Big Data for Public and Private Sectors
Parte 2
Facultad de Economía
Universidad del Rosario

Primer Semestre de 2017

Profesor: Jorge Gallego (jorge.gallego@urosario.edu.co)

1 Introducción

Las técnicas de aprendizaje automatizado, o *machine learning*, cada vez se consolidan más como un método poderoso para analizar los datos derivados de la denominada revolución del *big data*. Las aplicaciones trascienden el mundo de la electrónica y día a día ganan espacio en los sectores público y privado. Estudios de riesgo crediticio, detección de células cancerosas, predicción del crimen en las ciudades, decisiones de justicia criminal, pronóstico de deserción estudiantil, entre otras, son algunas de las aplicaciones directas de los métodos que estudiaremos en este curso. En la segunda parte del curso de *big data* para los sectores público y privado, estudiaremos las principales técnicas de aprendizaje supervisado, aplicando los métodos estudiados a diversos casos de ambos sectores.

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. Todos los materiales del curso se encuentran en el repositorio de GitHub https://github.com/jagallegod/Big-Data-4-Public-and-Private-Sectors

3 Contenido

1. **Introducción**
    - Causalidad vs. Predicción
    - ¿Qué es Big Data?
    - ¿Qué es Machine Learning?
    - Algoritmos de Machine Learning
    - Algunos ejemplos
2. **Clasificación usando vecinos más cercanos**
    (a) Algoritmo k-NN
    (b) Aplicaciones
3. **Aprendizaje probabilístico: *Naive Bayes***
    (a) Métodos bayesianos
    (b) Algoritmo *Naive Bayes*
    (c) Aplicaciones
4. ***Support Vector Machines***
    (a) *Support Vector Machines*
    (b) Aplicaciones
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. **Regresiones RIDGE y LASSO**

   (a) Regresión RIDGE

   (b) Regresión LASSO

   (c) Aplicaciones

8. **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 de esta sección del curso resulta del promedio simple de un examen final y una propuesta de trabajo.

- Examen final 50%
- Propuesta de trabajo 50%

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). 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.