

Banco Central del Uruguay

# Big Data, Aprendizaje y Minería de Datos

## Herramientas para Economistas

Walter Sosa Escudero

Este curso intensivo plantea una introducción al paradigma de aprendizaje automático y sus vínculos con la visión tradicional de la estadística y la revolución de *big data*. De carácter teórico, práctico y conceptual, el curso presenta una variedad de técnicas, discute casos recientes e incluye una introducción a R, en donde cada método es implementado usando datos reales.

**Destinatarios:** economistas y profesionales con formación similar, interesados en el análisis moderno de datos.

**Requisitos:** conocimientos básicos de modelos de regresión similares a los de un curso de econometría de pregrado.

**Duración y dinámica:** 12 horas, repartidos en dos días. Incluye teoría, discusión de casos y aplicaciones en R.

**Créditos:** 2.

### Temario

1. **Conceptos básicos:** Introducción al paradigma de ciencia de datos y sus relaciones con los roles estándar de la estadística y la econometría.
2. **Regresión para la predicción:** modelos lineales y linealizables. Polinomios, bases.
3. **Complejidad y regresión:** El *trade-off* sesgo y varianza. Overfit. Validación cruzada. Hiperparámetros y aprendizaje. LASSO y ridge.
4. **Clasificación:** Regresión logística y vecinos cercanos. Análisis ROC. Árboles decisorios. Bosques aleatorios. Bootstrap, bagging y boosting.
5. **Aprendizaje no supervisado:** Dimensionalidad. Componentes principales y factores. Clusters jerárquicos y no jerárquicos.
6. **Temas adicionales:** breve introducción a support vector machines, redes neuronales y deep learning.

**Walter Sosa Escudero:** especialista en econometría y estadística aplicada a cuestiones sociales. Sus trabajos de investigación se encuentran en publicaciones nacionales e internacionales incluyendo *Journal of Econometrics*, *Journal of Multivariate Analysis* y *Econometric Theory*. A su labor de investigación y docencia agrega una activa agenda de divulgación científica. Fue orador invitado en TEDXRioDeLaPlata, es autor de *Que es (y que no es) la Estadística (Siglo XXI, Colección Ciencia Que Ladra)* y escribe frecuentemente para los principales diarios argentinos. Ha recibido los premios Fulvio Pagani/ARCOR, Bernardo Houssay, Konex, el Premio EGSO como el mejor profesor del PhD en economía

de la Universidad de Illinois, entre otros. Fue presidente de la Asociación Argentina de Economía, es investigador principal del CONICET y miembro titular de la Academia Nacional de Ciencias Económicas.

## **Bibliografía**

### **Libros**

Ahumada, H., Gabrielli, F., Herrera, M. y Sosa Escudero, W., 2018, *Una Nueva Econometría: Automatización, Big Data, Econometría Espacial y Estructural*, EdiUNS, Buenos Aires (en prensa).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Murphy, K., (2012). *Machine learning: a probabilistic perspective*, MIT Press, Cambridge.

### **Artículos científicos**

Anastasopoulos, J., Badani, D., Lee, C., Ginosar, S. & Williams, J. R. (2018). "Political image analysis with deep neural networks". (Submitted).

Anselin, L., & Williams, S. (2015). Digital neighborhoods. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 1-24.

Askatas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2), 107-120.

Athey, S., & Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050, 5.

Athey, S. (2015, August). Machine Learning and Causal Inference for Policy Evaluation. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 5-6). ACM.

Bai, J. & Ng, S. (2008). "Forecasting economic time series using targeted predictors", *Journal of Econometrics*, vol. 146(2), pp. 304-317.

Baylé, Federico (2016) "Detección de villas y asentamientos informales en el partido de La Matanza mediante teledetección y sistemas de información geográfica" Tesis de Maestría. <https://drive.google.com/file/d/0ByPgZ6LNcIcGNW05YVNNMDVqOTA/view>

Belloni, V. Chernozhukov, C. Hansen: "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28 (2), Spring 2014, 29-50. <https://www.aeaweb.org/articles?id=10.1257/jep.28.2.29>

Chernozhukov, Victor, et al. "Double machine learning for treatment and causal parameters." arXiv preprint arXiv:1608.00060 (2016). <https://arxiv.org/pdf/1608.00060.pdf>

- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073-1076.
- Breiman, L. (2003). Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1), 81-82.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *The American Economic Review*, 105(5), 481-485.
- Calude, C. S., & Longo, G. (2016). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, 1-18.
- Caruso, G., Scartascini, C., & Tommasi, M. (2015). Are we all playing the same game? The economic effects of constitutions depend on the degree of institutionalization. *European Journal of Political Economy*, 38, 212-228.
- Caruso, G., Sosa-Escudero, W., & Svarc, M. (2015). Deprivation and the dimensionality of welfare: a variable-selection cluster-analysis approach. *Review of Income and Wealth*, 61(4), 702-722.
- Cavallo, A. (2013). Online and official price indexes: measuring Argentina's inflation. *Journal of Monetary Economics*, 60(2), 152-165.
- Cavallo, A. (2015). *Scraped data and sticky prices* (No. w21490). National Bureau of Economic Research.
- Cavallo, A. "Are Online and Offline Prices Similar? Evidence from Multi-Channel Retailers" *American Economic Review*- January 2017 - Vol 107 (1). [http://www.mit.edu/~afc/papers/Cavallo\\_Online\\_Offline.pdf](http://www.mit.edu/~afc/papers/Cavallo_Online_Offline.pdf)
- De Mol, C., Giannone, D. & Reichlin, L. (2008). "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?", *Journal of Econometrics*, Elsevier, vol. 146(2), pages 318-328.
- Donaldson, D. & Storeygard, A. (2016). "The View from Above: Applications of Satellite Data in Economics", *Journal of Economic Perspectives*, vol. 30(4), pp. 171-198.
- Einav, L., Knoepfle, D., Levin, J., & Sundaresan, N. (2014). Sales taxes and internet commerce. *The American Economic Review*, 104(1), 1-26.
- Gilchrist, D.S. & Sands, E. G. (2016). "Something to Talk About: Social Spillovers in Movie Consumption", *Journal of Political Economy*. vol. 24(105), pp. 1339-1382.
- Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S.; Brammer, Lynnette; Smolinski, Mark S.; Brilliant, Larry (19 February 2009). "Detecting influenza epidemics using search engine query data". *Nature*. 457 (7232): 1012-1014.
- Guvenen, F., Kaplan, G., & Song, J. (2014). How risky are recessions for top earners?. *The American Economic Review*, 104(5), 148-153.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. & Hadsell, R. (2017). "Overcoming catastrophic forgetting in neural networks", *PNAS*, vol. 114(13), pp. 3521-3526.

Leak, A. & Lansley, G. (2018). "Geotemporal Twitter Demographics", *Consumer Data Research*, capítulo 11, UCL Press.

Linden, A. & Yarnold, P. R. (2016). "Combining machine learning and matching techniques to improve causal inference in program evaluation", *J Eval Clin Pract.*, vol. 22(6), pp.:864-870.

Mittal, M., Mohan, L. & Hemanth, J. (2018). "Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning", *Computational Economics*, pp. 1-19.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106. <https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87>

Nickerson, D., & Rogers, T. (2014). "Political Campaigns and Big Data", *Journal of Economic Perspectives*, vol. 28(2), pp. 51-74.

Keely, L. C., & Tan, C. M. (2008). Understanding preferences for income redistribution. *Journal of Public Economics*, 92(5), 944-961.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *The American Economic Review*, 105(5), 491-495.

Kreiner, C. T., Leth-Petersen, S., & Skov, P. E. (2014). Year-end tax planning of top management: Evidence from high-frequency payroll data. *The American Economic Review*, 104(5), 154-158.

Radinsky, K., Davidovich, S. & Markovitch, S. (2012). "Learning causality for news events prediction".

Wager, S., & Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*.

### **Surveys y artículos de divulgación**

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of Economic Perspectives*, 24(2), 3-30.

Anderson, C. (2008). The end of theory. *Wired magazine*, 16(7), 16-07.

Aromí, D. (2016) Sobre árboles, bosques aleatorios y crisis de deuda soberana. *Alquimias Económicas Blog*.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324), 483-485. <http://science.sciencemag.org/content/sci/355/6324/483.full.pdf>

Biuk-Aghai, R. P., Kou, W. T., & Fong, S. (2016, May). Big data analytics for transportation: Problems and prospects for its application in China. In *2016 IEEE Region 10 Symposium (TENSYMP)* (pp. 173-178). IEEE.

