Big Data, Aprendizaje y Minería de Datos
Perspectivas, ideas y herramientas para economistas

Walter Sosa Escudero
wsosa@udesa.edu.ar
Asistente: Florencia Hnilo
fhnilo@udesa.edu.ar

Objetivos: quizás el título de este curso sea el primer acercamiento a la problemática del mismo. Es complejo aislar el ámbito de cada una de estas sub-disciplinas, que tienen más en común que diferencias. Este es un curso de predicciones analíticas que explora un conjunto de herramientas estadísticas, matemáticas y computacionales para hacer predicciones y clasificaciones confiables, independientemente de entender o no el fenómeno en cuestión. A modo de ejemplo, en este curso, el hecho de ver gente con paraguas será usado para predecir lluvia, aun cuando esté claro que es un disparate fomentar el uso de paraguas para hacer llover. Asimismo, el curso explora las posibilidades de interacción entre el paradigma de aprendizaje automático con los roles estándar de la econometría, en particular en lo que se refiere al análisis causal y estructural.

Perfil: el curso tiene un fuerte carácter técnico y computacional. Al asistente promedio de este curso le apasiona lo analítico, los datos, las fórmulas, los algoritmos, programar, la tecnología y las computadoras, las redes sociales, YouTube y revolver internet.

Requisitos: haber aprobado el curso de econometría de la licenciatura en economía. Alguna experiencia en manejo de datos (Stata, Eviews, R, etc.).

Habilidades computacionales: un objetivo explícito del curso es aprender a manejar con fluidez alguna herramienta computacional sofisticada. El curso de basa en R y Python, ambos lenguajes de programación estadísticos potentes y de amplio uso. No se requiere conocimiento previo, pero si la voluntad de aprender y experimentar más allá de la clase. Se espera que los asistentes sepan (o estén dispuestos a aprender) cómo manejarse con fluidez en las redes sociales (Facebook, Twitter, blogs, etc.) y, en general, afinidad con las computadoras, pues pasaremos una considerable cantidad de tiempo con ellas. Uno de los componentes de participación de este curso se basa en interacción en las redes sociales.

Material: todo el material del curso se encuentra en `http://bigdataudesa.weebly.com`. La página en Facebook del curso está en Big Data UdeSA (grupo cerrado), que servirá para una parte del contenido de "participación" del curso y como herramienta de difusión. Cada alumno debe tener una cuenta en Facebook y unirse al grupo. Se recomienda enfáticamente tener una cuenta en twitter.

Dinámica: este es un curso no estándar, computacionalmente intensivo, que requiere una gran cantidad de trabajo fuera del aula. Los alumnos trabajarán en grupos de tres personas (a determinar el primer día de clase). La aprobación del curso de basa en las siguientes actividades:

1. **Trabajos prácticos (30 % de la nota):** se basan en datos reales, requieren programación y entregar resultados estadísticos, discusión y código de programación en R o Python. Es requisito entregar y aprobar todos los trabajos prácticos.
2. **Participación (10 % de la nota):** se basa en dos actividades:
   - Cada grupo debe realizar una presentación breve (15') a lo largo del curso, a elección de cada grupo. Las opciones disponibles se listan más abajo, en la sección "Papers a presentar". Una vez elegido el paper, informar a la tutora, que les asignará una fecha de presentación. Se espera una presentación profesional, en donde importa tanto el contenido como la estética de la misma. Luego de la presentación, cada grupo debe entregar un powerpoint (o similar) que será subido a la página del curso.
   - Cada dos semanas los grupos deben postear un link relevante (nota, discusión, video, conferencia, base de datos, etc.) y no mencionado en este programa que utilice la metodología vista en la clase teórica, con un breve comentario acerca de su relevancia. No puede haber dos posteos sobre el mismo link.
3. **Elaboración de una propuesta de trabajo (20 % de la nota):** puede ser una aplicación o un trabajo de investigación. En las clases tutoriales se discutirá el formato de esta propuesta.
4. **Examen final (40 % de la nota):** evaluación integral e individual de todo el contenido del curso, incluyendo lecturas y habilidades computacionales. Importante: es condición necesaria aprobar el examen final.

Asistencia y plagio: como es práctica de UdeSA, se requiere asistir como mínimo al 75 % de las clases teóricas y tutoriales, si bien no tomamos asistencia. Velaremos por las cuestiones éticas en lo que se refiere a plagio y honestidad intelectual.

