A SHORT COURSE IN DATA MINING WITH APPLICATIONS TO PUBLIC POLICY

Institute for Capacity Development - International Monetary Fund

July 6, 2018

Version 1

Professor: Alvaro J. Riascos Villegas
Contact e-mail: ariascos@uniandes.edu.co
Personal website: www.alvaroriascos.com
Course website: http://www.alvaroriascos.com/data-mining-public-policy/

1. Objectives

This is a super short version of my data mining courses with applications taught in the last 5 years to master students and advanced undergraduate students of economics, mathematics and engineering at the University of los Andes (http://www.alvaroriascos.com/teaching/mineriadatos/). It also leverages on several applied courses to industry audiences. This version of the course places special emphasis on applications to public policy and economics. The course introduces participants to the central theoretical pillars of statistical learning theory as a theoretical framework for data mining: the learning problem, the compromise between bias and variance, approximation and error, risk, consistency, regularization, complexity, etc. At the same time, it introduces the main techniques of data mining: nearest neighborhood method, trees, random forests, boosting, support vector machines, neural networks, cross validation, clustering, k-means, association rules and text mining with a selected set of applications to public policy issues: public health policy, crime prediction, forecasting of macroeconomic variables, fraud detection, macroeconomic uncertainty quantification, etc.

The course is designed for a single session of 6 hours.

2. Program

Time	Topic
9:30AM - 9:50AM (20 minutes)	Data Mining: The Scientific and Industry Revolution
9:50AM - 10:30AM (40 minutes)	Statistical Learning: Models, Concepts, Fundamental Results, Prediction vrs. Causality
10:30AM – 11:00 AM (30 minutes)	KNN, Linear Methods and Regularization **Application: Crime prediction**
11AM – 11:15AM (15 minutes)	Coffee break
11:15AM – 12:00 PM (45 minutes)	Trees, Random Forests, Boosting Model Selection and Validation
12:00PM – 12:30PM (30 minutes)	**Application: Public health**
12:30PM -2PM	Lunch
2:00PM – 2:30PM (30 minutes)	Special techniques: Cross Validation, (Sub) Bagging, Bootstrapping.
2:30PM – 3:30PM (60 minutes)	Text Mining: Document vectorization, Word2Vec, Glove, LDA. **Application: Effects FOMC Communications and Economic Policy Uncertainty**
3:30PM - 3:45PM (15 minutes)	Coffee break
3:45PM – 4:15PM (30 minutes)	Unsupervised learning: Clustering, K-means, associative rules. **Application: Fraud detection**
4:15PM –	Advanced topics: Neural Networks and Deep

5:0PM (45 minutes)	Learning **Application: Forecasting inflation, unemployment and poverty characterization**

3. **Readings**

The first set of references is really the minimum, from the perspective of this course, to get a good and founded idea of what data mining, big data and/or machine lealvaarning is about. They are mostly non technical

The absolute minimum

Prediction Policy Problems. Jon Kleinberg. Jens Ludwig. Sendhil Mullainathan. Ziad Obermeyer. American Economic Review. Vol. 105, NO. 5, May 2015. (pp. 491-95).

McKinsey Global Institute: The age of analytics executive summary

Big Data: New Tricks for Econometrics. Hal R. Varian. Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 3–28.

Statistical Modeling: The Two Cultures Leo Breiman. Statistical Science, Vol. 16, No. 3. (Aug., 2001), pp. 199-215.

Economics in the age of big data. Liran Einav and Jonathan Levin. Science 346 , (2014).

Presentations references

Theory

Bishop. Pattern Recognition and Machine Learning. Springer.

Luxburg, U., B. Scholkopf. 2008. Statistical Learning Theory: Models, Concepts and Results. http://arxiv.org/abs/0810.4752

Introduction to Statistical Learning with Applications in R.
http://www-bcf.usc.edu/~gareth/ISL/

Hastie, T., Tibshirani, R. y J. Hastie. 2009. The Elements of Statistical Learning: Data Minning, Inference and Prediction. Segunda Edición. Springer. http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

Applications

Andrés Azqueta-Gavaldón. Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. Economics Letters 158 (2017) 47–50.

Bayesian Variable Selection for Nowcasting Economic Time Series. Steven L. Scott, Hal R. Varian. http://www.nber.org/chapters/c12995

Predicting the Present with Bayesian Structural Time Series. Steven L. Scott. Hal Varian. 2013.

Predicting the Present with Google Trends. Hyunyoung Choi, Hal Varian. December 18, 2011.

Predicting Initial Claims for Unemployment Benefits. Hyunyoung Choi, Hal Varian. July 5, 2009.

The Billion Prices Project: Using Online Prices for Measurement and Research. Alberto Cavallo and Roberto Rigobon. Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178.

Combining satellite imagery and machine learning to predict poverty. Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, Stefano Ermon. Science 2016 . Vol 353 Issue 6301.

The Effects of the Content of FOMC Communications on US Treasury Rates. Christopher Rohlfs, Sunandan Chakraborty, Lakshminarayanan Subramanian.

Machine Learning: An Applied Econometric Approach Sendhil Mullainathan and Jann Spiess. Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106.

Machine Learning Methods for Demand Estimation. Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. American Economic Review: Papers & Proceedings 2015, 105(5): 481–485.