Machine Learning
Maestria en Economía, UNLP

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

1 Objetivo del curso

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.

2 Contenidos del curso

- Introducción al Aprendizaje de Maquinas: Predecir, explicar. Causalidad y predicción. Aprendizaje supervisado y no supervisado. Trade-off Sesgo-Varianza. Sobreajuste. Métodos de remuestreo y validación cruzada.
- Problemas de Regresión. Modelos lineales, linealizables, y no lineales. Selección de modelos y regularización. Lasso, Ridge y Elastic Net.
- Problemas de Clasificación. Logit y Probit en predicción. Análisis discriminante. Aprendizaje no Balanceado.
- Árboles de decisión (CARTs). Bosques, Bagging, y Boosting. XGBoost, LightGBM, y SuperLearners.
- Aprendizaje profundo y redes neuronales.
- Machine Learning para Inferencia Causal.

3 Requisitos

Los estudiantes deben estar familiarizados con los conceptos de econometría de grado y se recomienda haber cursado econometría avanzada para sacarle más provecho a la clase. Del mismo modo, para la parte computacional es recomendable que los estudiantes se sientan cómodos manipulando datos y con software del estilo de `Python` o `R`. Aquellos estudiantes sin experiencia, con ganas y voluntad de aprender son bienvenidos al curso. ¡Estos programas (y todos) se aprenden utilizándolos!

4 Evaluaciones

La evaluación será a través de 2 (dos) trabajos prácticos grupales donde los grupos no podrán superar los 3 miembros. La participación de los estudiantes es fundamental para sacar el mayor provecho del curso. Por lo tanto, se espera que los estudiantes asistan y participen en todas las clases, lean el material asignado y repliquen las aplicaciones presentadas por el profesor.

5 Referencias

- 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.
- Breiman, L. (2001). "Random Forests". In: Machine Learning. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004 eprint: arXiv:1011.1669v3.
- 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.
- 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.
- 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.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Kasy M. (2019). Trees, forests, and causal trees. 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.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3):199–231, 2001b.
- Lundberg, I (2017). Causal forests. A tutorial in high dimensional causal inference. Mimeo
- 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.
- 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.
- Varian, Hal R. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28, no. 2 (2014): 3-28.
- Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal StatisticalSociety, Series B.67: pp. 301–320