TEMARIO:

1. Introducción: Predecir, explicar. Causalidad y predicción. Data mining, big data, learning, business analytics. Aprendizaje supervisado y no supervisado.
2. Regresión. Modelos lineales, linealizables y no lineales. Vecinos cercanos.
3. Clasificación. Análisis discriminante. Clasificador de Bayes. Regresión logística.
4. Remuestreo. Bootstrap y jacknife. Cross validation. Boostrap en big data. Bags of little boostraps.
5. Regularización y elección de modelos. Lasso y ridge.
6. Estrategias no lineales: saturación, funciones base, splines, regresión local, modelos aditivos.
7. Kernels, densidades y regresion no parametrica. La maldición de la dimensionalidad.
8. Arboles: arboles de regresión y clasificación. Bagging, boosting.
9. Support vector machines. Vector classifiers. Hiperplanos.
10. Reducción de dimensionalidad Componentes principales y factores.
11. Clusters. Metodos jerárquicos y no jerárquicos.

12. Redes neuronales y deep learning.

BIBLIOGRAFÍA:

Libros de consulta

Hastie, T., Tibshirani, R. & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer, New York.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). "An Introduction to Statistical Learning with Applications in R", Springer, New York .

Sosa Escudero, W. (2019). "Big Data: breve manual para conocer la ciencia de datos que ya invadió nuestras vidas", Siglo XXI, Buenos Aires.

Econometría clásica vs. Big Data

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.

Breiman, L. (2003). Statistical modeling: The two cultures. Quality control and applied statistics, 48(1), 81-82.

Calude, C. S., Longo, G. (2016). The Deluge of Spurious Correlations in Big Data. Foundations of Science, 1-18.

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`

Causalidad

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.

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

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.

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

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

Varian, H. R. (2014). Big data: New tricks for econometrics. The Journal of Economic

Perspectives, 28(2), 3-27.

Wager, S., Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests.

Urbanismo

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

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.

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

Precios online vs offline

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

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

Einav, L., Knoepfle, D., Levin, J., Sundaresan, N. (2014). Sales taxes and internet commerce. The American Economic Review, 104(1), 1-26.

Reducción de la dimensionalidad

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

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

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.

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.

New Data: nuevos datos para medir la desigualdad, pobreza y enfermedades

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

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/0ByPgZ6LNcIgGNWO5YVNNMDVqOTA/view

Blumenstock, J., Cadamuro, G., On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.

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.

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.

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.

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.

Random forests

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

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32

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

Limitaciones y desafíos de Big data

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

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

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 automatico: Ideas y desafios 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.

Videos

Susan Athey, Guido Imbens and NBER Organizers. Summer Institute 2015 Methods Lectures, July 18, 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 Chanching 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

PAPERS A PRESENTAR:

Athey, S. et al. (2018). "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data." *AEA Papers and Proceedings*, 108 : 64-67.

Barboza, F., Kimura, H. & Altman, E. (2017). "Machine learning models and bankruptcy prediction." *Expert Systems With Applications*, 83 : 405-417.

Blumenstock, J. E. (2018). "Estimating Economic Characteristics with Phone Data." *AEA Papers and Proceedings*, 108 : 72-76.

Cantú, F. & Saiegh, S. M. (2011). "Fraudulent Democracy? An Analysis of Argentina's Infamous Decade Using Supervised Machine Learning." *Political Analysis*, 19 : 409433.

Edelman, B. & Luca, L. (2014). "Digital Discrimination: The Case of Airbnb.com." Harvard Business School Working Paper, No. 14-054.

Giannone, D., Lenza, M. & Primiceri, G. E. (2018). "Economic Predictions with Big Data: The Illusion of Sparsity." Federal Reserve Bank of New York, Staff Report No. 847.

Glaeser, E., Kim, H. & Luca, M. (2018). "Measuring Gentrification: Using Yelp Data to Quantify Neighborhood Change" *AEA Papers and Proceedings*, 108 : 77-82.

Goolsbee, Austan D., and Peter J. Klenow. (2018). "Internet Rising, Prices Falling: Measuring Inflation in a World of E-Commerce." *AEA Papers and Proceedings*, 108 : 488-92.

Kleinberg, Jon et al. (2018). "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, Oxford University Press, vol. 133(1), pages 237-293.

Kokil Jaidka, Saifuddin Ahmed, Marko Skoric Martin Hilbert (2019). "Predicting elections from social media: a three-country, three-method comparative study", *Asian Journal of Communication*, 29(3) : 252-273.

Koustas, Dmitri K. (2019). "What Do Big Data Tell Us about Why People Take Gig Economy Jobs?" *AEA Papers and Proceedings*, 109 : 367-71.

Magua, W. et al. (2017). "Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers' Critiques."*Journal of Women's Health*, 26(4) : 560-570.