Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

1 Información del equipo pedagógico y horario atención a estudiantes

Profesor: Ignacio Sarmiento-Barbieri (i.sarmiento@uniandes.edu.co)

- Horario Clase: Martes, 6:00 p.m. – 8:50 p.m. Virtual
- Web del Curso: Bloque Neón
- Horario de atención a estudiantes: Via Slack o hacer cita en https://calendly.com/i-sarmiento/horarios-atencion-estudiantes

Profesores Complementarios:

- Lucas Gómez Tobón (l.gomezt@uniandes.edu.co)
  - Horario Clase: Jueves, 6:00 p.m. – 7:20 p.m. Virtual
  - Horario de atención a estudiantes: Via Slack
- Valentina Laverde (v.laverde@uniandes.edu.co)
  - Horario Clase: Viernes, 6:00 p.m. – 7:20 p.m. Virtual
  - Horario de atención a estudiantes: Via Slack

2 Descripción del curso

Este es un curso con un enfoque especial en herramientas relevantes para economistas y ciencias sociales. Está destinado a estudiantes interesados en investigación aplicada y/o análisis de datos grandes y no estructurados. Problemas de predicción e inferencia, con especial énfasis en inferencia causal, atraviesan transversalmente al curso.

Mediante una combinación de contenido asincrónico, clases sincrónicas magistrales y complementarias, talleres grupales, y quices los estudiantes adquirirán las herramientas estadísticas y computacionales necesarias para para responder varias preguntas en economía y en una gran cantidad de subcampos en investigación aplicada. Se hará énfasis especial en el análisis de datos reales, y la aplicación de metodologías específicas; ejemplos incluyen encuestas de hogares, precios de propiedades, datos de internet y redes sociales.

Es pre-requisito haber cursado Microeconomía 3, Econometría 1 y 2, o equivalentes. Se recomienda enfáticamente haber cursado Microeconometría de la MECA o Econometría Avanzada del PEG. Se necesita experiencia básica en manejo de datos y en software `R` o `Python`. El curso se basará principalmente en `R`. Aquellos estudiantes sin experiencia, y con ganas y voluntad de aprender son bienvenidos al curso previa consulta con el docente. ¡Estos programas (y todos) se aprenden utilizándolos!

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

3 Objetivos específicos y competencias

El objetivo de este curso es introducir a los alumnos a un conjunto de herramientas estadísticas, matemáticas, y computacionales para abordar problemas de gran cantidad/tipos/calidad de datos ("large n"), y cantidad de variables ("large p"). Problemas de predicción e inferencia, con especial énfasis en inferencia causal, atravesarán transversalmente al curso. Se buscará también familiarizar a los alumnos con la literatura reciente que utiliza estas herramientas.

Competencias

- Aplicar las técnicas provenientes de la ciencia de datos, la ciencia computacional, y la estadística desde una visión de economistas para resolver problemas puntuales identificados en el contexto.
- Contrastar los distintos algoritmos y su conveniencia para contestar preguntas económicas y sociales con base en criterios relacionados con la naturaleza de problemas económicos y sociales.
- Implementar procesos técnicos para el manejo cuantitativo de datos que surgen de distintas fuentes: páginas web, encuestas, geoespaciales, texto, etc, para resolver problemas económicos y sociales.
- Generar conclusiones y recomendaciones sobre preguntas relevantes a las ciencias sociales por medio del manejo, análisis y síntesis de bases de datos con gran número de observaciones y variables.
- Aplicar el software R y su ecosistema para análisis estadístico, de big data y machine learning.

4 Organización del curso

1. Introducción al aprendizaje estadístico: Predecir, explicar. Causalidad y predicción. Aprendizaje supervisado y no supervisado.
2. Regresión lineal. MCO. Propiedades numéricas. Teorema FWL. Sobreajuste. Métodos de resampleo y validación cruzada. Optimización. Máxima verosimilitud. Modelos lineales, linealizables, y no lineales. Vecinos cercanos. Obtención de datos de la web: scraping y APIs.
3. Selección de modelos y regularización. Lasso y Ridge.
4. Clasificación. Análisis discriminante. Clasificador de Bayes. Regresión logística. Aprendizaje no Balanceado.
5. Árboles de decisión (CARTs). Bosques, Bagging, y Boosting. XGBoost y Super Learners. Aplicaciones en inferencia causal.

6. Datos espaciales. Modelado de dependencia espacial, métodos no paramétricos y econometría espacial.

7. Texto como datos y aprendizaje no supervisado. Clústering, Modelos de categorización de tópicos. Word Embeddings.

8. Introducción a aprendizaje profundo. Redes neuronales.

5 Metodología

La metodología del curso combina contenido asincrónico, clases sincrónicas magistrales y complementarias, talleres grupales, y quices.

Se espera que previo a la sesión sincrónica magistral los estudiantes realicen las actividades asincrónicas, ya que son insumo de los debates y reflexiones que se suscitan durante el encuentro sincrónico.

