SEMINARIO MINERIA DE DATOS Y SUS APLICACIONES

ECON 4649

ariascos@uniandes.edu.co

2013-10

Profesor: Alvaro J. Riascos Villegas
Contacto: e-mail: ariascos@uniandes.edu.co, Skype: alvaro.riascos
Oficina: Bloque W, Oficina 918.
Horario de clase: Martes 2PM-3:50PM, W-201.
Horario de atención a estudiantes: 11:30AM – 12:30PM, con cita previa o por skype.
Página en Internet del curso: se anunciará el primer día de clase.

1. Objetivos de la materia

Este seminario introduce los estudiantes en los pilares teóricos fundamentales de la teoría de aprendizaje estadístico como marco teórico de la minería de datos (el problema de aprendizaje o *machine learning*, el compromiso entre sesgo y varianza, aproximación y error, riesgo, consistencia, regularización, complejidad, etc.) Paralelamente se van aprender las principales técnicas de minería de datos (método de vecindades, redes neuronales, redes bayesianas, árboles, *boosting*, *cross validation*, maquinas de vectores de soporte, *clustering*, etc.) a través de ejemplos y problemas que los estudiantes deberán implementar y resolver haciendo uso de un computador y, en lo posible, resolver problemas que sean de su interés (datos de redes sociales, reconocimiento de caracteres, extracción de señales, etc.).

Este curso es muy práctico (75% del curso) y requiere de la participación constante de los estudiantes, haciendo presentaciones de diferentes técnicas, mostrando avances en el proyecto final, etc. Para facilitar la aplicabilidad de las técnicas aprendidas habrá una introducción corta a R (http://www.r-project.org/) y parte de las técnicas se aprenderán de forma paralela con su implementación en este lenguaje de programación y sus respectivos paquetes.

2. Contenido

1. Ene. 23 **Minería de Datos y Aprendizaje estadístico**
   [LS]
   [HTF] - Capítulo 2.

2. Ene. 30 **Introducción a R y Rattle**
   Presenta: Elioth Sanabria y Miguel Bernal
   - Familiarización con competencia en Kaggel (Elioth)
   - Introducción a R (Elioth)
   - Introducción a Rattle (Miguel)

3. Feb. 5 **Aprendizaje Estadístico: Modelos, Conceptos y Resultados**
   - [LS]
   - [HTF] - Capítulo 2.

4. Feb. 12 **Redes Neuronales y Splines en R con Aplicaciones al Sector Eléctrico**
   - Presenta: Luis Felipe Parra

5. Feb. 19 **Análisis de Conglomerados**
   [W] – Capitulo 9.
   [HTF]
   Presentación Preliminar Proyecto: Grupo 1

6. Feb. 26 **Reglas de Asociación**
   [W] – Capitulo 10.
   [HTF]
   Presentación Preliminar Proyecto: Grupo 2

7. Mar. 5 **Árboles de Decisión**
   [W] – Capítulo 11.
   [HTF]
   Presentación Preliminar Proyecto: Grupo 3

8. Mar. 12 ***Random Forests***
   [W] – Capítulo 12
   [HTF]
   Presentación Preliminar Proyecto: Grupo 4

9. Mar. 19 ***Boosting***
   [W] – Capítulo 13
   [HTF]
   Presentación Preliminar Proyecto: Grupo 5

10. Mar. 26 **SEMANA SANTA**

11. Abr. 2 **Support Vector Machines**
    [W] – Capítulo 13
    [HTF]
    Presentación Preliminar Proyecto: Grupo 6

12. Abr. 9 **MCMC, algoritmos Metropolis-Hastings y Simulated Annealing**
    Presenta: Elioth Sanabria

13. Abr. 16 **Presentación Proyecto Final Grupo 1 y 2**

14. Abril 23 **Trabajo Individual**

15. Abr. 30 **Presentación Proyecto Final Grupo 3 y 4**

16. May. 7 **Presentación Proyecto Final Grupo 5 y 6**

3. Metodología

Este curso es muy práctico y requiere de la participación intensa de los estudiantes para su desarrollo. Los estudiantes tendrán que formar grupos (2 o máximo 3 personas) para hacer tres presentaciones: (1) Presentar una técnica estándar en minería de datos. (2) Hacer una presentación corta de la idea básica, motivación, datos, metodología inicial, etc. de su proyecto final y (3). Hacer una presentación detallada del proyecto final, resultados, etc. Adicionalmente, habrá algunas presentaciones teóricas del profesor y algunas presentaciones de invitados sobre técnicas importantes en minería de datos y sus aplicaciones a la industria.

4. Sistema de evaluación

La evaluación del curso consiste de cuatro notas. Las tres primeras evalúan las tres presentaciones más el informe final sobre el proyecto de investigación que cada estudiante escogió. La cuarta evalúa el desempeño en la competencia de Kagel (http://www.kaggle.com/): MD 2013.

Estos se avaluarán así:

1. Presentación de la técnica de minería de datos que escogió el grupo. (25%)
2. Presentación: Avance proyecto individual. Definición del problema, exploración o investigación a realizar. (15%)
3. Presentación e informe. Máximo seis páginas incluyendo tablas, gráficos, bibliografía, etc. (35%)
4. Resultado concurso Kaggel. (25%).

5. 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.

6. Bibliografía

Referencias obligatorias

[LS]: Luxburg, U., B. Scholkopf. 2008. Statistical Learning Theory: Models, Concepts and Results.

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

[W]: Williams, G. Data Minning with Rattle and R. Springer.

Fecha de entrega del 30% de las notas: Marzo 22 de 2013
Fecha límite para retiros: Abril 05 de 2013