

Profesores:

Adolfo J. Quiroz (aj.quiroz1079@uniandes.edu.co)

Oficina: H-004

Horario de atención a estudiantes:

Alvaro J. Riascos Villegas (ariascos@uniandes.edu.co)

Skype: alvaro.riascos

Oficina:

Horario de atención a estudiantes:, con cita previa o por skype.

Página en Internet del curso: www.alvaroriascos.com

1. Motivación

El inmenso potencial de aplicación del área de Reconocimiento de Patrones en medicina, en verificación de firmas, huellas dactilares, en identificación de señales eléctricas, etc., por un lado, y los grandes avances realizados recientemente tanto en la práctica como en la teoría probabilística de los procedimientos del área, tales como máquinas de soporte vectorial, árboles de clasificación, clasificadores de vecinos más cercanos, clasificadores basados en estimación de densidades por núcleos, además del desarrollo reciente de la Teoría de Aprendizaje Estadístico y el impacto que esta teoría ha tenido sobre la práctica de Reconocimiento de Patrones, motivan la inclusión de esta asignatura entre las electivas del área de Estadística del Pregrado y la Maestría en Matemáticas de Uniandes. Una posible ventaja de ofrecer esta asignatura es que pudiera atraer a estudiantes de posgrados en Ingeniería y Computación, promoviendo así la interacción enriquecedora entre estudiantes de distintos programas.

Este curso 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, minería de texto, 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 y requiere de la participación activa de los estudiantes, haciendo presentaciones de diferentes técnicas, talleres, haciendo un proyecto final y participando de un concurso en Kaggle

(<http://www.kaggle.com/>). 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. Objetivos y competencias

- ✓ Que el estudiante pueda enunciar el planteamiento probabilístico del problema de Reconocimiento de Patrones y su aplicabilidad a problemas de interés práctico.
- ✓ Que el estudiante entienda como implementar y como evaluar la precisión del Clasificador de Bayes cuando puede estimarse paramétricamente la verosimilitud de los parámetros de interés.
- ✓ Que el estudiante maneje el Análisis de Componentes Principales y el Análisis Discriminante como herramientas estadísticas auxiliares en el contexto de Reconocimiento de Patrones.
- ✓ Que el estudiante pueda implementar y conozca las ventajas y limitaciones teórico-prácticas de los métodos de discriminación lineal de Fisher, y otros discriminadores basados en planos separadores, como por ejemplo el procedimiento de Ho-Kashyap.
- ✓ Que el estudiante conozca las ventajas teóricas del procedimiento de k-vecinos- más-cercanos en el contexto de clasificación y conozca los criterios prácticos para su implementación.
- ✓ Que el estudiante sea capaz de seleccionar un clasificador siguiendo los criterios de Minimización de Riesgo Empírico y Validación Cruzada.
- ✓ Que el estudiante conozca la noción de Minimización de Riesgo Estructural, pueda estimar el riesgo estructural en ejemplos específicos y pueda utilizar este criterio en la selección de la familia de clasificadores apropiada a un problema. (Se incluirá este tópico si el tiempo lo permite).
- ✓ Que el estudiante pueda explicar las ventajas del método de Máquinas de Soporte Vectorial en el problema de clasificación y pueda explicar sus bases teóricas. Que el estudiante pueda implementar algoritmos de Máquinas de Soporte Vectorial en ejemplos concretos.
- ✓ Que el estudiante pueda utilizar las herramientas de análisis de conglomerados (clustering) como apoyo en la resolución de problemas de clasificación.
- ✓ Que el estudiante pueda aplicar todos estos conocimientos y habilidades técnicas para atacar un buen problema practico en el que la minería de datos pueda hacer una diferencia importante en comparación a otras técnicas.
- ✓ Que el estudiante aplique los criterios desarrollados en la solución de problemas tipo proporcionados por el profesor de la materia.

