AN INTRODUCTION TO MACHINE LEARNING USING STATA

Berlin, 15-17 June 2020

Recent years have witnessed an unprecedented increase in the availability of information on social, economic and health-related phenomena. Today researchers, professionals and policy makers have therefore, access to enormous databases (so-called Big Data) containing detailed information on individuals, companies and institutions and use of mobile devices. Machine learning is a relatively new approach to data analytics, which lies at the intersection between statistics, computer science and artificial intelligence. Its primary objective is that of *turning information into knowledge and value* by "letting the data speak". In contrast to the more tradition approach of data analysis focusing on prior assumptions relating to data structure and the derivation of analytical solutions, Machine Learning techniques rely instead on a model-free philosophy development of algorithms, computational procedures, and graphical inspection of the data in order to more accurately predict outcomes. Computationally infeasible until very recently, Machine Learning is itself a product of the latest advancements in IT technology, of the computing power and the learning capabilities of today's computers, of hardware development, and continuous software development.

This intensive introductory course offers therefore an introduction to the standard machine learning algorithms currently applied to social, economic and public health data in order to illustrate (using a series of both official and user written Stata commands), how Machine Learning techniques can be applied to search for patterns in large (often extremely "noisy") databases, which can subsequently be used to make both decisions and predictions.

In common with TStat's training philosophy, each individual session is composed of both a theoretical component (in which the techniques and underlying principles behind them are explained), and an extensive applied (hands-on) segment, during which participants have the opportunity to implement the techniques using real data under the watchful eye of the course tutor. Throughout the course, theoretical sessions are reinforced by case study examples, in which the course tutor discusses and highlights potential pitfalls and the advantages of individual techniques. The intuition behind the choice and implementation of a specific technique is of the utmost importance. In this manner, the course leader is able to bridge the "often difficult" gap between abstract theoretical methodologies, and the practical issues one encounters when dealing with real data.

At the end of the course, participants are expected to be able to: i) autonomously implement (with the help of the Stata routine templates developed for the course) the appropriate Machine Learning algorithms, given both the nature of their data and the analysis in hand, and ii) to have mastered the concepts of: factor-importance detection, signal-from-noise extraction, correct model specification and model-free classification, from both a data-mining an causal perspective.

https://www.tstattraining.eu/training/machine-learning-stata/

CODE

D-EF36

DATE AND LOCATION

Berlin, 15-17 June 2020

TARGET AUDIENCE

Researchers and professionals working in biostatistics, economics, epidemiology, social and political sciences and public health wishing to implement Machine Learning techniques in Stata.

PREREQUISITES

Participants should be familiar with the statistical software Stata. An introductory knowledge of econometrics and/or statistics is also required.

PROGRAMME | DAY 1

SESSION I: THE BASICS OF MACHINE LEARNING

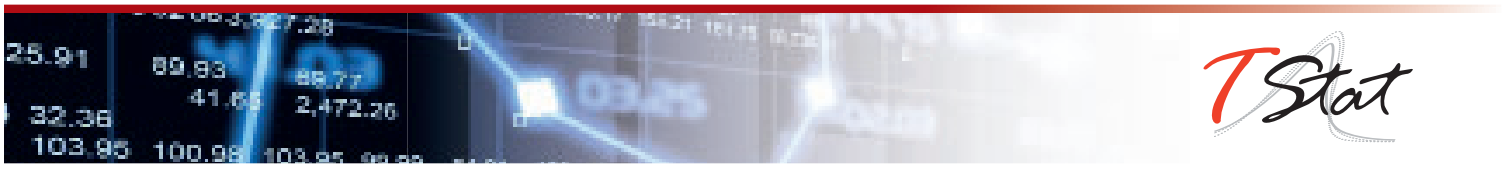
1. *Machine Learning: definition, rational, usefulness*
   - Supervised vs. unsupervised learning
   - Regression vs. classification problems
   - Inference vs. prediction
   - Sampling vs. specification error
2. *Coping with the fundamental non-identifiability of E(y|x)*
   - Parametric vs. non-parametric models
   - The trade-off between prediction accuracy and model interpretability
3. *Goodness-of-fit measures*
   - Measuring the quality of fit: in-sample vs. out-of-sample prediction power
   - The bias-variance trade-off and the Mean Square Error (MSE) minimization
   - Training vs. test mean square error
   - The information criteria approach
4. *Machine Learning and Artificial Intelligence*
5. *The Stata/Python integration: an overview*

SESSION II: RESAMPLING AND VALIDATION METHODS

1. Estimating training and test error
2. *Validation*
   - The validation set approach
   - Training and test mean square error
3. *Cross-Validation*
   - K-fold cross-validation
   - Leave-one-out cross-validation
4. *Bootstrap*
   - The bootstrap algorithm
   - Bootstrap vs. cross-validation for validation purposes

