PRØMiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Métodos Avanzados en Minería de Datos

(506) 2268.8823 - (506) 8708.9091 info@promidat.com facebook.com/oldemarrodriguez www.promidat.com

Tutor: El curso será impartido por Dr. Oldemar Rodríguez graduado de la Universidad de París IX y con un postdoctorado en Minería de Datos de la Universidad de Stanford.

Duración: Cuatro semanas.

Descripción:

En este curso se estudiarán en detalle las técnicas de Validación Cruzada (Cross-Validation) y Remuestreo (boostrapping) con el objetivo de calibrar y seleccionar el mejor método de Minería de Datos para un problema y juego de datos dado.

Se estudiará como programar y automatizar la Validación Cruzada (Cross-Validation) y Remuestreo (boostrapping) en el lenguaje R.

Los ejemplos de este curso estarán motivados por problemas reales en el campo. Por lo tanto, los estudiantes adquieren conocimientos de muchas herramientas diferentes que pueden combinarse para resolver problemas reales.

Objetivos:

El objetivo fundamental de este curso es que el estudiante sea capaz de manipular y generar datos con facilidad, para esto será capaz de:

1. Escribir programas en R para Validación Cruzada (Cross-Validation) y Remuestreo (bootsrapping).
2. Utilizar bases de datos como MySql, SQLite, PostgreSQL y SQLServer para procesar datos.
3. Aplicar adecuadamente una Validación Cruzada y un Remuestreo.
4. Calibrar adecuadamente tanto métodos exploratorios como métodos predictivos en Minería de Datos.
5. Seleccionar el mejor modelo predictivo dado un conjunto de datos.
6. Instalar y utilizar paquetes avanzados en R.

Metodología:

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas.

- Una vídeo conferencia semanal, las cuales quedarán grabadas en Webex, para que los alumnos la puedan acceder en cualquier momento.
- Trabajos prácticos semanales.
- Foros para plantear dudas al tutor y compañeros.
- Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren métodos exploratorios y predictivos avanzados directamente usando el lenguaje R.

Contenido:

1) Implementación de métodos Métodos Descriptivos (Clustering o Aprendizaje no Supervisado)
    a. Biplots Usando prcomp de {stats}
    b. Análisis en Componentes Principales
    c. Clustering Jerárquico
    d. El método de K-medias
2) Implementación de Métodos de Predictivos (Clasificación o Aprendizaje-Supervisado)
    a. El método de los K vecinos más cercanos
    b. Método de Bayes
    c. Análisis Discriminante Lineal y Cuadrático
    d. Máquinas Vectoriales de Soporte
    e. Árboles de Decisión
    f. Bosques Aleatorios (Random Forest)
    g. Métodos de Potenciación (Boosting)
    h. Redes Neuronales
    i. Guardando y Recuperando en disco un Modelo para su posterior uso
    j. Implementación de Curvas ROC
3) Validación Cruzada (cross-validation) y Remuestreo (bootstrapping)
    a. Enfoque: “tabla de aprendizaje y tabla de testing” (the validation test approach)
    b. Validación cruzada dejando uno fuera (Leave-one-out cross-validation - LOOCV)
    c. Validación cruzada usando K grupos (K-fold cross-validation)
    d. El enfoque del “Bootstrap”
    e. Un ejemplo de Remuestreo (bootstrapping)

4) Calibración y Selección de Métodos
    a. Calibración de Métodos Exploratorios (descriptivos)
        - Calibrando el método de las k-medias
    b. Calibración de Métodos Predictivos
        - Calibrando el método Máquinas Vectoriales de Soporte
    c. Seleccionando el mejor método predictivo

Evaluación:

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínimima de aprobación es de 70.

Bibliografía:

1. John M. Chambers. Programming with R: Software for Data Analysis. Springer, Stanford University, Palo Alto, 2008.
2. Graham Williams, Data Mining with Rattle and R. Springer, New York, 2011.
3. Owen Jones, Robert Maillardet and Andrew Robinson. Introduction to Scientific Programming and Simulation using R. Chapman & Hall/CRC Taylor & Francis Group, FL. 2009.
4. R Development Core Team. “R: A Programming Environment for Data Analysis and Graphics”. The R Project for Statistical Computing, 2010. http://www.r-project.org/
5. R Development Core Team. “Writing R Extensions”. The R Project for Statistical Computing, 2010. http://www.r-project.org/