Machine learning

Clément Gorin
University of Toronto
gorinclem@gmail.com

Course's objective

Many important economic questions remain unanswered, partly because the data necessary to address them is encoded into high-dimensional data structures such as text or images. Applied economists have become increasingly interested in using machine learning models to transform these data into simpler representations, which can be used as inputs for subsequent economic analysis. After introducing statistical learning, this course provides a comprehensive understanding of some of the most capable supervised learning models including random forest, gradient boosted trees, and specific neural network architectures to make sense of these complex forms of data and perform original economic analysis. From a strong base in theory and mathematical formalisation, focus is kept on intuition and effective implementation using Python, both to illustrate abstract statistical concepts using simulated data and to implement the models studied in class.

Outline

The course is organised in four three-hours sessions. Each session articulates a theoretical lecture and practical applications using Python. The sessions are organised as follows:

Lecture 1: Statistical learning
Function approximation – Inference and prediction – Bias and variance – Resampling – Generalised additive models

Lecture 2: Trees and ensembles
Prediction trees – Estimation – Pruning – Random forest – Gradient boosting – Application

Lecture 3: Neural networks
Units and layers – Estimation – Back-propagation – Better optimisation – Representations

Lecture 4: Image and text data
Image processing – Convolutional networks – Text processing – Embeddings networks – Recurrent networks

Course materials

Lectures, papers exercises, solutions and resources will be made available on the Dropbox of the course, which is updated before every lecture. Students should bring their personal computer with administrator rights and a working internet connection. For the practice sessions, we use the Miniconda or Anaconda distribution of Python, which provides basic packages and simplify management. A Conda environment containing the necessary packages will be provided before the class.

Grading

Students will be graded on the basis of a final assessment. They will have to hand in a short report (10 pages maximum) along with a script reproducing the results. The report must implement machine learning methods to shed light on a research question of interest to economists. Students are encouraged to work on a question related to their PhD thesis.

Handbooks

- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2009.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.
- Nielsen, Michael. *Neural Networks and Deep Learning*. Determination Press, 2019.

Papers

- Breiman, Leo. *Statistical modeling: The two cultures*. Statistical Science, 16(3): 199-231, 2001
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. *Deep learning*. Nature, 521: 436-444, 2015.
- Natekin, Alexey and Alois Knoll. *Gradient boosting machines, a tutorial*. Frontiers in neurorobotics, 7(21), 2013.
- Sendhil, Mullainathan and Spiess Jann. *Machine learning: An applied econometric approach*. Journal of Economic Perspective, 31(2): 87-106, 2017.
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. *Text as data*. Journal of Economic Literature, 57(3): 535–574, 2019.
- Lones, Michael A. *How to avoid machine learning pitfalls*: A guide for academic researchers. *CoRR, 2021.*

Specific references will be given in class.