Postgraduate Course
Predictive and Descriptive Learning (MSc)

Instructor Information

Eduardo López Gonzalo

E-mail: eduardo.lopez@upm.es
Work Phone: +34 91 549 57 00 Ext: 4082

Course Information

Course Description

This course covers the concepts and principles of a large variety of Machine Learning methods. The course introduces main principles in Machine Learning: supervised, unsupervised and reinforcement learning, though only predictive and descriptive learning is addressed and reinforcement learning is covered in a subsequent course. Methodological issues such as model assessment and selection, and overfitting are discussed. The predictive or supervised techniques include various types of regression, generalized linear models, k-nearest neighbor classifier, classification and regression trees, ensemble methods (Bagging, Random Forests and Boosting) and kernel methods and Support Vector Machines. The course also addresses descriptive or unsupervised techniques: principal components analysis and clustering methods (k-means and hierarchical clustering). The course places special emphasis on presenting and discussing each technique through the analysis of practical use cases. Complementary use cases and experiments for large-scale scenarios are addressed in the Machine Learning Lab.

Prerequisites

Elementary course in statistics

Previous exposure to a programming language, such as MATLAB, R or Python.

Course Goal

This course covers the principles and methodology for the design, evaluation and selection of a large variety of Machine Learning methods for supervised and unsupervised learning.

Summary of intended course outcomes

The students will understand the fundamentals and important topics in statistical machine learning. This outcome represents a fundamental ingredient in the training of a modern data scientist providing a solid base for its use on a wide range of applications

in science and industry. In particular students will understand the ideas behind the most used and widely applicable techniques for regression, classification and clustering. The course will start presenting simple machine learning techniques in order to grasp the more sophisticated ones. Through several examples and use cases students will also learn how important is to accurately assess the performance of a method. They will also acquire solid criteria on what could be best model for a given data and task. By the end of the course, students should be able to:

- Understand the fundamentals of the most used models and techniques for predictive and descriptive learning.
- Design a proper methodology for accurately assessing and gaining knowledge from the use of each one of the particular machine learning techniques.
- Know the strengths and weaknesses of the various approaches in order to choose the best models for a given data and application scenario.

Syllabus

Introduction to Machine Learning
What and Why?
Statistical Learning
Assessing Model Accuracy

Linear Regression
Simple and Multiple Linear Regression
Linear Regression and Distributed Machine Learning Principles
Interpreting Regression Coefficients
Model Selection and Qualitative Predictors
Interactions and Nonlinearity
Comparison of Linear Regression with KNN

Classification
Logistic Regression
Bayes classifier and Linear Discriminant Analysis
Classification error analysis
Quadratic Discriminant Analysis
K-Nearest Neighbors
A Comparison of Classification Methods: Logistic Regression, LDA, QDA and KNN

Resampling methods
Cross-validation
Bootstrap

Linear Model Selection and Regularization
Feature selection
Optimal Model selection

Regularization
Dimension Reduction
High-Dimensional Data

Moving Beyond Linearity
Polynomial Regression
Step functions
Splines
Local Regression
Generalized Linear Models and Generalized Additive Models

Tree-Based Methods
Decision tress
Bagging
Random Forests
Boosting

Support Vector Machines
Maximal Margin Classifier
Support Vector Classifiers
Kernels and Support Vector Machines
Relationship to Logistic Regression

Descriptive Learning
Supervised vs Unsupervised learning
Principal Components Analysis
Clustering Methods
K-means
Hierarchical Clustering
Practical Issues in Clustering

Textbook:

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Bibliography:

- Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second Edition. Springer Series in Statistics, 2009.

- Konstantinos Koutroumbas, Sergios Theodoridis. *Pattern Recognition*, 4th Edition. Academic Press, 2008.
- Rajaraman, Anand, and Jeffrey D. Ullman. *Mining of massive datasets*. Vol. 77. Cambridge: Cambridge University Press, 2012.
- Bekkerman, Ron, Mikhail Bilenko, and John Langford, eds. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

Student Assessment Criteria

Final Exam	40%
Practical exercises	20%
Group projects	20%
Personal project	20%