

DIPLOMADO DATA SCIENCE- MACHINE LEARNING E INTELIGENCIA ARTIFICIAL, DEEP LEARNING

» MODALIDAD ON-LINE

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

» **Presentación del diplomado**
Lunes 15 de Noviembre, 19:00 hrs.

FECHA DE CLASES

17, 22, 29 de noviembre
1, 6, 13, 15, 20, 22 de diciembre
3, 5, 10, 12, 17, 19 de enero
21, 23, 28 de febrero
2, 7, 9, 14, 16, 21, 23 de marzo
4, 6, 11 y 13 de abril

» Horario 19:00 a 22:00 hrs.

CONTACTO:
diplomado.estadistica@pucv.cl

INSTITUTO DE
ESTADÍSTICA



PONTIFICIA
UNIVERSIDAD
CATÓLICA DE
VALPARAÍSO

DIPLOMADO DATA SCIENCE- MACHINE LEARNING E INTELIGENCIA ARTIFICIAL, DEEP LEARNING TEMARIO

SOFTWARE: R, PYTHON, SPARK, SQL.*

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

PRESENTACIÓN DEL
DIPLOMADO
15 DE NOVIEMBRE
A LAS 19.00 hrs

CONTACTO
diplomado.estadistica@pucv.cl

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: 17, 22, 29 de noviembre, 1, 6, 13, 15, 20, 22 de diciembre, 3, 5, 10, 12, 17, 19 de enero, 21, 23, 28 de febrero, 2, 7, 9, 14, 16, 21, 23 de marzo y 4, 6, 11 y 13 de abril. Todas las clases son de 3 horas y empiezan a las 19 hrs en modalidad "online".** TOTAL: 87 horas.

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.
- Tiempo 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. REDUCCIÓN DE LA DIMENSIÓN: ANÁLISIS POR COMPONENTES PRINCIPALES (ACP)

3. MODELOS PARA DATOS TEMPORALES

- 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
- Big data aplicada a la finanza: Uso de Elastic Cloud Computing (EC2) de AWS Amazon.

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. d. 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.

12. MANEJO DE HERRAMIENTAS DE AWS

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

13. SISTEMAS DE RECOMENDACIÓN

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

14. DEEP LEARNING

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

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.
- Experimentos de Monte Carlo.

*No se necesitan 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 SIMULADOS

INSTITUTO DE
ESTADÍSTICA



PONTIFICIA
UNIVERSIDAD
CATÓLICA DE
VALPARAÍSO