Diplomado Data Science - Machine Learning e Inteligencia Artificial

Temario

Hamdi Raissi, PhD Universidad de Lille, Francia, Profesor Auxiliar PUCV
Patricio Videla, Profesor de planta PUCV, Jefe de docencia del Instituto de Estadística
Mario Guzmán, Data Scientist
Pablo Rodríguez, MBE U. de Chile

Software: R, Python, Spark, SQL. **Se incluye también 50USD de uso del servicio Elastic Cloud Computing (EC2) de Amazon Web Services (AWS).** No se necesita conocimientos previos de los softwares dado que una introducción será hecha para cada software ocupado. Los códigos listos para el uso y comentados en la clase.

Fechas: 20, 24, 27, 31 de agosto, 3, 7, 10, 14, 21, 24, 28 de septiembre, 1, 5, 8, 15, 19, 22, 26, 29 de octubre y 2 de noviembre. Todas las clases son de 3 horas y empiezan a las 19hrs.

Los conceptos presentados en clase serán cada vez ilustrados con datos reales o simulados.

Temas
Básicos

1. Estadística descriptiva y introducción a R (2 horas)
    a. Como utilizar R, funciones básicas, estrategias para elegir los paquetes R.
    b. Estadísticas descriptivas y su visualización.
    c. Tipos de variables en los datos.

2. Toma de decisión en un entorno aleatorio (2 horas)
    a. Test estadístico.
    b. Intervalos de confianza para pronósticos.

3. Análisis de asociación de variables (5 horas)
    a. Estrategias para medir la correlación entre variables: Pearson, Spearman o Kendall?

b. Modelos lineales simples: Estimación MCO, Diagnóstico de bondad. Test de normalidad.
c. One way ANOVA y two way ANOVA, razón de correlación.

4. Reducción de la dimensión: Análisis por Componentes Principales (ACP) (3 horas)

Temas Avanzados

1. Modelos lineales múltiples (12 horas)
    a. Estimación MCO, diagnóstico de bondad (t-test, test de Fisher) y tipos de predicción (individual y del fenómeno estudiado).
    b. Test de homogeneidad poblacional de Chow
    c. Identificación de las variables pertinentes (Cp de Mallows, Criterios de información, algoritmos de selección forward, stepwise y backward). Como introducir las variables categóricas en un modelo lineal.
    d. Problema de colinealidad y soluciones (regresión PCR, regresión Ridge)
    e. Datos outliers (atípicos): detección y diagnostico (leverages, residuos studentizados, distancia de Cook, DFBETAS). Solución con la estimación robusta de Theil-Sen y Siegel.
    f. Heteroscedasticidad y autocorrelación: diagnóstico (test de Durbin Watson, tests deBreusch-Pagan) y estimación MCG.

2. Métodos numéricos de alto nivel computacional (6 horas)
    a. Introducción a EC2 de AWS.
    b. Métodos bootstrap.
    c. Experimentos de Monte Carlo.

3. Introducción a SQL (3 horas)
    a. Modelos relacionales.
    b. Transformación de la información.
    c. Conexión con diferentes bases de datos.
    d. Depuración.
    e. Estudio de caso.

4. Introducción a Spark (3 horas)
    a. Tratamiento de data frame.
    b. Análisis descriptivo.
    c. Categorización de bases.

d. Rutinas de Spark.

5. Algoritmo de K-medias (3 horas)
    a. Medidas de similaridades.
    b. Identificación del número de conglomerados.
    c. Métricas de validación.

6. Árboles de decisión (3 horas)
    a. Clasificación del árbol.
    b. Requisitos y supuestos de los datos.
    c. Interpretación de los resultados.
    d. Predicción y Evaluación.
    e. Aplicación de un caso real en R.

7. Random Forest (3 horas)
    a. Introducción al Random Forest.
    b. Entrenamiento de un modelo Random Forest.
    c. Evaluación de out-of-bab error.
    d. Evaluación del rendimiento del modelo Random Forest.
    e. Estudio de caso en R.

8. Modelo de Regresión Logística (3 horas)
    a. Presentación del modelo e interpretación.
    b. Validación de supuestos.
    c. Ajuste del Modelo e interpretación de resultados.
    d. Estudio de caso aplicado en R: Evaluación y Construcción.

9. Máquinas de vectores de soporte (3 horas)
    a. Definición de hiperplano de separación.
    b. Clasificador de margen máximo.
    c. SVM para clasificador linealmente separable.
    d. SVM para clasificador linealmente no separable.
    e. Extensión de las máquinas de vectores de soporte.
    f. Métricas de validación.

10. Redes neuronales (3 horas)
    a. Arquitectura de una red.
    b. Perceptrón.
    c. Función de activación.
    d. Back-propagation.
    e. Métricas de validación.

11. Text mining (3 horas)
    a. Homologación de textos en base a cercanía de textos.
    b. Arquitectura del web scraping.
    c. Aplicaciones de web scraping y cercanía de textos.

12. Web marketing (3 horas)
    a. Marketing en un mundo digital.
    b. Analítica digital.
    c. Aplicaciones de estadística en web marketing.
    d. Estrategias de marketing usando big data.