Econ 4676: Big Data and Machine Learning for Applied Economics
Syllabus (Preliminar)

Ignacio Sarmiento-Barbieri

Información del Curso

Clases: Virtuales, Martes y Jueves de 11:00 a 12:30

Sitio web: https://github.com/ECON-4676-UNIANDES

Horario de Oficina: A concretar vía correo electrónico.

email: i.sarmiento@uniandes.edu.co

Descripción General

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. Mediante una combinación de conjuntos de talleres, presentaciones, exámenes, y un trabajo final grupal, los estudiantes adquirirán las herramientas estadísticas y computacionales necesarias para hacer uso de big data y machine learning en investigación empírica.

Prerrequisitos

Microeconomía 3 y Econometría 1. Se recomienda experiencia con programación en `R`, aunque no es requisito. Si bien no es requisito tener experiencia con `R` si es requisito *tener mucha voluntad de aprender y experimentar*. Este programa (y todos) se aprende utilizándolos!

Evaluación

- 10% Participación
- 40% Talleres
- 25% Propuesta de trabajo
- 25% Examen Final

Participación. La participación de los estudiantes es fundamental para sacar el mayor provecho del curso. La virtualidad impone nuevos desafíos y es importante mantenerse conectados para crear las sinergias que surgen de las interacciones humanas. Si bien participación es la actividad con menos peso en la composición final, será el *"tiebreaker"* por el cual decidiré la nota final. Participación no incluye solamente la asistencia a clases, sino también actividades fuera de clase. Una vez registrados en el curso los estudiantes recibiran invitación al canal de Slack, al aula virtual de AWS y a github. La participación será juzgada en función a la participación en las discusiones, en los trabajos grupales, de las interacciones en el canal de Slack, el aprovechamiento de AWS y se espera que estudiantes encuentren al menos un error de tipeo o cualquier otro tipo y los arregrlen a traves de *pull requests* en github

Talleres. Los estudiantes realizarán trabajos prácticos grupales para evaluar su aprendizaje. Los groupos no podrán superar los 4 miembros. Habrán 4 talleres durante el semestre. Se dedicarán al menos 4 clases para la discusión y presentación de los talleres. Los talleres serán submitidos via github y parte de nota de la particpación saldra de la evaluación de la historia del repositorio donde se verá la contribución de cada estudiante.

Propuesta de trabajo. El producto final de este curso es un plan de trabajo con una propuesta de cómo implementar los conceptos y herramientas aprendidas a un problema concreto. La actividad es grupal y puede estar constituida por los mismos miembros del grupo de taller. La actividad estará dividida en 3 entregas. En la primera entrega los grupos se reunirán conmigo y presentaran brevemente (maximo 5 slides) la idea y como planean llevarla a cabo. En una segunda entrega donde se expondrán los datos propuestos. La entrega final será al concluir el curso que consolida toda el trabajo. Se otorgarán bonos a los estudiantes que además de presentar el plan de trabajo o propuesta, entreguen resultados concretos.

Examen final. Este examen pretende evaluar los conceptos y habilidades aprendidas en el curso. Va a ser un examen domiciliario, de tiempo fijo, entre 48-72 horas.

Libros y Recursos (*Preliminar y sujeto a cambios*)

- Farrell, D., Greig, F., and Deadman, E. (2020). Estimating family income from administrative banking data: A machine learning approach. *AEA Papers and Proceedings*, 110:36–41.
- Glaeser, E. L., Kominers, S. D., Luca, M., and Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Temario (*Preliminar y sujeto a cambios*)

1. Introduction to ML: prediction and inference. Supervised and unsupervised learning. MCO revision. Goodness of fit. Introduction to R, Jupyter Lab, Github, and AWS
2. New Economic Observation. Search and computer-mediated behavior. Text Data: news media and social media. Large N Problems: compute and processing. Web scrapers and APIs.
3. Observing from above: Introduction to spatial econometrics. Modeling spatial dependence. Processing big spatial/satellite data, raster data.
4. Intro to non parametric econometrics. Kernels, densities, and non parametric regressions. The curse of dimensionality.
5. Classification: Bayes Risk, Logit Models, ROC analysis.
6. Non lineal methods: Clusters, PCA, K-means, Tress, Boosting andRandom Forests, Support vector machines.

7. Bonus Track: Machine Learning for Causal Inference and Deep Learning.