Machine Learning for Social Sciences

Reto Wüest
University of Geneva

Instructor Biography

Reto Wüest is a postdoctoral researcher in the Department of Political Science and International Relations at the University of Geneva. After pre-doctoral fellowships at Princeton University and New York University, he received his PhD from the University of Geneva in 2016. He studies legislative behavior and political representation using quantitative methods and machine learning techniques. His research has been published in *West European Politics* and *Swiss Political Science Review*.

Course Description

With ever more data available in electronic form, automated methods of data analysis become increasingly important also in the social sciences. Machine learning refers to a set of methods that can automatically detect patterns in data, or "learn" from data. The uncovered patterns can then be used by the analyst to make accurate predictions and decisions under uncertainty.

This course will introduce participants to the fundamentals of machine learning. Students will leave the course with a thorough understanding of the core issues in machine learning (prediction and inference, supervised and unsupervised learning, overfitting, bias-variance trade-off), knowledge of some of the most widely used machine learning methods, and the ability to apply these methods in their own research. All course materials are available at http://retowuest.net/recsm-2018/.

Software

The course will use the open-source software R, which is freely available for download at https://www.r-project.org/. We will interact with R through the user interface RStudio, which can be downloaded at https://www.rstudio.com/products/rstudio/download/.

Prerequisites

Participants are expected to have a solid understanding of linear and binary regression models. The course will also assume at least a basic familiarity with the R statistical programming language.

Schedule

Session 1: Introduction to Machine Learning

(July 2, 2018, 9:00-13:00)

The first session will provide an introduction to machine learning. We will discuss the goals of machine learning (prediction, inference, or both), the difference between supervised and unsupervised machine learning, the problem of overfitting, and the bias-variance trade-off. We will then get to know a first class of important supervised learning methods, namely shrinkage methods (ridge regression and the lasso).

Class Schedule

Time	Topic
09:00-09:30	Introductions and course overview
09:30-10:00	General introduction to machine learning (prediction and inference, supervised and unsupervised learning)
10:00-10:45	Assessing model accuracy (overfitting, bias-variance trade-off, cross-validation)
10:45-11:15	Break
11:15-11:45	Shrinkage methods I: ridge regression
11:45-12:15	Shrinkage methods II: the lasso
12:15-13:00	Application of ridge regression and the lasso

Main Readings

- James et al., *An Introduction to Statistical Learning*, ch. 2 (pp. 15-42) and ch. 6 (pp. 214-228)

Recommended Readings

- James et al., *An Introduction to Statistical Learning*, ch. 5 (pp. 175-186)
- Hastie et al., *The Elements of Statistical Learning*, ch. 3 and ch. 7
- Shalev-Shwartz and Ben-David, *Understanding Machine Learning*, ch. 5 and ch. 13
- Bishop, *Pattern Recognition and Machine Learning*, ch. 12

Session 2: Classification and Regression Trees (CART)

(July 3, 2018, 9:00-13:00)

The second session will deal with tree-based methods, which are another important and highly flexible class of supervised learning methods. After an introduction to the basics of decision trees and a general discussion of the advantages and disadvantages of tree-based models, we will look at three specific widely-used tree-based methods: bagging, random forests, and boosting.

Class Schedule

Time	Topic
09:00-09:30	Introduction to classification and regression trees
09:30-10:00	Advantages and disadvantages of trees
10:00-10:45	Bagging, random forests
10:45-11:15	Break
11:15-12:00	Boosting
12:00-12:30	Application I: regression and classification trees
12:30-13:00	Application II: bagging, random forests, boosting

Main Readings

- James et al., *An Introduction to Statistical Learning*, ch. 8

Recommended Readings

- Hastie et al., *The Elements of Statistical Learning*, ch. 9, ch. 10, and ch. 15
- Shalev-Shwartz and Ben-David, *Understanding Machine Learning*, ch. 18
- Lantz, *Machine Learning with R*, ch. 11

Session 3: Unsupervised Learning

(July 4, 2018, 9:00-13:00)

In the third session, we will move to unsupervised machine learning methods. We will cover two important unsupervised learning techniques: principal components analysis (PCA) and clustering analysis (K-means clustering and hierarchical clustering).

Class Schedule

Time	Topic
09:00-09:30	Introduction to unsupervised learning
09:30-10:15	Principal components analysis (PCA)
10:15-10:45	K-means clustering
10:45-11:15	Break
11:15-12:00	Hierarchical clustering
12:00-12:30	Application I: PCA
12:30-13:00	Application II: clustering methods

Main Readings

- James et al., *An Introduction to Statistical Learning*, ch. 10

Recommended Readings

- Hastie et al., *The Elements of Statistical Learning*, ch. 14
- Shalev-Shwartz and Ben-David, *Understanding Machine Learning*, ch. 22 and ch. 23
- Bishop, *Pattern Recognition and Machine Learning*, ch. 12
- Barber, *Bayesian Reasoning and Machine Learning*, ch. 15
- Lantz, *Machine Learning with R*, ch. 9

References

Barber, David. 2016. *Bayesian Reasoning and Machine Learning.* New York: Cambridge University Press. Available for free as a PDF.
URL: *http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.HomePage*

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning.* New York: Springer.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer. Available for free as a PDF.
URL: *https://web.stanford.edu/ hastie/ElemStatLearn/*

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R.* New York: Springer. Available for free as a PDF.
URL: *http://www-bcf.usc.edu/ gareth/ISL/*

Lantz, Brett. 2015. *Machine Learning with R.* 2nd ed. Birmingham: Packt Publishing.

Shalev-Shwartz, Shai and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms.* New York: Cambridge University Press. Available for free as a PDF.
URL: *http://www.cs.huji.ac.il/ shais/UnderstandingMachineLearning/*