PRØMiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Métodos Exploratorios 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 presentarán los principales conceptos y métodos en Minería de Datos. El énfasis principal del curso será examinar dichos métodos desde un punto geométrico y de sus aplicaciones concretas. Se le dará especial importancia al uso de los conceptos de minería de datos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán los programas especializados en Minería de Datos como **discove*R*** sobre la plataforma de software libre *R* y RStudio.

Objetivos:

En este curso el estudiante será capaz de:

1. Entender la necesidad de la utilización de modelos, algoritmos, software especial para el descubrimiento de conocimiento en grandes volúmenes de datos.
2. Conocer la Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.
3. Conocerá la metodología del ciclo de desarrollo usado para el descubrimiento del conocimiento en grandes bases datos (KDD – “Knowledge Discovery in Databases”).
4. Entender las diferencias entre: estadística, análisis de datos, recuperación de la información, ML – “Machine Learning”, minería de datos y Ciencia de Datos.
5. Conocer los principales modelos, técnicas y algoritmos utilizados para descubrir el conocimiento en grandes volúmenes de datos.
6. Utilizar el **discove*R*** sobre la plataforma *R* para analizar ejemplos con datos reales.

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 segmetación de carteras de clientes.

Contenido:

1. Conceptos de la Minería de Datos

a. Definiciones básicas en Ciencia de Datos
b. Instalación de la Plataforma R y RStudio
c. Instalación del paquete **discove*R***

2. Análisis Exploratorio de Datos

a. Tipos de variables
b. Estadísticas básicas y matriz de correlaciones
c. Tablas de datos y datos atípicos
d. Aplicaciones en casos reales con el paquete **discove*R*** sobre la plataforma *R*

3. Métodos de condensación de la información

a. Análisis en Componentes Principales – ACP (PCA, Karhunen-Loeve o K-L Method)
    - Plano principal
    - Círculo de correlaciones
    - Dualidad y sobre-posición de gráficos

- Análisis Factorial de Correspondencias Múltiples

b. Aplicaciones en casos reales con el paquete **discove*R***

4. Clustering Jerárquico Aglomerativo

a. ¿Qué es “cluster analysis”?
b. Clustering Jerárquica Aglomerativa
    - Distancias y matrices de distancias
    - Agregaciones
    - Jerarquías binarias
    - Jerarquías Binarias sobre las Componentes Principales
c. Aplicaciones en casos reales con **discove*R***

5. Método de k-medias (k-means)

a. Inercia total, inercia inter-clases e inercia intra-clases
b. Teorema de Fisher
c. Problema combinatorio
d. Método de Forgy
e. Método de las nubes dinámicas
f. Aplicaciones en casos reales con R y el paquete **discove*R***

Bibliografía:

1. Berry M. and Linoff G. “Data Mining Techniques”. John Wiley & Sonsa, 1997.
2. Bry X. “Analyses factorielles simples”, Ed. Economica, Paris, 1995.
3. Hastie, Tibshirani and Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, 2009.
4. John M. Chambers. Programming with R: Software for Data Analysis. Springer, Stanford University, Palo Alto, 2008.
5. Giudici Paolo. “Applied Data Mining: Statistical Methods for Business and Industry”. Wiley, 2005.
6. Graham Williams, Data Mining with Rattle and R. Springer, New York, 2011.
7. Dunhan M. “Data mining: Introductory and Advanced Topics”. Prentice Hall, 2002.

8. Han J. and Kamber M. “Data Mining: concepts and techniques”, Morgan Kaufman Publishers 2001.
9. Jambu M. “Introduccion au Data Mining: Analyse Intelligente des données”. Eyrolles, Paris, 1999.
10. Mirkin Boris. “Clustering for Data Mining, a data recovery approach”. Chapman & Hall. Boca Raton FL, 2005.
11. Owen Jones, Robert Maillardet and Andrew Robinson. Introduction to Scientific Programming and Simulation using R. Chapman & Hall/CRC Taylor & Francis Group, FL. 2009.
12. 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/
13. R Development Core Team. “Writing R Extensions”. The R Project for Statistical Computing, 2010. http://www.r-project.org/
14. Rodríguez O, P.J.F. Groenen S. Winsberg and E. Diday. "I-Scal: Symbolic Multidimensional Scaling of Interval Dissimilarities". COMPUTATIONAL STATISTICS & DATA ANALYSIS the Official Journal of the International Association for Statistical Computing, London, 2006.
15. W. J. Braun and D. J. Murdoch, A First Course in Statistical Programming with R. The University of Western Ontario, 2007.