SESSION III: MODEL SELECTION AND REGULARIZATION

1. Model selection as a correct specification procedure
2. The information criteria approach
3. *Subset Selection*
   - Best subset selection
   - Backward stepwise selection
   - Forward stepwise Selection
4. *Shrinkage Methods*
   - Lasso and Ridge, and Elastic regression
   - Adaptive Lasso
   - Information criteria and cross validation for Lasso
5. Stata implementation



DAY 2

SESSION IV: DISCRIMINANT ANALYSIS AND NEAREST-NEIGHBOR CLASSIFICATION

1. The classification setting
2. Bayes optimal classifier and decision boundary
3. Misclassification error rate
4. *Discriminant analysis*
   - Linear and quadratic discriminant analysis
   - Naive Bayes classifier
5. *The K-nearest neighbors classifier*
6. Stata implementation

SESSION V: NONPARAMETRIC REGRESSION

1. Beyond parametric models: an overview
2. Local, semi-global, and global approaches
3. *Local methods*
   - Kernel-based regression
   - Nearest-neighbor regression
4. *Semi-global methods*
   - Constant step-function
   - Piecewise polynomials
   - Spline regression
5. *Global methods*
   - Polynomial and series estimators
   - Partially linear models
   - Generalized additive models
6. Stata implementation

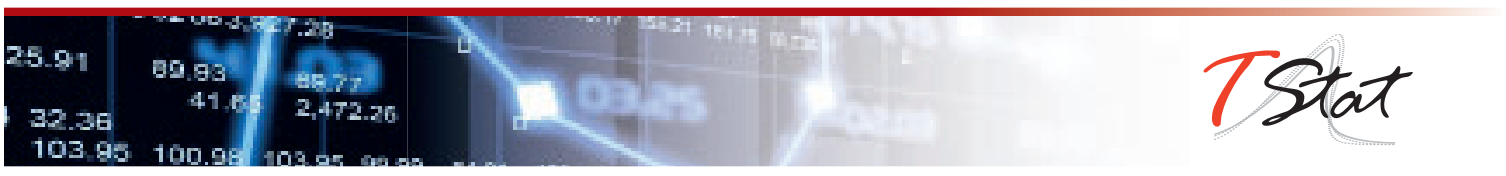
DAY 3

SESSION VI: TREE-BASED REGRESSION

1. Regression and classification trees
   - Growing a tree via recursive binary splitting
   - Optimal tree pruning via cross-validation
2. Tree-based ensemble methods
   - Bagging, Random Forests, and Boosting
3. Stata implementation

SESSION VII: NEURAL NETWORKS

1. *The neural network model*
   - neurons, hidden layers, and multi-outcomes
2. *Training a neural networks*
   - Back-propagation via gradient descent
   - Fitting with high dimensional data
   - Fitting remarks
3. *Cross-validating neural network hyperparameters*
4. Stata implementation



COURSE REFERENCES

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013), An Introduction to Statistical Learning with Applications in R, Springer, New York, 2013.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2008), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second edition, Springer.

REGISTRATION FEES

Full-Time Students*: € 1155.00
Academic: € 1605.00
Commercial: € 2110.00

*To be eligible for student prices, participants must provide proof of their full-time student status for the current academic year.

Fees are subject to VAT (applied at the current Italian rate of 22%). Under current EU fiscal regulations, VAT will not however applied to companies, Institutions or Universities providing a valid tax registration number.

Please note that a non-refundable deposit of €100.00 for students and €250.00 for Academic and Commercial participants, is required to secure a place and is payable upon registration. The number of participants is limited to 10. Places, will be allocated on a first come, first serve basis. The course will be officially confirmed, when at least 5 individuals are enrolled.

Course fees cover: teaching materials (handouts, Stata do-files, program templates and datasets to use during the course), a temporary course licence of Stata valid for 30 days from the beginning of the course, light lunch and coffee breaks.

To maximize the usefulness of this course, we strongly recommend that participants bring their own laptops with them, to enable them to actively participate in the empirical sessions.

Individuals interested in attending this course must return their completed registration forms by email (training@tstat.eu) to TStat by the 26st May 2020.

Further details regarding our registration procedures, including our commercial terms and conditions, can be found at https://www.tstattraining.eu/training/machine-learning-stata/

CONTACTS

Monica Gianni
TStat S.r.l. | Via Rettangolo, 12-14
I-67039 Sulmona (AQ)
T. +39 0864 210101

TStat Training | Kleebergstraße, 8
D-60322 Frankfurt am Main

formazione@tstat.it

www.tstat.it
www.tstattraining.eu