

Profesor: Alvaro J. Riascos Villegas

Contacto: e-mail: ariascos@uniandes.edu.co, Skype: alvaro.riascos

Oficina: Bloque W, Oficina 918.

Horario de clase: Martes 2:00 PM-3:20 PM, W-205.Salón: W-506

Horario de atención a estudiantes: 11:30AM – 12:30PM, con cita previa o por skype.

Página en Internet del curso: www.alvaroriascos.com

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*, máquinas de vectores de soporte, *clustering*, reglas de asociación, etc.) a través de ejemplos y problemas que los estudiantes deberán implementar y resolver haciendo uso de un computador y, en lo posible, resolviendo problemas que sean de su interés (análisis de datos de redes sociales, reconocimiento de caracteres, extracción de señales, pronósticos, clasificación de clientes, alertas de fraude, anomalías, 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. 21 **Minería de Datos y Aprendizaje estadístico**
[LS]
[HTF] - Capítulo 2.
2. Ene. 28 **Aprendizaje Estadístico I: Modelos, Conceptos y Resultados**
[LS]
[HTF] - Capítulo 2 y 7.
3. Feb. 4 **Aprendizaje Estadístico II: Modelos, Conceptos y Resultados**
[LS]
[HTF] - Capítulo 2 y 7.
4. Feb. 11 **Introducción a R**
Presenta: Miguel Bernal

5. Feb. 18 **Métodos Lineales de Regresión**
[HTF] – Capítulo 3
Presentación Preliminar Proyecto: Grupo 1
6. Feb. 25 **Métodos lineales para Clasificación**
[HTF] – Capítulo 4
Presentación Preliminar Proyecto: Grupo 2
7. Mar. 4 **Evaluación y Selección de Modelos**
[HTF] - - Capítulo 7
Presentación Preliminar Proyecto: Grupo 3
8. Mar. 11 **Inferencia y Promedios**
[HTF] – Capítulo 8
Presentación Preliminar Proyecto: Grupo 4
9. Mar. 18 **Boosting**
[HTF] – Capítulo 10
Presentación Preliminar Proyecto: Grupo 5
10. Mar. 25 **Support Vector Machines**
[HTF] – Capítulo 13
Presentación Preliminar Proyecto: Grupo 6
11. Abr. 1 **Análisis No Supervisado**
[HTF] – Capítulo 14
Presentación Preliminar Proyecto: Grupo 7
12. Abr. 8 **Random Forests**
[HTF] – Capítulo 15
Presentación Preliminar Proyecto: Grupo 8
13. Abr. 15 **SEMANA SANTA**
14. Abr. 22 **Presentación Proyecto Final Grupo 1 y 2**
15. Abr. 29 **Presentación Proyecto Final Grupo 3, 4, 5**
16. May. 6 **Presentación Proyecto Final Grupo 6, 7, 8**

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 2014.

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 Kagel. (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 Mining, Inference and Prediction. Segunda Edición. Springer

Fecha de entrega del 30% de las notas: Mar. 21 de 2014

Último día para solicitar retiros (no genera devolución): Mar. 28 de 2014

Último día para subir notas finales en banner: Jun. 03 de 2014