

DIPLOMADO DATA SCIENCE

*Se incluye acceso a la plataforma cloud de Amazon Web Services (AWS).

PRESENTACIÓN DEL DIPLOMADO

Miércoles 3 :: Mayo :: 2023
19.00 HRS.

CONTACTO

diplomado.estadistica@pucv.cl

PROFESORES

HAMDI RAISSI, PhD Universidad de Lille, Francia, Profesor Adjunto PUCV
PATRICIO VIDELA, Profesor Auxiliar PUCV, Jefe de docencia del Instituto de Estadística
MARIO GUZMÁN, Data Scientist

CLASES

Mayo 8 - 10 - 15 - 17 - 22 - 24 - 29 - 31

Junio 5 - 7 - 12 - 14 - 19 - 28

Julio 3 - 5 - 10 - 12 - 17 - 19 - 24 - 26 - 31

Agosto 2 - 7 - 9 - 21 - 23 - 28

Todas las clases son de 3 horas y empiezan a las 19 hrs. en modalidad "online"***

MACHINE LEARNING E INTELIGENCIA ARTIFICIAL, DEEP LEARNING

SOFTWARE: R, PYTHON, SPARK, SQL.*

TEMARIO

TEMAS BÁSICOS

1. ESTADÍSTICA DESCRIPTIVA Y INTRODUCCIÓN A R

- Como utilizar R, funciones básicas, estrategias para elegir los paquetes R.
- Estadísticas descriptivas y su visualización.
- Tipos de variables en los datos.

2. TOMA DE DECISIÓN EN UN ENTORNO ALEATORIO

- Test estadístico.
- Intervalos de confianza para pronósticos.

3. ANÁLISIS DE ASOCIACIÓN DE VARIABLES

- Estrategias para medir la correlación entre variables: Pearson, Spearman o Kendall?
- Modelos lineales simples: Estimación MCO, Diagnóstico de bondad. Test de normalidad.
- One way ANOVA y two way ANOVA, razón de correlación.

4. MÉTODOS MULTIVARIADOS EN ESTADÍSTICA

- Análisis por componentes principales (ACP)
- Análisis de correspondencias (AC)
- Análisis de correspondencias Múltiples (ACM)

TEMAS AVANZADOS

1. MODELOS LINEALES MÚLTIPLES

- Estimación MCO, diagnóstico de bondad (t-test, test de Fisher) y tipos de predicción (individual y del fenómeno estudiado).
- Test de homogeneidad poblacional de Chow
- 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.
- Problema de colinealidad y soluciones (regresión PCR, regresión Ridge)
- Datos outliers (atípicos): detección y diagnóstico (leverages, residuos studentizados, distancia de Cook, DFBETAS). Solución con la estimación robusta de Theil-Sen y Siegel.
- Heteroscedasticidad y autocorrelación: diagnóstico (test de Durbin Watson, tests de Breusch-Pagan) y estimación MCG.

2. MÉTODOS NUMÉRICOS DE ALTO NIVEL COMPUTACIONAL

- Introducción a EC2 de AWS.
- Métodos bootstrap.

3. MODELOS PARA DATOS TEMPORALES (6 HORAS)

- Modelamiento univariado de datos temporales con modelos AR, MA y ARMA.
- Una caja de herramientas para el modelamiento de datos temporales:
 - Identificación: Autocorrelaciones (ACF), Autocorrelaciones parciales (PACF), Criterios de información
 - Estimación: Menos Cuadrados Ordinarios (MCO), Máximo de verosimilitud
 - Diagnóstico y predicción
- Modelos SARIMA

4. MODELIZACIÓN DE RENDIMIENTOS FINANCIEROS

- Hechos estilizados de las series de tiempo
 - Reagrupación de los valores extremos (Volatility clustering)
 - Leptocurticidad
 - Asimetría
- Modelos GARCH y extensiones
- Detección de la naturaleza financiera de datos dependientes
- Medir los riesgos en finanza:
 - Valor en Riesgo (Value-at-Risk, VaR), VaR condicional
 - Técnicas bootstrap y Monte Carlo para mediciones de riesgos a horizonte más grande que uno
 - Backtesting de las medidas de riesgo

5. INTRODUCCIÓN A SQL

- Modelos relacionales.
- Transformación de la información.
- Conexión con diferentes bases de datos.
- Depuración.
- Estudio de caso.

6. INTRODUCCIÓN A SPARK

- Tratamiento de data frame.
- Análisis descriptivo.
- Categorización de bases.
- Rutinas de Spark.

7. ALGORITMO DE K-MEDIAS

- Medidas de similitudes.
- Identificación del número de conglomerados.
- Métricas de validación.

8. ÁRBOLES DE DECISIÓN

- Clasificación del árbol.
- Requisitos y supuestos de los datos.
- Interpretación de los resultados.
- Predicción y Evaluación.
- Aplicación de un caso real en R.

9. RANDOM FOREST

- Introducción al Random Forest.
- Entrenamiento de un modelo Random Forest.
- Evaluación de out-of-bag error.
- Evaluación del rendimiento del modelo Random Forest.
- Estudio de caso en R.

10. MODELO DE REGRESIÓN LOGÍSTICA

- Presentación del modelo e interpretación.
- Validación de supuestos.
- Ajuste del Modelo e interpretación de resultados.
- Estudio de caso aplicado en R: Evaluación y Construcción.

11. MÁQUINAS DE VECTORES DE SOPORTE

- Definición de hiperplano de separación.
- Clasificador de margen máximo.
- SVM para clasificador linealmente separable.
- SVM para clasificador linealmente no separable.
- Extensión de las máquinas de vectores de soporte.
- Métricas de validación.

12. REDES NEURONALES

- Arquitectura de una red.
- Perceptrón.
- Función de activación.
- Back-propagation.
- Métricas de validación.

13. TEXT MINING

- Homologación de textos en base a cercanía de textos.
- Arquitectura del web scraping.
- Aplicaciones de web scraping y cercanía de textos.

14. MANEJO DE HERRAMIENTAS DE AWS (6 HORAS)

- Introducción a S3.
- Gestión de permisos con IAM.
- Redes virtuales en la nube VPC.
- Introducción a SageMaker.
- Rutinas de modelos de ML en SageMaker.

15. SISTEMAS DE RECOMENDACIÓN (6 HORAS)

- Filtros colaborativos
- Sistema basado en usuarios e items.
- Aplicaciones de sistemas de recomendación.

16. DEEP LEARNING (6 HORAS)

- Introducción al Deep Learning.
- Redes convolucionales (CNN).
- Arquitectura Alexnet.
- Aplicaciones de CNN con torch.h

* No se necesita conocimientos previos de los software dado que una introducción será hecha para cada software ocupado. Los códigos listos para el uso y comentados en la clase.

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