**Statistical Learning**

Prof. Alessia Pini

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

The course aims to provide students with knowledge and understanding of the principal notions of statistical learning, that is the principal techniques for the analysis of complex data. The theoretical presentation of the methods is supported by examples and real applications to Finance and Economics. In addition, students will learn the computational tools to perform and evaluate all presented methods with the R statistical software.

The following learning abilities are provided and expected to be achieved by participants at the end of the course:

1. Knowledge of concepts, terms and methods of statistical learning, which constitute a basis for applying or developing original ideas, often in a research context (DD1 - Knowledge and understanding);
2. Ability to correctly apply methods of statistical learning to real economics and to manage complex problems (DD2 - Applying knowledge and understanding);
3. Quantitative and critical thinking addressed to make independent judgments about the models’ performances on real-world applications, capability of selecting the most adequate model among the studied ones (DD3 - Making judgments).
4. Ability to communicate in a clear and objective way the results of a statistical analysis and quantitatively motivate the choices that were made to specialist and non-specialist audience (DD4 - Communication).
5. Master data analytic study in a manner that may be largely self-directed or autonomous in future careers involving management of data, rigorous reasoning and data-driven decision-making (DD5 - Lifelong learning skills).

***COURSE CONTENT***

*Introduction to statistical learning*. Introduction and description of the principal aims of statistical learning. Difference between supervised and unsupervised methods. Prerequisites: short summary of linear regression and logistic regression.

*Unsupervised learning methods*. Principal component analysis. Clustering methods (k-means and hierarchical clustering). Model-based clustering and EM algorithm.

*Introduction to supervised learning methods*. Bias-variance tradeoff and metrics for model evaluation in regression and classification. Short summary of linear and logistic regression.

*Linear model selection and classification.* Variable selection techniques. Ridge and Lasso regression. Dimensional reduction methods.

*Non-linear regression and classification*. Polynomial regression, regression splines and local regression. Generalized additive models.

*Linear and non-linear classification methods.* Naïve Bayes classifier. K-nearest neighbors, support vector classifiers and support vector machines. Relationship to logistic regression.

*Tree-based methods*. Regression and classification trees. Bagging, random forests and boosting.

*High-Dimensional inference.* Control of the family-wise error rate and false discovery rate. Methods for p-value adjustment.

***READING LIST***

**Main textbooks:**

G. James-D. Wittens-T. Hastie-R. Tibshirani, *An Introduction to Statistical Learning. Springer,* 2013. Chapters 6 - 10. Available at: <http://www-bcf.usc.edu/~gareth/ISL/>

R.A. Johnson-D.W. Wichern,  *Applied Multivariate Statistical Analysis (6th edition). Pearson,* 2002. Chapters 8, 11, 12.

**Further readings:**

T. Hastie-R. Tibshirani-J. Friedman, *The Elements of Statistical Learning (2nd edition). Springer,* 2009. Available at: <https://web.stanford.edu/~hastie/ElemStatLearn/>

C.M. Bishop, *Pattern Recognition and Machine Learning. Springer,* 2006. Available at: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf>

**Further support material** (lecture slides, notes, exercises) will be available on the course’s Blackboard page.

***TEACHING METHOD***

Teaching method will include:

* Frontal lectures,
* PC-labs using the software R,
* Case study discussion.

***ASSESSMENT METHOD AND CRITERIA***

The final grade for this course will be assigned by the evaluation of the following two parts. Is necessary that both grades are sufficient in ordet to pass the exam.

1. **A written exam that accounts for 75% of the final grade**.

The exam will be composed of a set of theoretical questions on the topic of the course and exercises regarding the analysis of a data set. Exercises have to be solved using the software R. Aim of the exam is to assess knowledge, reasoning abilities and analytic abilities on the course subjects.

Students attending class regularly have the possibility to take a midterm partial exam which accounts for 37.5% of the final score. Details will be announced by the lecturer and posted on Blackboard. The second partial exam (37.5% of the final score) will take place together with the first full exam.

1. **A project that accounts for 25% of the final grade**.

Students will be asked to prepare and discuss a project in teams of 2-3 people. The project will regard the analysis of a real data set. The data set to analyze will be discussed with the professors at the end of the first module. Aim of the project is to assess reasoning abilities for a real data analysis as well as communication abilities.

***NOTES AND PREREQUISITES***

Students enrolling in this course are expected to know foundations of algebra, probability, data analysis, statistical inference, applied linear models and computational methods for statistical analyses, i.e., the topics covered in the courses “Mathematical methods and probability”, “Statistical Inference”, “Applied linear models”, and “Computational statistics” taught in the first year. Moreover, students are expected to be familiar with the statistical software R.

 *Office hours*

The teacher receive the students in her personal office (Lanzone 18, third floor) every Tuesday from 14 to 15, after requesting an appointment by email. Note that it will be possible to ask for an online appointment as well. Any changes to that time will be reported on the teacher’s personal web page and on the Blackboard page of the course.