Financial Intermediation Network of European Studies

W I N T E R   S C H O O L

Machine Learning using STATA

Module outline

Dr. Giovanni Cerulli
21 – 22 December 2020
University of Roma Tre, Italy

Organizer

Prof. Ornella Ricci
Professor of Banking and Finance, University of Rome Tre, Italy

Lecturer

Dr. Giovanni Cerulli
Research Institute on Sustainable Economic Growth, National Research Council of Italy, Italy
giovanni.cerulli@ircres.cnr.it

Short bio

Giovanni Cerulli holds a degree in Statistics and Economics, and a PhD in Economics (both at the Sapienza University of Rome). His research interests are about statistic and econometric modeling, with a special focus on the econometrics of program evaluation and causal inference. Also, he does research on "machine learning" and computational econometrics. Giovanni developed some original models for quantitative program evaluation, such as models for continuous treatment, neighborhood interaction, as well as nonparametric extensions of existing models. In applications, my research activity has focused mainly on measuring the effects of technological policies on firm economic and technological performance. He also covered applications in banking and finance.

As of May 2019, the Research Papers in Economics (RePEc) archive ranks his profile among the top 5% authors. Furthermore, RePEc ranks his profile in the 26th position among the Italian economists for the scientific production of the last 10 years.

Giovanni is Editor-in-Chief of the International Journal of Computational Economics and Econometrics, and an associate editor of the World Review of Science, Technology and Sustainable Development. He is also coordinator of GRAPE - Research Group on the Analysis of Economic Policies. Giovanni also authored the book: Econometric Evaluation of Socio-Economic Programs: Theory and Applications (Springer, 2015).

When

21, 22 December 2020

Where

Microsoft Teams

Overall aims and purpose

The overall aim is to provide an introduction to machine learning models and to show how these can be applied for empirical research to banking data using Stata.

Course description

Recent years have witnessed an unprecedented availability of information on social, economic, and health-related phenomena. Researchers, practitioners, and policymakers have nowadays access to huge datasets (the so-called "Big Data") on people, companies and institutions, web and mobile devices, satellites, etc., at increasing speed and detail.

Machine learning is a relatively new approach to data analytics, which places itself in 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". To this purpose, machine learning limits prior assumptions on data structure, and relies on a *model-free* philosophy supporting algorithm development, computational procedures, and graphical inspection more than tight assumptions, algebraic development, and analytical solutions. Computationally unfeasible few years ago, machine learning is a product of the computer's era, of today machines' computing power and ability to learn, of hardware development, and continuous software upgrading.

This course is a primer to machine learning techniques using Stata. Stata owns today various packages to perform machine learning which are however poorly known to many Stata users. This course fills this gap by making participants familiar with (and knowledgeable of) Stata potential to draw knowledge and value form row, large, and possibly noisy data. The teaching approach will be mainly based on the graphical language and intuition more than on algebra. The training will make use of instructional as well as real-world examples, and will balance evenly theory and practical sessions.

After the course, participants are expected to have an improved understanding of Stata potential to perform marching learning, thus becoming able to master research tasks including, among others: (i) factor-importance detection, (ii) signal-from-noise extraction, (iii) correct model specification, (iv) model-free classification, both from a data-mining and a causal perspective.

Teaching and learning strategy

The course will last for two full days (9.00-17.00)

Pre-requisites

A good knowledge of panel data.
A good knowledge of Stata.
Students need to have their own laptop with Stata (version 15 or 16).

DAY 1
(21 December 2020)

1. The basics of Machine Learning

Machine Learning: definition, rational, usefulness

Supervised vs. unsupervised learning
Regression vs. classification problems
Inference vs. prediction
Sampling vs. specification error

Coping with the fundamental non-identifiability of $E(y|x)$

Parametric vs. non-parametric models
The trade-off between prediction accuracy and model interpretability

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

Machine Learning and Artificial Intelligence

The Stata/Python integration: an overview

2. Resampling and validation methods

Estimating training and test error

Validation

The validation set approach
Training and test mean square error

Cross-Validation

K-fold cross-validation
Leave-one-out cross-validation

Bootstrap

The bootstrap algorithm
Bootstrap vs. cross-validation for validation purposes

3. Model Selection and regularization

Model selection as a correct specification procedure
The information criteria approach

Subset Selection

Best subset selection
Backward stepwise selection
Forward stepwise Selection

Shrinkage Methods

Lasso and Ridge, and Elastic regression
Adaptive Lasso
Information criteria and cross validation for Lasso

Stata implementation

4. Discriminant analysis and nearest-neighbor classification

The classification setting
Bayes optimal classifier and decision boundary

Misclassification error rate
Discriminant analysis
- Linear and quadratic discriminant analysis
- Naive Bayes classifier

The K-nearest neighbors classifier
Stata implementation

DAY 2
(22 December 2020)

5. Nonparametric regression
Beyond parametric models: an overview
Local, semi-global, and global approaches
Local methods
- Kernel-based regression
- Nearest-neighbor regression

Semi-global methods
- Constant step-function
- Piecewise polynomials
- Spline regression

Global methods
- Polynomial and series estimators
- Partially linear models
- Generalized additive models

Stata implementation

6. Tree-based regression
Regression and classification trees
- Growing a tree via recursive binary splitting
- Optimal tree pruning via cross-validation

Tree-based ensemble methods
- Bagging, Random Forests, and Boosting

Stata implementation

7. Neural networks
The neural network model
- neurons, hidden layers, and multi-outcomes

Training a neural networks
- Back-propagation via gradient descent
- Fitting with high dimensional data
- Fitting remarks

Cross-validating neural network hyperparameters
Stata implementation