Diplomado Introducción Analítica a la Ciencia de Datos

Coordinador: M. en C. Jaime Vázquez Alamilla.

Cuerpo docente

1. Actuario Eduardo Selim Martínez Mayorga. Profesor de Asignatura, Facultad de Ciencias, UNAM.

2. Maestro José Salvador Zamora Muñoz. Profesor de Asignatura, Facultad de Ciencias, UNAM.

3. Doctor Carlos Erwin Rodríguez Hernández-Vela. Investigador de Carrera, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.

4. Doctor Jorge Ignacio González Cázares. Investigador de Carrera, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.

5. Doctor Alan Riva Palacio Cohen. Investigador de Carrera, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.

6. Doctora María Fernanda Gil Leyva Villa. Investigadora de Carrera, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.

7. Doctora Claudia Ivonne Juárez Gallegos. Profesora de Asignatura, Facultad de Ciencias, UNAM.

8. Doctora Sandra Palau Calderón. Investigadora de Carrera, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.

TEMARIO

MÓDULO 1. Programación para Ciencia de Datos (24 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
I. Fundamentos de R	1. R como calculadora 2. Objetos básicos de R 3. Sentencias de control 4. if y else 5. switch 6. ifelse 7. Ciclos 8. Funciones adhoc	3	Eduardo Selim Martínez Mayorga
II. Iteraciones en R	1. Ciclos for 2. Ciclos while 3. Iteración la familia de funciones apply 4. Iteración con librería {purrr} 5. Funciones de mapeo 6. Comparación con la familia apply 7. Manejo de listas	6	Eduardo Selim Martínez Mayorga
III. Tidyverse y flujo de ciencia de datos	1. Importación 2. Limpieza 3. Transformación 4. Modelación 5. Visualización 6. Comunicación	9	Eduardo Selim Martínez Mayorga
IV. Sistemas de manejo de versiones	1. Git 2. GitHub 3. GitLab	3	Eduardo Selim Martínez Mayorga
V. Algunas integraciones populares	1. Integración con python vía {reticulate} 2. Nociones de programación en paralelo en R	3	Eduardo Selim Martínez Mayorga

Módulo 2. Lectura y limpieza de datos (20 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
I. Lectura de diversas fuentes de datos en R	1. SQL y MariaDB 2. noSQL: Mongo, Casandra, 3. STATA, SAS, SPSS 4. PDF, EXCEL 5. XML 6. JSON	6	Eduardo Selim Martínez Mayorga
II. Datos web y web scraping	1. Un centavo de HTML 2. Librería {rvest} 3. Selector Gadget de Google Chrome 4. Introducción a expresiones regulares	6	Eduardo Selim Martínez Mayorga
II. Principios de datos limpios	1. tidyr 2. Transformación y limpieza con dplyr 3. Transformación y limpieza con data.table 4. Transformación y limpieza con janitor	8	Eduardo Selim Martínez Mayorga

Módulo 3. Análisis exploratorio de datos (26 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
I. Medidas de tendencia central	1. Media 2. Media truncada 3. Mediana 4. Cuantiles	3	Eduardo Selim Martínez Mayorga
II. Medidas de variación	1. Rango 2. Rango intercuartil	3	Eduardo Selim Martínez Mayorga

	3. Varianza y desviación estándar		
III. Medidas de asociación	1. Covarianza y correlación 2. Correlación de Spearman rank-order 3. Relative risk y odds ratio 4. V de Cramer 5. Faceting	6	Eduardo Selim Martínez Mayorga
IV. Datos categóricos	1. Conteos 2. Frecuencias relativas 3. Tabla de contingencia	4	Eduardo Selim Martínez Mayorga
V. Análisis exploratorio de datos automatizado	1. dataMaid 2. DataExplorer 3. SmartEDA 4. skimr	4	Eduardo Selim Martínez Mayorga
VI. Graficación y visualización de datos	1. ggplot2 2. Plotly 3. Leaflet (mapas) 4. Grafos	6	Eduardo Selim Martínez Mayorga

