

Métodos de Aprendizaje Estadístico y Machine Learning

Profesor: Carlos Felipe Valencia Arboleda

cf.valencia@uniandes.edu.co

Of. ML 718

Instructor: Manuel Bolívar

ma.bolivar643@uniandes.edu.co

Of. ML 334

Clase: Martes y Jueves de 3:30pm a 4:50pm. Salón: TBA

A. DESCRIPCIÓN GENERAL Y OBJETIVOS DEL CURSO

Este curso presenta conceptos y métodos estadísticos para el análisis de datos bajo el enfoque de modelamiento predictivo (“machine learning”). Los modelos predictivos buscan aproximar el valor de una variable de interés en términos de valores particulares de las variables predictoras, usando una función que debe ser estimada (aprendida) a partir de datos. Incluye problemas de regresión (predicción de variable continua, por ejemplo, predecir el monto de una transacción electrónica), clasificación (predicción de variable categórica, por ejemplo, predecir si la transacción es fraude o no), ranking (ordenamiento de la variable de respuesta) y aprendizaje no-supervisado (cuando la respuesta no es observable). A diferencia de modelos estadísticos clásicos, los modelos predictivos se evalúan a partir de su eficacia para pronosticar la respuesta y no necesariamente en su capacidad para explicar las relaciones entre las variables. En el curso se incluyen técnicas de minería de datos, de “machine learning” y de reconocimiento de patrones. El enfoque central es cómo usar estas técnicas para crear métodos que sean buenos predictores desde el punto de vista estadístico.

Los objetivos primarios del curso son:

- Comprender los conceptos generales del modelamiento predictivo: buen uso de los datos, dimensión y complejidad del problema, evaluación y selección de mejores modelos.
- Comprender los conceptos fundamentales de cada técnica vista en el curso, incluyendo sus supuestos y limitaciones.
- Aprender a utilizar herramientas computacionales que permitan la correcta aplicación de los métodos vistos.
- Desarrollar habilidades para el análisis, comprensión y comunicación de resultados de los modelos.

B. CONTENIDO

I. Preliminares y conceptos generales en métodos de aprendizaje estadístico y Machine learning

- Introducción a modelos predictivos: clasificación y regresión. Estimación de funciones
- Aprendizaje supervisado y no-supervisado
- Medidas de Precisión y selección de modelos (sesgo vs. varianza)

II. Modelos lineales para regresión: Selección de modelos y regularización

- Conceptos básicos de modelos de regresión lineal. Explicación vs. Poder predictivo
- Selección de modelos lineales y métodos de regularización (manejo de la dimensionalidad)
- Comparación con modelos no-estructurados

III. Métodos de remuestreo y separación de datos

- Conceptos generales de remuestreo y separación de muestras
- “Cross-validation” y método de “bootstrapping”

IV. Métodos de regresión no-lineales

- Modelos polinomiales y de bases finitas
- Modelos no-paramétricos (suavización y penalización)
- Modelos tipo caja negra: Redes Neuronales
- Modelos aditivos generalizados (GAMS)

V. Modelos de clasificación lineales

- Conceptos básicos de clasificación
- Análisis discriminante lineal y cuadrático
- Regresión Logística

VI. Modelos de clasificación no-lineales

- Máquina de Soporte Vectorial (SVM) y aprendizaje con “kernels”
- Redes neuronales, “Naïve Bayes” y modelos aditivos

VII. Árboles, métodos basados en Reglas, “boosting” y “bagging”

- Árboles para regresión y clasificación y Métodos basados en reglas
- “Bagged trees” y método de “Boosting”
- Random Forest

VIII. Aspectos finales

- Algunos métodos no-supervisados
- Medición de la Importancia de variables predictoras
- Métodos de ensamblaje de meta-algoritmos (boosting)
- Problema de imbalanceo de clases en clasificación

C. BIBLIOGRAFÍA

El curso utilizará la guía de varios libros de referencia que se pueden acceder a través de la biblioteca general (excepto el último).

- **James, G.; Witten, D.; Hastie, T.; Tibshirani, R.** An Introduction to Statistical Learning with Applications in R. Springer, 2013.
- **Kuhn, M.; Johnson, K.** Applied Predictive Modeling. Springer, 2013.
- **Hastie, T.; Tibshirani, R.; Friedman, J.** The Elements of Statistical Learning. Second Edition. Springer, 2011.
- **Schölkopf, B.; Smola, A.** Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Second Edition. MIT Press, 2002.

D. EVALUACIÓN

La nota final del curso se determinará a partir del desempeño en tres tipos de evaluaciones: exámenes escritos presenciales, tareas y quices de refuerzo y proyecto final del curso. Todas las evaluaciones son individuales, aunque el trabajo conjunto es permitido para las tareas si se otorga el crédito correspondiente y se entrega el reporte individual.

- Examen I (20%).
- Examen II (20%).
- Tareas y Quices (40%).
- Proyecto Final (20%).

E. ASPECTOS GENERALES

- **Software y uso de herramientas computacionales.** Durante el curso se hará énfasis en el aspecto práctico de los modelos y herramientas vistas en clase, por lo cual, es indispensable el manejo de software estadístico que permita solucionar problemas basados en datos reales. El paquete estadístico que se utilizará mayormente será **R**, el cual se puede obtener de manera gratuita desde <http://cran.r-project.org> . Otros softwares que pueden ser utilizados en menor medida son SAS y módulos de Phyton. El estudiante es libre de utilizar el software de su preferencia que le permita solucionar los problemas prácticos.
- **Metodología.** El contenido de los temas será presentado por el profesor durante las clases magistrales, además de casos de aplicación y uso de software presentados por el instructor. Se espera que el estudiante realice un trabajo individual de manera semanal para afianzar los conceptos y obtener las habilidades requeridas. Solución de dudas o asesorías extras por parte de los profesores pueden ser programadas por correo electrónico o durante el tiempo de clase.
- **Prerrequisitos.** Conocimientos básicos de probabilidad, estadística y modelos de regresión lineal.
- **Conducta y Código de Honor.** Se espera que todos los estudiantes tengan un comportamiento respetuoso frente a sus compañeros de clase y a sus profesores. Así mismo, se espera completa honestidad en las evaluaciones de forma que estas reflejen el trabajo individual y el conocimiento que se tiene sobre el material del curso. Cualquier violación al reglamento general de estudiantes será notificada y tratada mediante el conducto regular.