Universidad de los Andes Colombia | Facultad de Economía

BIG DATA - MECA

MECA 4107 – Sección 1
Juan José Ospina
2020-13

Clase magistral del 8 de junio al 22 de julio. Horario de clases lunes y miércoles de 6 a 9 p.m. Clases Complementarias, viernes de 7:30 a 9 pm.

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

Profesor: Juan José Ospina
Email: jospina@uniandes.edu.co
Email2: juan.jose.ospina@gmail.com
Atención a estudiantes: con cita previa

Profesor complementario:
Douglas Newball Ramírez
Email: d.newball10@uniandes.edu.co
Atención a estudiantes: con cita previa

2. Introducción y descripción general del curso

Este es un curso en Analítica y Big Data en el cual los estudiantes aprenderán: 1) cómo explorar y analizar conjuntos de datos grandes y de altas dimensiones 2) como convertirse en constructores de sistemas potentes para predecir y 3) desarrollarán el entendimiento necesario para interpretar estructura en los 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. Es una clase que se enfoca en temas que son útiles en el mundo laboral.

Este curso 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, arboles de decisión y bosques aleatorios, análisis de texto y procesamiento de lenguaje natural.

3. Objetivos

Se aprenderán tanto los conceptos básicos subyacentes al igual que habilidades computacionales, incluyendo técnicas para análisis de datos distribuidos. 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.

4. Pre-requisitos

Fundamentos de probabilidad y estadística y entendimiento de regresión lineal.

5. Organización del curso

Temas del curso (sujeto a cambio):

1. Introducción: Computación, dispersión, principios y tasa de falso descubrimiento.
2. Datos: (a) preparación, limpieza. (b) tipos de datos disponibles
3. Regresión: Revisión general, lineal y logística.
4. Selección de Modelo: Penalidades, criterios de información, validación cruzada.
5. Efecto de Tratamiento: Controles, propensity scores, bootstrap.
6. Clasificación: Multinomiales, KNN, sensibilidad/especificidad, DMR.
7. Redes: co-ocurrencia, gráficos dirigidos, rango de página.
8. Arboles: CART y bosques aleatorios, conjuntos (ensambles.)
9. Agrupamiento: Mescla de modelos, k-medias, y reglas de asociación.
10. Factores: Variables latentes, PCA, PCR, y PLS
11. Minería de texto: modelos de temas, predicción de sentimiento, aprendizaje profundo.

6. Metodología

Computación

Esta clase utiliza R, el cual está disponible de forma gratuita vía www.r-project.org El estudiante puede descargar e instalar el software siguiendo las direcciones en cran.us.r-project.org. Se recomienda que esto se haga antes de la primera clase, sin embargo, el profesor asistente los ayudará en su primera clase.

R es una plataforma de análisis ampliamente utilizada y tremendamente flexible. Tiene una interface de comandos. Muchos estudiantes encuentran dificultades para recorrer la curva de aprendizaje de la programación. En este curso el profesor provee una enseñanza limitada del software, se hacen algunas demostraciones en clase y se entregan los códigos que acompañan las clases y las tareas. No es un pre-requisito haber utilizado R antes

Sin embargo, es importante saber que esta no es una clase sobre R, Como cualquier lenguaje de programación, R sólo se aprende haciendo. El estudiante debe instalar R lo antes posible y familiarizarse con las operaciones básicas. Idealmente, el estudiante empezaría el curso siendo capaz de replicar en R un análisis de una clase previa de estadística básica.

Una gran forma de aprender es conseguir un libro y empezar a revisar tutoriales. Una buena guía es "R in a Nutshell" de Adler que tiene varios tutoriales. Si el estudiante es nuevo en R debería hacer un tutorial completo para familiarizarse con el lenguaje. Una gran opción es la escuela de programación TryR disponible en http://tryr.codeschool.com.

7. Criterios de evaluación

Tareas: 80% (Excelente = 5, Muy buena = 4.5, Buena = 4, Ok = 3.00 y Mala = 2.5)

Proyecto final: 20%

No se aceptan tareas o proyectos que sean entregados tarde. Deben ser entregados a través de SICUA.

8. Horarios de clase

Fechas del curso: junio 8 a julio 24 de 2020

Número de horas de clase en ese periodo: 30 horas

Días de clase: lunes y miércoles de 6:00 pm a 9:00 pm

Clase complementaria: viernes 7:30 pm a 9:00 pm

Plataforma: Teams, Zoom o Webex.

9. Asistencia a clases

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 Articulo 44 del Reglamento.

10. Bibliografía

No hay un libro de texto requerido. Todo el material del curso estará disponible en SICUA. La mejor preparación para la clase es revisar el código y trabajar los ejemplos.

Un libro bueno para empezar es An Introduction to Statistical Learning, de James, Witten, Hastie, y Tibshrani. Sin embargo, toma una aproximación diferente a la nuestra.

Un muy buen texto avanzado es Elements of Statistical Learning de Hastie, Tibshrani, and Friedman, pero requiere cierta sofisticación matemática y va más allá del material que estaremos cubriendo en el curso. El libro se puede encontrar en http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

También bueno, pero avanzados son los libros Pattern Recognition & Machine Learning de Bishop y Machine Learning de Murphy, Los dos presentan el material desde una perspectiva más de ingeniería de computación, más que de una perspectiva estadística.