Módulo 4: Elementos de probabilidad y estadística para ciencia de datos (18 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
1.Fundamentos de Probabilidad	1. Fenómenos aleatorios, incertidumbre y variabilidad 2. Fundamentos de probabilidad 3. Variables aleatorias y sus distribuciones 4. Límite de promedios de variables aleatorias	6	Carlos Erwin Rodríguez Hernández-Vela
2. Fundamentos de Inferencia estadística	1. De probabilidad a estadística 2. Inferencia no paramétrica 3. Inferencia clásica 4. Inferencia bayesiana 5. Ejemplos con R	6	Carlos Erwin Rodríguez Hernández-Vela

3. Introducción a Modelos Lineales	1. Modelo de regresión lineal simple 2. Modelo de regresión lineal múltiple 3. Ejemplos con R	6	Carlos Erwin Rodríguez Hernández-Vela

Módulo 5: Aprendizaje supervisado (44 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
1.Introducción al aprendizaje supervisado	1. Datos en entrenamiento y prueba 2. Validación cruzada y estimación del error de predicción 3. Selección de modelos	10	Jorge Ignacio González Cázares
2. Modelos lineales	1. Regresión ● Categorical inputs ● Regresión regularizada (ridge, LASSO, etc.) ● Regresión no lineal ● Árboles de regresión ● Modelos lineales generalizados 2. Clasificación ● Regresión logística ● Análisis de discriminante lineal ● Bayes ingenuo ● Árboles de clasificación	12	Alan Riva Palacio Cohen
3. Árboles de decisión y bosques aleatorios	1. Bagging 2. Ensemble methods 3. Boosting: adaboost & xgboost	12	Jorge Ignacio González Cázares

4. Redes neuronales	1. Redes superficiales y profundas 2. Redes convolucionales y recurrentes 3. Autoencoders y redes generativas 4. Modelos de difusión 5. Support vector machines	10	Alan Riva Palacio Cohen

Módulo 6. Aprendizaje no-supervisado (30 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
I. Clustering	Determinación del número de clusters Vecinos más cercanos Clustering jerárquico	10	José Salvador Zamora Muñoz
II. Reducción de dimensionalidad	Descomposición SVD PCA (Componentes principales) t-SNE	10	José Salvador Zamora Muñoz
III. Análisis de factores	Análisis de factores exploratorio Análisis de factores confirmatorio	10	José Salvador Zamora Muñoz

Módulo 7: Optimización en aprendizaje de máquina (40 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
1. Introducción a la optimización	1. Modelos y funciones de pérdida 2. Panorama de optimización 3. Búsqueda en rejilla (grid search) 4. Métodos de primer orden	10	Sandra Palau Calderón

	5. Métodos de segundo orden		
2. Descenso del gradiente y sus variaciones	1. Descenso del gradiente 2. Descenso del gradiente estocástico 3. Variaciones adaptativas 4. Variaciones aleatorizadas	10	María Fernanda Gil Leyva Villa
3. Optimización convexa y no-convexa	1. Optimización de funciones convexas 2. Optimización de funciones no-convexas 3. Efecto de la inicialización en la optimización 4. Efecto de la tasa de aprendizaje en la optimización 5. Efecto de otros hiperparámetros en la optimización	10	Sandra Palau Calderón
4. Otros paradigmas de optimización.	1. Aprendizaje en línea 2. Aprendizaje por refuerzo	10	María Fernanda Gil Leyva Villa

Módulo 8: Comunicación de resultados (40 horas)

TEMA	Subtemas	No. de horas	Ponente(s)
1. Generación de documentos con RMarkdown	1. Sintaxis básica de lenguaje de marcado ligero 2. Generación de reportes HTML/PDF/WORD 3. Generación de presentaciones 4. Generación de notebooks de R	6	Claudia Ivonne Juárez Gallegos
2. Generación de dashboard simples (flexdashboards)	1. Introducción a flexdashboard 2. Estructura general de un dashboard 3. Personalización y diseño de dashboards	8	Claudia Ivonne Juárez Gallegos

3. Documentos de larga extensión con bookdown	1. Introducción a bookdown 2. Estructura general de de un documento con bookdown 3. Referencias y citas	6	Claudia Ivonne Juárez Gallegos
4. Introducción a blogdown	1. Introducción a blogdown 2. Creación y personalización de sitio web 3. Publicación del sitio web	8	Claudia Ivonne Juárez Gallegos
5. Aplicaciones interactivas y dashboards interactivos con Shiny	1. Introducción a Shiny 2. Elementos para la visualización dinámica 3. Publicación de la aplicación	12	Claudia Ivonne Juárez Gallegos