Universidad de los Andes
Facultad de Economía

BIG DATA
2021-Verano

PROGRAMA (JUNIO 20 DE 2021)

Profesor: Alvaro Riascos
Contacto: e-mail: ariascos@uniandes.edu.co

1. Objetivos de la materia

Este curso introduce a los estudiantes a algunas de las aplicaciones más importantes de la minería de datos y el aprendizaje de máquinas (Big Data) a problemas empresariales y económicos. El énfasis es aprender las herramientas más comunes que se usan en la práctica y mostrar una gran variedad de ejemplos de cómo estas enriquecen el análisis de problemas específicos de las ciencias sociales.

Para esto los estudiantes serán introducidos muy rápidamente a los elementos más básicos del aprendizaje estadístico (i.e. riesgo, el problema de aprendizaje de máquinas, error de estimación y aproximación), las principales técnicas algorítmicas: el modelo de regresión lineal, vecinos más cercanos, el modelo logístico, árboles y bosques aleatorio y las principales técnicas para el análisis de textos, los elementos básicos de los sistemas de información geográfica, redes y lo que se conoce como análisis no supervisado. Con este conjunto de herramientas haremos muchas aplicaciones a problemas sociales en salud, desarrollo, mercado laboral, crimen, redes sociales, etc.

Además de aprender los fundamentos teóricos, principales técnicas y aplicaciones, los estudiantes aprenderán herramientas básicas de programación en R, lo que les permitirá una vez finalizado el curso, hacer un trabajo final en el que se plantee un problema que utilice diversas fuentes de datos, aplicar las técnicas aprendidas en clase y ofrecer una solución novedosa. Este trabajo final podrá estar basado en problemas específicos del lugar de trabajo de los estudiantes, tesis o intereses particulares.

2. Contenido

Sesión	Tema	Referencias
1		
	Teoría: Fundamentación estadística del aprendizaje de máquinas y la minería de datos. **Aplicación:** Breve introducción al Aprendizaje de máquinas en economía: predicción vs. Causalidad (AR).	**Teoría:** [LS] [JWHT]: Capítulo 1,2., [HTF]: Capítulo 1,2. **Aplicaciones:** Athey, Imbens [2019]. Kleinberg, Ludwig, Mullainathan, Obermeyer [2015].
2 - 4	**Teoría:** Métodos Lineales de clasificación (regresión lineal, modelo logístico). Métodos no lineales: árboles, bosques aleatorios, kernels. Selección y validación de modelos: Regularización, validación cruzada, (sub) bagging, boosting, curva ROC. **Aplicación:** Predicción de hospitalizaciones innecesarias y riesgo moral (AR)	**Teoría:** [JWHT]: Capitulos 3,4,5,6 **Aplicaciones:** Riascos, A., N. Serna (2018).
5 - 6	**Teoría:** Análisis no supervisado: **Aplicación:** Estimación de densidad de Kernels y representación espacio temporal del crimen en Bogota (Isabella Rodas)	**Teoría:** [JWHT]: Capitulos 10 **Aplicaciones:** https://centroanalitic app.org/proyectos/cr imen/
7	**Teoría:** Análisis de texto **Aplicación:** Ofertas laborales y sesgos implicitos (Jose Sebastian Ñungo)	**Teoría:** Presentaciones **Aplicaciones:** https://centroanalitic app.org/proyectos/4 0mil-primeros-empleos/
8	**Teoría**: Introducción a los sistemas de información geográfica **Aplicación:** Reconstrucción de red de vías terciarias de Colombia (Camilo Erasso)	
9 - 10	**Teoría:** Teoría de grafos **Aplicación:** Por definir (AR)	**Teoría: [J]:** Capítulos 1, 2 **Aplicaciones:** Por definir

3. Metodología

Este curso es muy práctico y requiere de la participación intensa de los estudiantes para su desarrollo. Las clases magistrales tendrán una duración de tres horas complementadas con una hora y media adicional por semana de programación en R. No es necesario saber de programación en estos lenguajes aunque es muy deseable haber estado expuesto a un nivel introductorio cualquier software estadístico (Stata, SPSS, EViews,etc) o lenguaje de programación (Python, Matlab, Mathematica, etc).

Los estudiantes tendrán que formar grupos (de mínimo dos personas a máximo cuatro personas) para hacer las siguientes entregas:

- Taller 1 (30% de la nota). **Entrega Lunes 5 de julio.**
- Prepropuesta de proyecto final (10% de la nota): **Entrega Lunes 12 de junio.**
- Taller 2 (30% de la nota). **Entrega Lunes 19 de julio.**
- Proyecto final (30% de la nota). **Entrega Miércoles 28 de Julio.**

4. Sistema utilizado para aproximar la nota definitiva

El sistema de notas definitivas es el siguiente: las notas totales con decimales en 0 o en .5 no se modificarán. Las notas totales con decimales entre .25 a .49 y entre .75 a .99, se aproximarán a la nota definitiva siguiente. Las notas con decimales entre .01 a .24 y entre .51 a .74, se aproximarán a la nota definitiva anterior.

5. Protocolo MAAD: Conductas de maltrato, acoso, amenaza, discriminación, etc.

El miembro de la comunidad que sea sujeto, presencie o tenga conocimiento de una conducta de maltrato, acoso, amenaza, discriminación, violencia sexual o de género (MAAD) deberá poner el caso en conocimiento de la Universidad. Ello, con el propósito de que se puedan tomar acciones institucionales para darle manejo al caso, a la luz de lo previsto en el protocolo, velando por el bienestar de las personas afectadas.

Para poner en conocimiento el caso y recibir apoyo, usted puede escribir a la ínea MAAD: lineamaad@uniandes.edu.co

6. Bibliografía

Las referencias principales del curso son:

- [LS]: Luxburg, U., B. Scholkopf. 2008. Statistical Learning Theory: Models, Concepts and Results.
  http://arxiv.org/abs/0810.4752

- [JWHT]: Introduction to Statistical Learning with Applications in R.
  http://www-bcf.usc.edu/~gareth/ISL/

- [HTF]: Hastie, T., Tibshirani, R. y J. Hastie. 2009. The Elements of Statistical Learning: Data Minning, Inference and Prediction. Segunda Edición. Springer.

http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

- [J] Jackson, M. 2018. Social and Economic Networks.

- Athey,S.,and G. Imbens. 2019. Machine Learning Methods Economists Should Know About.

- Dulce, M., Ramirez, S. y A. Riascos (2018). Efficient allocation of law enforcement resources using predictive police patrolling.
https://arxiv.org/pdf/1811.12880.pdf

- Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction Policy Problems

- Alvaro J. Riascos (University of los Andes and Quantil, Bogotá, Colombia) and Natalia Serna (University of Wisconsin-Madison, Madison, USA) International Journal of Knowledge Discovery in Bioinformatics (IJKDB) 8(2). A Machine Learning Based Program to Prevent Hospitalizations and Reduce Costs in the Colombian Statutory Health Care System.