Introducción a técnicas avanzadas para el análisis de datos

Carlos S. Pérez

Otoño, 2018

Lugar: CC101 Horario: Lu 19:00-20:30

Descripción del Curso

En la era digital, individuos y organizaciones producen conjuntos de datos de forma masiva, lo cual implica una necesidad de técnicas y metodologías que permitan analizar los datos de forma sistemática. El uso creciente de herramientas inteligentes y automatizadas provee mejoras claras en la toma de decisiones en términos del volumen y velocidad de datos que pueden manejarse.

En este curso se llevará a cabo un recorrido de principio a fin en la generación de un producto de datos, en el cual se revisarán las tecnologías actuales que permiten recolectar, transformar, visualizar, extraer y modelar información clave para la toma de decisiones. Se tendrá un enfoque práctico en el que el alumno resolverá casos reales en materia de ciencias sociales a través del uso de herramientas y técnicas avanzadas de análisis de datos.

Objetivos del Curso

Al finalizar este curso el alumno deberá

- Haber desarrollado una forma de pensamiento analítico que permite la formulación de problemas de interés bien definidos y el planteamiento de soluciones a través del modelado de datos.
- Poder identificar las fuentes de datos necesarias, así como utilizar herramientas de recolección y almacenamiento que correspondan adecuadamente al problema en cuestión y al alcance de la solución propuesta.
- Realizar visualizaciones, exploración y transformaciones de los datos que permitan identificar anomalías y extraer las características esenciales del fenómeno con el fin de que cumplan con criterios fundamentales para el modelado de datos.
- Identificar los algoritmos y herramientas más apropiados para implementar una solución y asimismo determinar métricas de desempeño que permitan evaluar distintos modelos desde una perspectiva de negocio.
- Entender los procesos de estimación/optimización de parámetros y selección de modelos e interpretar los resultados de forma accionable.
- Implementar una solución reproducible que resuelva una problemática real a través de herramientas de cógido abierto.
- Conocer los errores más comúnes y las implicaciones éticas que existen en el modelado de datos.

Prerrequisitos

- Conocimiento básico de Matemáticas en Álgebra Lineal, Cálculo y Probabilidad
- Experiencia previa con herramientas computacionales (opcional)

Bibliografía

- TBD

Estructura del curso

Este itinerario es tentativo y está sujeto a cambios. Los objetivos de aprendizaje son conceptos claves que ayudan a delinear una guía para las evaluaciones.

Semana 1, 08/10 - 12/10: El pensamiento analítico

- Panorama histórica del análisis de datos
- Perfiles, herramientas y metodología en el proceso analítico
- Productos de datos y entendimiento del negocio
- Formulación de problemas y diseño de experimentos
- Identificación de insumos y propuestas de valor (mediante máquinas)
- Lab: Problem Framing

Semana 2, 15/10 - 19/10: Recolección y Almacenamiento de Datos

- Propiedades básicas de los datos estructurados y no estructurados
- Breviario sobre Gobernanza de Datos, Big Data y Arquitecturas de Información
- Tipos de Funciones y Semántica de manipulación en bases de datos
- APIs y protocolos de comunicación web
- Lab: Data Scrapping and Wrangling (Python/SQL)

Semana 3, 22/10 - 26/10: Exploración de Datos

- Motivación y objetivos del análisis exploratorio
- Estadística descriptiva y detección de valores atípicos
- Técnicas de visualización y otras consideraciones sobre la experiencia del usuario
- Lab: Exploratory Data Analysis (R/Python)

Semana 4, 29/10 - 02/11: Manipulación de Datos

- Transformaciones lineales y no lineales
- Extracción de variables y reducción de dimensiones
- Selección de variables: Medidas de información y Técnicas de regularización*
- Lab: Feature Engineering & Feature Selection (R/Python/SQL)

Semana 5, 05/11 - 09/11: Aprendizaje de Máquina

- Panorama histórico del aprendizaje de máquina
- Metodología general, optimización e inferencia causal en el análisis de datos
- Aprendizaje supervisado: regresión, naive bayes, árboles, SVMs y redes neuronales
- Aprendizaje no supervisado: K-means, Métodos jerárquicos, LDA (Text Classification)
- Lab: Unsupervised Learning (R/Python)

Semana 6, 12/11 - 16/11: Selección de Modelos y Ensemble Learning

- Funciones de pérdida y métricas de desempeño
- Fuentes del sesgo y varianza en la estimación de los errores
- Elasticidades, Estabilidad y Causalidad
- Ensemble Learning: Bagging and Boosting
- Lab: Supervised Learning (R/Python)