as the ability to communicate the relevant findings. The students are expected to produce a technical report no longer than 8 pages.

The final exam is a written test of 120 minutes containing 3 main questions and 10-15 short questions.

**CONTENTS**

1. Introduction to data mining. Tools for data analysis, visualisation and description.
2. The linear regression model.
3. Model selection and evaluation: bias-variance trade-off, model complexity and goodness of fit. Cross-validation. Selection using information criteria.
4. Regularization and shrinkage methods: rigde regression, lasso, forward stagewise regression.
5. Linear methods for classification: Bayes Classification Rule. Discriminant analysis. Canonical variates.
6. Linear methods for classification: logistic regression.
7. Semiparametric regression: Regression splines and smoothing splines.
8. Kernel smoothing methods: Local polynomial regression.
9. Density estimation. Nearest neighbour classification.
10. Additive Models, tree-based methods. GAM, Regression and classification trees. Boosting.

**TEACHING MATERIAL**

The course material will be made available during the course: slides, suggested readings, datasets, supplementary materials (script of Matlab, R and SAS).

**SUGGESTED READING**

The main reference for the course is G James, D Witten, T Hastie, and R Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, Springer Series in Statistics, 2014. Dowloadable at <http://www-bcf.usc.edu/~gareth/ISL/>

Slides, readings, datasets and supplemental material will be made

available in the course website.

**ADDITIONAL SUGGESTED TEXTBOOKS**

T Hastie, R Tibshirani and J Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, Springer Series in Statistics, 2009.

Website: <http://www-stat.stanford.edu/ElemStatLearn/>