Programa abierto de **Complementación y ampliación de la currícula de la Maestría 2017**

Maestría en Generación y Análisis de Información Estadística

UNTREF
UNIVERSIDAD NACIONAL
DE TRES DE FEBRERO

Programa abierto de
Complementación y ampliación de la currícula de la Maestría 2017

Maestría en Generación y Análisis de Información Estadística

La Maestría en Generación y Análisis de Información Estadística, en el marco de su Programa de Actualización Permanente de las Orientaciones de Estadísticas Económicas, Estadísticas Sociodemográficas y Estadísticas de Opinión y Mercado, en coordinación con la Secretaría de Extensión Universitaria y la Dirección de Posgrado, presenta el Ciclo de Seminarios y Cursos extracurriculares en Estadística.

Coordinador: Jorge Fernández Bussy

Curso: "Fundamentos de la programación estadística y Data Mining en R. Estadística descriptiva, modelos de regresión y árboles de decisión"

Docente: Dr. Germán Rosati

Presentación y objetivo del curso

Debido a su carácter de software libre y a la creciente comunidad de usuarios el lenguaje R se ha convertido en algo así como la *lingua* franca dentro del análisis estadístico. El presente seminario se propone realizar una introducción a algunos conceptos fundamentales de la programación estadística en R. A su vez, se hará énfasis en la implementación de análisis estadísticos básicos (descriptivos y regresiones) en R, se presentarán algunos elementos teóricos de la minería de datos/aprendizaje automático (balance sesgo-variancia, overfitting, etc.) y se revisarán algunos algoritmos para la estimación de árboles (ID4, C4.5, CART y random forests).

UNTREF
UNIVERSIDAD NACIONAL
DE TRES DE FEBRERO

El curso se propone que los alumnos

• Se familiaricen con aspectos relevantes de la programación estadística en lenguaje R

• Logren implementar e interpretar análisis estadísticos descriptivos y modelos de regresión en lenguaje R

• Incorporen algunos conceptos fundamentales del data mining/aprendizaje automático

• Conozcan generalidades de algunos algoritmos para la generación de árboles de decisión (ID4, C4.5,CART y random forests) y su implementación en lenguaje R

• Logren identificar situaciones de aplicación de este tipo de modelos a problemas de investigación básica y aplicada.

Destinatarios

Estudiantes avanzados de carreras de grado y posgrado, técnicos, profesionales, investigadores, docentes y no docentes.

Temario de clases

Unidad 1. Elementos de programación estadística en R. Objetos en R (vectores, matrices, data frames y listas). Estructuras de control (loops –for, while, repeat– if, ifelse). Implementación de funciones ad-hoc. Generación de números aleatorios y distribuciones de probabilidad. Análisis estadístico básico en R. Generación de gráficos y visualización de datos (función plot y paquete gráfico base). Importación y exportación de datos (.csv, .txt, .tab, .sav, etc.).

Unidad 2. Nociones básicas de data mining/aprendizaje automático. Tipos de problemas en aprendizaje supervisado: clasificación y regresión. Error de entrenamiento (training error), error de prueba (test error). Sobre-ajuste. Balance entre el sesgo y la variancia de un modelo. Métodos de estimación del error: partición del dataset, validación cruzada. Aplicaciones en R.

Unidad 3. Estadística descriptiva. Implementación y análisis de modelos de regresión lineal y logística. Evaluación del modelo: supuestos, ajuste, estimación de error de generalización. Extensiones del modelo lineal y logístico: variables cualitativas, no linealidad, etc. Funciones lm, glm y predict.

Unidad 4. Clasificadores basados en árboles: generalidades. Algoritmos lD4, C4.5 y CART. Partición múltiple y binaria, medidas de pureza de nodos. Crecimiento (growing) y podado (prunning) de árboles de decisión. Balance entre costo y complejidad del árbol. Introducción a los modelos de random forests. Aplicaciones en R (paquetes tree, rpart y random forests).

Bibliografía básica de referencia

Breiman, L., Friedman, J., Stone, C. y Olshen, R. (1984), *Classification and Regression Trees,* New York: Champan & Hall/CRC.

Breiman, Leo (2001), "Statistical modelling. The two cultures", *Statistical Science,* Vol. 16, nº3: 199-215.

Hastie, T.; Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction,* Berlin: Springer.

Hernández Orallo, J., Ramírez Quintana, J., Ramírez, C., (2004). *Introducción a la minería de datos,* España: Pearsons Editorial.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R,* Berlin: Springer.

R Core Development Team, (2000), *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos,* (disponible en https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf)

Tetor, P. (2011), *R Cookbook. Proven recipes for data analysis, statistics and graphics,* New York: O´Reilly.

Requisitos para la cursada y aprobación

Conocimientos básicos de estadística descriptiva y cierta familiaridad con modelos de regresión lineal y logística. Será útil (pero no absolutamente necesario) alguna experiencia en programación estadística (sea en SPSS, Stata o similar)

Para la aprobación del curso se requiere

1) un mínimo de asistencia del 80% sobre el total de clases y
2) la entrega y aprobación de una monografía final.

Organización del curso

Modalidad: Presencial
Días y horario: Miércoles 19/07, viernes 21/07, miércoles 26/07, viernes 28/07, jueves 03/08 en el horario de 18:00 a 22:00 hs.
Lugar de cursada: Sede Centro Cultural Borges, Viamonte y San Martín, Pabellón de las Naciones, 3er piso, CABA.
Fecha de inicio: 19/07/17
Fecha de finalización: 03/08/17
Cantidad de clases: 5
Total de horas: 20 hs

Informes e inscripción a: *maestriaestadistica@untref.edu.ar*