

1. Horario atención a estudiantes, correos electrónicos y nombres de los profesores complementarios

Clase magistral

Profesor: María Paula Ríos Arango maria.rios@uniandes.edu.co

Horario de atención a estudiantes: martes 5 p.m.

Lugar de atención a estudiantes: W-705

Profesor complementario: Manuel Felipe Avella Niño, mf.avella47@uniandes.edu.co

Horario de atención a estudiantes: después de la clase complementaria

Lugar de atención a estudiantes: Pendiente

Monitor: Andrés Octavio Davila, ao.davila10@uniandes.edu.co

2. Introducción y descripción general del curso

Este es un curso en Analítica y Big Data, en donde los estudiantes aprenderán: 1) como explorar y analizar conjuntos de datos grandes y de varias dimensiones 2) como convertirse en adeptos en construir sistemas potentes para predecir y 3) desarrollarán el entendimiento necesario para estructurar interpretaciones en estos modelos. Este curso incluye conceptos y herramientas clave que los científicos de datos necesitan en ambientes de negocios. Es una primera aproximación y les da las bases para que continúen estudiando estos temas a mayor profundidad.

No es una introducción a computer science o machine learning. Tampoco es una clase en econometría o estadística de altas dimensiones. Por el contrario, es una clase que toma elementos de varias disciplinas, tal y como lo hacen los buenos científicos de datos.

Las técnicas que se cubrirán a lo largo del curso incluyen una revisión de regresión lineal y regresión logística, selección de modelos y tasas de falso descubrimiento, criterios de información y validación cruzada, regresión regularizada y lasso, bagging y bootstrap, experimentos y estimación causal, regresión multinomial y binaria, clasificación, modelos de variables latentes, análisis de componentes principales, modelos tópicos, árboles de decisión y bosques aleatorios, análisis de texto y procesamiento de lenguaje natural.

Se aprenderán tanto los conceptos básicos subyacentes al igual que habilidades computacionales. Trabajaremos analizando datos reales. Entre algunos de los ejemplos se consideran minería en bases de datos de consumidores, monitoreo de internet y redes sociales, valoración de activos, análisis de redes, análisis de deportes y minería en textos.

3. Objetivos de la materia

El curso tiene por objetivo brindar a los estudiantes nuevas herramientas y métodos para la analítica de grandes datos. Al final del curso los estudiantes sabrán 1) como explorar y analizar conjuntos de datos grandes y de varias dimensiones 2) como convertirse en adeptos en construir sistemas potentes para predecir y 3) desarrollarán el entendimiento necesario para estructurar interpretaciones en estos modelos. Además, a través de ejercicios con datos reales, los estudiantes realizarán de primera mano los análisis para desarrollar las competencias necesarias para convertirse en científicos de datos.

Con respecto al entorno computacional, los estudiantes estarán familiarizados con plataformas de análisis como R que permiten hacer analítica y trabajar con Big Data. También conocerán de otras plataformas que permiten trabajar con Big Data, como Distributed File Systems (DFS).

4. Organización del curso

Los temas por cubrirse durante el curso son (sujeto a cambios):

1. Datos: Principios, gráficas, temas computacionales y falsos descubrimientos
2. Preparación de Datos: exploración, limpieza,
3. Regresión: Una revisión de regresión lineal y logística
4. Selección de modelos: Penalidades, criterios de información, validación cruzada
5. Efectos de tratamientos: controles, puntajes de propensidad y bootstrap
6. Clasificación: Multinomiales, KNN, sensibilidad y especificidad
7. Redes: co-ocurrencias, gráficas dirigidas, Page Rank
8. Clústeres: modelos mixtos, k medias y asociación
9. Factores: Variables latentes, PCA, PCR y PLS
10. Árboles. CART y bosques aleatorios
11. Minería de texto: modelos tópicos, predicción, Deep learning

Antes de dar inicio al curso se llevará a cabo un taller de R opcional, en un horario diferente al de la clase, para ayudar a los estudiantes a familiarizarse con el software usado a lo largo del curso. Probablemente un sábado.

5. Metodología

El contenido se cubrirá con clases magistrales en la que se presentará la teoría y se discutirán ejemplos prácticos.

No obstante, en esta materia es importante aprender haciendo. Por eso en cada clase se asignará una tarea para que los estudiantes puedan aterrizar los conceptos estudiados durante la clase. Además, se llevarán a cabo talleres para ayudar a los estudiantes en el entendimiento de las clases y en la preparación de las tareas.

Al final de la clase habrá un proyecto final

6. Competencias

Al final del curso, los estudiantes estarán en capacidad de:

- a. Manejar la plataforma R para visualizar datos.
- b. Explorar las distintas herramientas analíticas que ofrece R para explorar conjuntos de datos.
- d. Aplicar distintos métodos de analítica de datos según el tipo de problema y de estructura de datos.
- e. Entender los límites y posibilidades del Big Data.