Se espera que los estudiantes asistan a todas las clases sincrónicas, estudien el material asincrónico y repliquen las aplicaciones presentadas por el profesor, A su vez, los estudiantes realizarán quices semanales individuales y cuatro talleres prácticos grupales para evaluar su aprendizaje.

6 Evaluaciones

Table 1: Puntajes

	Puntaje Individual	Puntaje Total
Quices Semanales	5%	40%
Talleres	15%	60%
Total		100%

- Quices. Los estudiantes tendrán 8 quices individuales semanales en Bloque Neón que evaluarán el aprendizaje individual.

- Talleres. Los estudiantes realizarán trabajos prácticos grupales donde los grupos no podrán superar los 4 miembros. Habrá 4 talleres durante el cursado. Los talleres serán entregados vía Bloque Neón y deberán contar con un repositorio en GitHub. Se espera que todos los miembros hagan contribuciones al repositorio del taller. La calificación del taller se verá reducida si no hay evidencia de contribución de todos los miembros.

Sistema de aproximación de notas definitiva

Las calificaciones definitivas de las materias serán numéricas de uno cinco (1,50) a cinco (5,00), en unidades, décimas y centésimas. La calificación aprobatoria mínima será de tres (3,0). En este

curso se aproximará la nota a la centésima más cercana. Por ejemplo, si el cálculo del cómputo es 3.245, la nota final se aproximará a 3.25; si el resultado del cálculo es 2.994 la nota final será de 2.99

Excusas

Los estudiantes que no presenten las actividades y evaluaciones del curso en la fecha establecida previamente recibirán una calificación de cero (0), a menos que justifiquen su ausencia ante el profesor dentro de un término no superior a ocho (8) días hábiles. Para excusas validas ver Artículo 45 del Reglamento General de Estudiantes de Maestría

7 Asistencias

Se espera que los estudiantes asistan a todas las clases. Si los asistentes faltan a más del 20% de las clases en las que nos reuniremos se penalizará la nota final del curso.

8 Fechas Importantes

- Inicio de clases: 23 de enero.
- Plazo para subir las notas parciales a MiBanner (mínimo el 30%): 24 de febrero.
- Último día de clases: 18 de marzo
- Último día para subir notas finales a MiBanner: 23 de marzo.
- Último día para solicitar retiros: 24 de marzo a las 6:00 p.m.

9 Reclamos y fraude académico

Según los Artículos 62, 63 y 64 del Reglamento General de Estudiantes de Maestría el estudiante tendrá cuatro días hábiles después de la entrega de la evaluación calificada para presentar un reclamo. El profesor magistral responderá al reclamo en los cinco días hábiles siguientes. Si el estudiante considera que la respuesta no concuerda con los criterios de evaluación, podrá solicitar un segundo calificador al Consejo de la Facultad de Economía dentro de los cuatro días hábiles siguientes a la recepción de la decisión del profesor.

Fraude académico

Las conductas que se consideran fraude académico se encuentran en el artículo 4 del Régimen Disciplinario.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

10 Políticas de bienestar

Ajustes razonables

Se entiende por ajustes razonables todas "las modificaciones y adaptaciones necesarias y adecuadas que no impongan una carga desproporcionada o indebida, cuando se requieran en un caso particular, para garantizar a las personas con discapacidad el goce o ejercicio, en igualdad de condiciones con las demás, de todos los derechos humanos y libertades fundamentales" Convención sobre los Derechos de las Personas con Discapacidad, art. 2.

Si requiere ajustes razonables, lo invitamos a buscar asesoría y apoyo en la Coordinación de su programa o en la Decanatura de Estudiantes.

Más información aquí.

Momentos difíciles

Siéntase en libertad de hablar con su profesor si sus circunstancias personales transitorias constituyen un obstáculo para su aprendizaje. En estos casos es responsabilidad del estudiante dar información completa y oportuna al equipo pedagógico para que se evalúe si procede algún ajuste.

Más información aquí.

Cláusula de respeto por la diversidad

Todos debemos respetar los derechos de quienes integran esta comunidad académica. Consideramos inaceptable cualquier situación de acoso, acoso sexual, discriminación, matoneo, o amenaza. Cualquier persona que se sienta víctima de estas conductas puede denunciar su ocurrencia y buscar orientación o apoyo ante alguna de las siguientes instancias: el equipo pedagógico del curso, la Coordinación o la Dirección del programa, la Decanatura de Estudiantes, la Ombudsperson o el Comité MAAD. Si requiere más información sobre el protocolo MAAD establecido para estos casos, puede acudir a Nancy García (n.garcia@uniandes.edu.co) en la Facultad de Economía. Más información sobre el protocolo MAAD: https://agora.uniandes.edu.co/wp-content/uploads/2020/09/ruta-maad.pdf.