3. Contenido

Semana	Tema	Expositor	Referencias
1	Introducción a la Minería de Datos Aprendizaje Estadístico: Modelos, Conceptos y Resultados	Alvaro Riascos	[HTF]: Capítulo 1,2
2	Aprendizaje Estadístico: Modelos, Conceptos y Resultados	Alvaro Riascos	[HTF]: Capítulo 1,2

3	Métodos lineales de regresión clasificación Aplicaciones		Alvaro Riascos	[HTF]: Capítulo 3,4
4	Métodos no lineales y regularización Aplicaciones		Alvaro Riascos	[HTF]: Capítulo 5,6
Entrega Taller I				
5	<i>Arboles, Boosting de Árboles y Random Forests</i>		Alvaro Riascos	[HTF]: Capítulos 9, 10 y 15
6	Selección de Modelos e Inferencia Minería de Texto Entrega Taller II		Alvaro Riascos	[HTF]: Capítulos 7 y 8
7	Minería de Texto Última Semana 30% Notas		Alvaro Riascos	[HTF]: Capítulos 9, 10 y 15
8	Análisis no supervisado		Alvaro Riascos	[HTF]: Capítulos 14
9 SEMANA DE TRABAJO INDIVIDUAL				
10	Primera evaluación concurso Kaggle Estimación de densidades por núcleos. Propiedades asintóticas. Selección del tamaño de la ventana. Aplicación a Clasificación		Adolfo Quiroz	Notas de clase. DHS Secciones 4.1 a 4.3.3, 4.3.6
11	Clasificación por vecinos más ceranos. Propiedades. Selección del número de vecinos. Cros-validación.		Adolfo Quiroz	DHS: Cap. 4. Secciones 4.4 y 4.5. HTF: Cap. 6. DGL: Cap. 5 Sección 5.5.
12	Leyes Fuertes Uniformes de Grandes Números. Dimensión VC. Convergencia del Clasificador de		Adolfo Quiroz	DGL: Capítulos 12 y 13.
Mínimo Error Empírico. Examen Parcial 1				
13	Minimización de Riesgo Estructural. Comparación con otros criterios de comparación de modelos. Información de Akaike		Adolfo Quiroz	DGL: Cap. 18. HTF: Cap. 7
14	Propiedades teóricas de las Máquinas de Soporte Vectorial.		Adolfo Quiroz	HTF: Cap. 12 CST: Cap. 4

15	Máquinas de Soporte Vectorial. Implementación.	Adolfo Quiroz	CST: Caps. 5 y 6. HTF: Cap. 14
16	Otros aspectos de Máquinas de Soporte Vectorial. Examen Parcial 2. Segunda evaluación concurso Kaggle	Adolfo Quiroz	CST: Caps. 6 y 7

4. Metodología

El curso consiste de 4 horas semanales de clases en aula, donde el profesor expone el contenido de la materia, incentivando la participación de los alumnos a través de preguntas. Este curso es muy práctico y requiere de la participación intensa de los estudiantes para su desarrollo.

5. Formato de evaluación

Primera parte del curso (60% de la nota)

- Taller 1 (15% de la nota)
- Taller 2 (15% de la nota)
- Primera evaluación concurso Kaggle (15% de la nota): <http://www.kaggle.com/>
- Segunda evaluación concurso Kaggle (15% de la nota): Una semana antes del último día para entregar notas.

Segunda parte del curso (40% de la nota)

La evaluación consta de dos exámenes parciales, que corresponden a un 20% de la evaluación c/u.

6. Sistema utilizado para aproximar la nota definitiva

Las notas finales sobre 5 se aproximarán a la décima más cercana. Por ejemplo 3.77 se aproximará a 3.8, mientras que 3.72 a 3.7. La nota definitiva mínima aprobatoria es 3.0.

7. Bibliografía

Las referencias principales del curso son:

[LS]: Luxburg, U., B. Scholkopf. 2008. Statistical Learning Theory: Models, Concepts and Results.
<http://arxiv.org/abs/0810.4752>

[JWHT]: Introduction to Statistical Learning with Applications in R.

¹ El mejor grupo en el concurso tendrá una nota de 5, el segundo mejor de 4.5, el tercer 4.0, el cuarto 3.5 y el quinto 3.0.

<http://www-bcf.usc.edu/~gareth/ISL/>

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

http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

[DHT] R. O. Duda, P. E. Hart, D. G. Stork (2001) Pattern Classification. John Wiley. New York.

[DGL] L. Devroye, L. Györfi, G. Lugosi. (1991). A Probabilistic Theory of Pattern Recognition.

[CST] N. Cristianini, J. Shawe-Taylor (2004). An Introduction to Support Vector Machines and other kernel based learning methods. Cambridge University Press.

[TK] S. Theodoridis, K. Koutroubas (2008). Pattern Recognition. 4th edition. Elsevier

Las referencias de las aplicaciones se irán subiendo a la página web del curso a lo largo del semestre.