Big Data: Machine Learning
Facultad de Economía
Universidad del Rosario

Primer Semestre de 2017

Profesor: Jorge Gallego (jorge.gallego@urosario.edu.co)
Profesor Asistente: Juan D. Martínez (jumartinez@dnp.gov.co)

1 Introducción

Carros sin conductor, recomendaciones de Netflix, correos spam en Gmail, diagnóstico de cáncer en hospitales, clientes que no pagan su tarjeta de crédito, sospechosos enviados a la cárcel. Todos estos ejemplos tienen algo en común: las herramientas del aprendizaje automatizado permiten hacer predicciones precisas que antes no eran posibles. Con la proliferación de nuevas tecnologías y la consolidación de nuevas fuentes de información, como las redes sociales, las ciencias sociales logran acceder a nuevos y grandes volúmenes de datos. Algunos llaman a este fenómeno Big Data. Tan importante como la información es el hecho de que para analizarla se han desarrollado nuevas técnicas y herramientas matemáticas, estadísticas y computacionales. El objetivo de este curso es introducir gentilmente al estudiante a algunas de estas técnicas. El curso girará en torno al método de *machine learning*, o aprendizaje automatizado, que ha sido aplicado a campos tan variados como el marketing, las finanzas, el comercio, la medicina, la justicia y por supuesto, las políticas públicas.

2 Metodología

El curso será teórico y práctico, con presentaciones magistrales en las que el profesor expondrá los principales conceptos estadísticos y matemáticos subyacentes a estas

técnicas, seguidas de talleres prácticos en los que se enseñará a los estudiantes a resolver problemas concretos. Para tal propósito, en el curso usaremos el paquete estadístico R e introduciremos los conocimientos básicos necesarios para aplicar las técnicas que estudiaremos en el curso.

3 Contenido

1. **Introducción**

   - ¿Qué es Big Data?
   - ¿Qué es Machine Learning?
   - Causalidad vs. Predicción
   - Usos y abusos de Machine Learning
   - Consideraciones éticas

2. **Fundamentos básicos de R**

   (a) Vectores y factores

   (b) Listas, bases de datos y matrices

   (c) Funciones

   (d) Gráficas

   (e) Iteraciones

   (f) Programación básica

   (g) Aplicación: *web scraping* usando R

3. **Clasificación usando vecinos más cercanos**

   (a) Algoritmo k-NN

   (b) Aplicaciones

4. **Aprendizaje probabilístico: *Naive Bayes***

   (a) Métodos bayesianos

   (b) Algoritmo *Naive Bayes*

(c) Aplicaciones

(d) Aplicación: Análisis automatizado de texto

5. **Árboles de decisión**

   (a) Algoritmo de árboles de decisión C5.0

   (b) Reglas de clasificación

   (c) *Random Forests*

   (d) Aplicaciones

6. **Métodos de predicción: Regresión**

   (a) Regresión simple, múltiple y correlaciones

   (b) Regresión Logit

   (c) Árboles de regresión

   (d) Aplicaciones

7. **Redes neuronales y *Support Vector Machines***

   (a) Redes neuronales

   (b) *Support Vector Machines*

   (c) Aplicaciones

   (d) Aplicación: análisis de sentimientos

8. **Patrones: Análisis de canastas de mercado**

   (a) Reglas de asociación

   (b) Aplicaciones

9. ***Clustering***

   (a) Agrupación con *k-means*

   (b) Aplicaciones

10. ***Deep Learning***

(a) Introducción al *Deep Learning*

(b) Aplicaciones

11. **Desempeño de modelos**

    (a) Evaluación del desempeño

    (b) Mejoramiento del desempeño

    (c) *Feature extraction, feature selection, feature engineering*

    (d) Ensamblaje de modelos, *overfitting*

4 Método de Evaluación

La nota final se basará en un examen parcial, un examen final, talleres prácticos y un trabajo final. La nota final se promediará de la siguiente manera:

- Examen parcial 20%
- Examen final 20%
- Talleres 30%
- Trabajo 30%

5 Políticas del curso

1. Se espera que los estudiantes realicen las lecturas asignadas antes de clase.
2. La asistencia a clase es obligatoria.
3. Se espera que haya una participación activa durante las clases.
4. Aunque no es obligatorio, se recomienda a los estudiantes traer sus computadores portátiles para el curso.

6 Bibliografía

Por su simplicidad, claridad y aplicaciones en R, el libro guía del curso es Lantz (2015). Sin embargo, el nivel de profundización (especialmente teórica) de este libro no es el más

alto. Para explicaciones más avanzadas y profundas, remitirse a Hastie et al. (2009) o James et al. (2013). Una buena guía para comenzar a programar en R se encuentra en Matloff (2011).

- Conway, D. y J. Myles (2012). *Machine Learning for Hackers*. O'Reilly Media

- Hastie, T., R. Tibshirani y J. Friedman (2009). *The Elements of Statistical Learning*. Springer

- James, G., D. Witten, T. Hastie y R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer

- Lantz, B. (2015). *Machine Learning with R*. Packt Publishing.

- Matloff, N. (2011). *The Art of R Programming*. No Starch Press.

- Siegel, E. (2013). *Predictive Analytics*. John Willey & Sons.

- Zumel, N. y J. Mount (2014). *Practical Data Science with R*. Manning Publications Co.