Booth, Adrian; Mohr, Niko y Peters, Peter (2016) "The Digital utility: New opportunities and challenges".

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. *The Journal of Economic Perspectives*, 30(2), 151-178.

Einav, L., & Levin, J. D. (2013). *The data revolution and economic analysis* (No. w19035). National Bureau of Economic Research.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210), 1243089.

Fan, J. (2013). Features of big data and sparsest solution in high confidence set. *Past, present, and future of statistical science*, 507-523.

Glaeser, Edward, Andrew Hillis, Scott Duke Kominers, and Michael Luca. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy." *American Economic Review: Papers and Proceedings* (forthcoming). <http://www.nber.org/papers/w22124.pdf>

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how?. *Journal of Economic Literature*, 51(1), 162-172.

Heffetz, O., & Ligett, K. (2014). Privacy and data-based research. *The Journal of Economic Perspectives*, 28(2), 75-98.

Jeske, M., Grüner, M., & Weiß, F. (2013). BIG DATA IN LOGISTICS: A DHL perspective on how to move beyond the hype. *DHL Customer Solutions & Innovation*, 12.

Lane, J. (2016). BIG DATA FOR PUBLIC POLICY: THE QUADRUPLE HELIX. *Journal of Policy Analysis and Management*, 35(3), 708-715.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, 343(6176), 1203-1205.

Lazer, W. & Kennedy, R.. (2015). What We Can Learn From the Epic Failure of Google Flu Trends, *Wired*, 10.01.15.

Lohr, Steve. (2014) Google Flu Trends: The Limits of Big Data. *The New York Times*.

Manyika, J., Lund, S., Bughin, J., Woetzel, J., Stamenov, K., Dhingra, D., ... & Al-Jaghoub, S. (2016). Digital globalization: The new era of global flows. *McKinsey Global Institute, February*.

Marcus, G. & Davis, E. Eight (No, Nine!) Problems With Big Data. (2014) *The New York Times*.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. *Harvard Bus Rev*, 90(10), 61-67.

NewVantage Partners (2016) *Big Data Executive Survey 2016. An Update on the Adoption of Big Data in the Fortune 1000*.

Riggins, F. J., & Wamba, S. F. (2015, January). Research directions on the adoption, usage, and impact of the internet of things through the use of big data analytics. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 1531-1540). IEEE.

Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. *Nature neuroscience*, 17(11), 1440-1441.

Sharda, R., Delen, D., & Turban, E. (2013). *Business Intelligence: A Managerial Perspective on Analytics*. Prentice Hall Press.

Sosa Escudero, W. (2014). Big data: otra vez arroz?, *Diario Clarin*, 6/4/2014.

Sosa Escudero, W. (2016). Al infinito y más allá: Funes, Borges y big data, *Diario La Nacion*, 12/6/2016.

Sosa Escudero, W. (2017). Big data y aprendizaje automático: Ideas y desafíos para economistas, mimeo.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?. *Big Data & Society*, 1(2), 2053951714536877.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2), 3-27.

## Videos

Susan Athey, Guido Imbens and NBER Organizers. Summer Institute 2015 Methods Lectures, July 18, 2015 [http://www.nber.org/econometrics\\_minicourse\\_2015/](http://www.nber.org/econometrics_minicourse_2015/)

Hal R. Varian, Susan Athey and Larry Wasserman and University of Chicago Organizers. "How Big Data is Changing Economies" April 10, 2015 <https://bfi.uchicago.edu/events/how-big-data-changing-economies>

World Economic Forum. Imagine you could measure supply and demand from space. Satellite imagery is being used to help track poverty. <https://www.facebook.com/worldeconomicforum/videos/10153680368831479/>

Tim Harford, The Big Data Trap <https://www.youtube.com/watch?v=0cizsKDn3TI>

Phil Evans, How data will transform business [https://www.ted.com/talks/philip\\_evans\\_how\\_data\\_will\\_transform\\_business?language=es](https://www.ted.com/talks/philip_evans_how_data_will_transform_business?language=es)

## Método de Evaluación

La evaluación del curso consistirá en un trabajo final aplicado utilizando las herramientas presentadas en el curso.