Nota sobre R:

Como software utilizaremos R, una plataforma para realizar análisis estadístico. Es recomendado que los estudiantes aprendan los comandos básicos del software y lenguaje de programación, antes de que la clase inicie.

R es un software gratuito disponible en www.r-project.org. Pueden descargar e instalar el software en cran.us.r-project.org. Es recomendable hacer esto lo antes posible.

R es una plataforma para análisis muy flexible y frecuentemente utilizada. Tiene una interface de comandos (se escriben comando para obtener lo que uno desea). Algunos estudiantes se sentirán retados por la curva de aprendizaje de “programación”. En las clases daremos explicaciones limitadas de R y se brindará código para acompañar las clases y tareas. No obstante, esta no es una clase sobre R.

7. Criterios de evaluación (Porcentajes de cada evaluación)

En este curso las evaluaciones serán solamente las tareas. No habrá proyectos o exámenes.

Todas las clases se asignará una tarea la cual deberá ser entregada la clase siguiente. Todas las tareas se evaluarán y todas tienen el mismo peso relativo para la clase.

De acuerdo con el Reglamento de Estudiantes de Maestría, Art. 41 a 44, el estudiante debe asistir como mínimo al 80% de las clases. Es facultativo de cada profesor controlar la asistencia a sus alumnos y determinar las consecuencias de la inasistencia si esta es superior al 20%. Se aceptan solamente excusas estipuladas en el Artículo 44 del Reglamento.

8. Sistema de aproximación de notas definitiva

Las notas se aproximarán a la centésima con la siguiente excepción: Nota final entre 2.90 y 2.99 se aproxima a 3.0 si el estudiante ha presentado todas las tareas, y ha aprobado al menos el 50% de ellas.

Reclamos (artículo 62 y 63 del reglamento general de estudiantes de maestría):

ART. 62. Todo estudiante que desee formular un reclamo sobre las calificaciones de cualquier evaluación o sobre la nota definitiva del curso deberá dirigirlo por escrito y debidamente sustentado al profesor responsable de la materia, dentro de los ocho (8) días hábiles siguientes a aquel en que se dan a conocer las calificaciones en cuestión. El profesor dispone de diez (10) días hábiles para resolver el reclamo formulado; vencido el término informará al estudiante la decisión correspondiente.

ART. 63. Si el estudiante considera que la decisión no corresponde a los criterios de evaluación, podrá solicitar la designación de un segundo calificador mediante un escrito debidamente sustentado, dirigido al consejo de facultad, dentro de los ocho (8) días hábiles siguientes al conocimiento de la decisión. Si el consejo encuentra fundada la solicitud, procederá a designar, solamente para tal efecto, un segundo calificador cuya decisión debidamente sustentada será definitiva e inmodificable. En ningún caso, el segundo calificador podrá desmejorar la nota inicialmente asignada por el profesor.

Fraude:

Según el artículo 105 del reglamento general de estudiantes de maestría el fraude académico y/o ayuda a cometer fraude son faltas disciplinarias sancionadas con una nota de 0.0 en la prueba en cuestión. En caso de desacuerdo o si se repite la falta se llevará el caso al Consejo de la Facultad de Economía para estudio.

9. Bibliografía

La bibliografía requerida para el curso la encontrarán en la página de SICUA del curso.

A continuación, presento textos referencia en temas de Big Data, pero no son los que seguiremos en detalle para la clase.

1. An Introduction to Statistical Learning, de James, Witten, Hastie, and Tibshirani. No obstante, el libro aproxima el tema de manera un poco diferente y solo hay puntos de encuentro parciales en el material
2. Un texto más avanzado es Elements of Statistical Learning por Hastie, Tibshirani, and Friedman. Requiere de sofisticación matemática y va más allá de los materiales que cubriremos en esta clase. El libro lo pueden descargar de forma gratuita en https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf
3. Otro libro avanzado es are Pattern Recognition & Machine Learning por Bishop, que toma una perspectiva desde ingeniería de sistemas (o Inteligencia Artificial) más que desde una perspectiva estadística
4. Al igual que el 3 el clásico Machine Learning por Murphy presenta un enfoque avanzado desde lo computacional

Para los que no han usado R (y también para los que lo han usado y quieren refrescar) se recomienda familiarizarse con el idioma de programación. Un excelente recurso es *TryR code school* en <http://tryr.codeschool.com>. Aquí hay un demo gratis que permite aprender gratis por 10 días.

Recursos adicionales de R son:

- Los tutoriales en data.princeton.edu/R son muy buenos
- La serie de videos "intro to R" de Google Developers. El primer video de la serie lo encuentran en buscando "R 1.1 Initial Setup and Navigation Google Developers"