11 Referencias Complementarias (sujeta a cambios)

- Albouy, D., Christensen, P., & Sarmiento-Barbieri, I. (2020). Unlocking amenities: Estimating public good complementarity. Journal of Public Economics, 182, 104110.
- Anselin, L. (1982). A note on small sample properties of estimators in a first-order spatial autoregressive model. Environment and Planning A, 14(8), 1023-1030.
- Anselin, Luc, & Anil K Bera. 1998. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics." Statistics Textbooks and Monographs 155. MARCEL DEKKER AG: 237–90.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

- Arbia, G. (2014). A primer for spatial econometrics with applications in R. Palgrave Macmillan.
- Ash, E., Chen, D. L., & Ornaghi, A. (2020). Stereotypes in High-Stakes Decisions: Evidence from US Circuit Courts (No. 1256). University of Warwick, Department of Economics.
- Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola (2020) Dive into Deep Learning. Release 0.15.1.
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685-725.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29-50.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608-650.
- Bivand, R. S., & Pebesma, E. J. (2020). Spatial Data Science (Chapter 8)
- Bivand, R. S., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).
- Breiman, L. (2001). "Random Forests". In: Machine Learning. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. eprint: arXiv:1011.1669v3.
- Carneiro, A., Guimarães, P., & Portugal, P. (2012). Real Wages and the Business Cycle: Accounting for Worker, Firm, and Job Title Heterogeneity. American Economic Journal: Macroeconomics, 4 (2): 133-52.
- Casella, G. (1985). An introduction to empirical Bayes data analysis. The American Statistician, 39(2), 83-87.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.
- Charpentier, Arthur (2018). Classification from scratch, boosting.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
- Chen, T., He, T., & Benesty, M. (2018). XGBoost Documentation.
- Chernozhukov, V., Hansen, C., & Spindler, M (2016). hdm: High-Dimensional Metrics R Journal, 8(2), 185-199.
- Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. The Quarterly Journal of Economics, 133(3), 1163-1228.
- Christensen, P., Sarmiento-Barbieri, I., Timmins C. (2020). Housing Discrimination and the Pollution Exposure Gap in the United States. NBER WP No. 26805
- Clark, M (2018). An Introduction to Text Processing and Analysis with R.Rstudio (2020). Tutorial TensorFlow
- Constantine, P. G., & Gleich, D. F. (2011, June). Tall and skinny QR factorizations in MapReduce architectures. In Proceedings of the second international workshop on MapReduce and its applications (pp. 43-50).
- Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods (Vol. 5). New York: Oxford University Press.
- Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters.
- Efron, B., & Hastie, T. (2016). Computer age statistical inference (Vol. 5). Cambridge University Press.
- Einav, Liran, and Jonathan D. Levin. The data revolution and economic analysis. No. w19035. National Bureau of Economic Research, 2013.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. Econometrica, 78(1), 35-71.
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

- Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public opinion quarterly, 76(3), 491-511.
- Greene, W. H. (2003). Econometric analysis fifth edition. New Yersey: Prentice Hall.
- Gu, J., & Koenker, R. (2017). Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data. Journal of Applied Econometrics, 32(3), 575-599.
- Hayashi, F. (2000). Econometrics. 2000. Princeton University Press. Section, 1, 60-69.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning . New York: springer.
- Kasy M. (2019). Trees, forests, and causal trees. Mimeo.
- Koenker, R. (2013) Economics 508: Lecture 4. Model Selection and Fishing for Significance. Mimeo
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
- Kuhn, M. (2012). The caret package. R Foundation for Statistical Computing, Vienna, Austria.
- Lee, K., & Braithwaite, J. (2020). High-Resolution Poverty Maps in Sub-Saharan Africa. arXiv preprint arXiv:2009.00544.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3):199–231, 2001b.
- Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press. (Chapters 2 & 6)
- Lundberg, I (2017). Causal forests. A tutorial in high dimensional causal inference. Mimeo
- McMillen, D., Sarmiento-Barbieri, I., & Singh, R. (2019). Do more eyes on the street reduce Crime? Evidence from Chicagos safe passage program. Journal of urban economics, 110, 1-25.
- Mullainathan, S. and Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), pp.87-106.
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
- Robinson, D. (2017). Introduction to Empirical Bayes: Examples from Baseball Statistics. 2017.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning for Applied Economics
ECON 4676
Créditos: 4
2023-1

- Rstudio (2020). Tutorial TensorFlow
- Ruud, P. A. (2000). An introduction to classical econometric theory. OUP Catalogue
- Sarmiento-Barbieri, I. (2016). An Introduction to Spatial Econometrics in R.
- Sosa Escudero, W. (2019). Big Data. Siglo Veintiuno Editores
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.
- Tobler, WR. 1979. "Cellular Geography." In Philosophy in Geography, 379–86. Springer.
- Van Loan, C. F., Golub, G. H. (2012). Matrix Computations. United States: Johns Hopkins University Press.
- Varian, Hal R. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28, no. 2 (2014): 3-28.
- Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 114(25), 6521-6526.
- Wasser, L. GIS With R: Projected vs Geographic Coordinate Reference Systems Last Access September 10, 2020
- Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal StatisticalSociety, Series B.67: pp. 301–320