

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-205.

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

Página en Internet del curso: <http://www.webpondo.org/ariascos/teaching/mineriadatos.htm>

1. Objetivos de la materia

Este seminario introduce los estudiantes en los pilares teóricos fundamentales de la moderna 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.) Posteriormente, y en su mayor parte (tres cuartas partes del seminario), 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, problemas que sean de su interés (datos de redes sociales, reconocimiento de caracteres, extracción de señales, etc.).

2. Contenido

1. Ene. 24 **Aprendizaje estadístico: Introducción**
[LS]
[HTF] - Capítulo 2.
2. Ene. 31 **Introducción a R**
Presenta: Ivan Cadena
 - Familiarización con competencia en Kaggel
 - Lectura capítulos [HTF]
 - Proyecto individual (capítulos [HTF], libros o artículos de minería de datos)
3. Feb. 7 **Algoritmo de vecindad más cercana**
Presenta: Ivan Cadena
 - Lectura capítulos [HTF]
 - Proyecto individual (capítulos [HTF], libros o artículos de minería de datos)
4. Feb. 14 **Aprendizaje supervisado**
[LS]
[HTF] – Capítulo 2.

- | | |
|-------------|--|
| 5. Feb. 21 | Sesión computacional: Complejidad, sesgo y varianza
[LS]
[HTF] – Capítulo 2. |
| 6. Feb. 28 | Métodos de Regresión Lineal
[LS]
[HTF] – Capítulo 3. |
| 7. Mar. 6 | Sesión computacional: Selección de variables y determinantes cáncer |
| 8. Mar. 13 | Métodos lineales de clasificación
[LS]
[HTF] – Capítulo 4. |
| 9. Mar. 27 | Sesión computacional: Modelos de scoring para otrogamiento de crédito |
| 10. Abr. 3 | SEMANA SANTA |
| 11. Abr. 10 | Evaluación y selección de modelos
[LS]
[HTF] – Capítulo 7. |
| 12. Abr. 17 | Sesión computacional: <i>Cross validation, bootstrap, bagging.</i>
[LS]
[HTF] – Capítulo 7. |
| 13. Abr. 24 | Presentaciones individuales I |
| 14. May. 1 | Presentaciones individuales II |
| 15. May. 8 | Presentaciones individuales III |

3. Metodología

El curso se desarrollará en una sesión magistral por semana intercalando una semana de teoría con una semana de sesión computacional donde el profesor enseñará a aplicar algunos de los métodos estudiados en las clases teóricas.

4. Sistema de evaluación

La evaluación de cuatro partes. Las tres primeras partes están asociadas a un proyecto de investigación o exploración de datos que elegirá cada estudiante (o pareja de estudiantes) con la ayuda del profesor. La cuarta parte consiste de un concurso de clasificación administrado por Kaggle. Estos se evaluarán así:

1. Avance proyecto individual. Definición del problema, exploración o investigación a realizar. Máximo una página. (20%).
2. Presentación relacionado con el proyecto de investigación. (25%).
3. Presentación informe. Máximo cinco páginas incluyendo tablas, gráficos, bibliografía, etc. (25%)
4. Resultado concurso Kaggle. (30%).

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