PRØMiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

"Big Data Analysis"

(Métodos especiales para Datos Masivos)

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

Tutor: El curso será impartido por M.Sc. Yersinio Jiménez quien cuenta con una especialización en Cloud Computing y Big Data por parte de Infosys, en Mysore India. Además tiene una maestría con honores en Ciencias de la Computación y su bachillerato en Informática Empresarial, ambos en la Universidad de Costa Rica, donde también trabaja como profesor.

Duración: Cuatro semanas.

Descripción:

En este curso se presentarán técnicas y modelos especiales para la manipulación y la aplicación de la Minería de Datos en bases de datos gigantes, para esto se hará uso de paquetes especialmente diseñados en **R** para el manejo de este tipo de bases de datos.

Además se hará uso de modelos del Análisis de Datos Simbólicos debido a que es una herramienta muy poderosa para poder resumir las bases de datos lo cual permite ejecutar modelos descriptivos y predictivos en este tipo datos.

Objetivos:

En este curso el estudiante será capaz de:

1. Comprender la necesidad de usar paquetes especializados de **R** para procesar grandes bases de datos.
2. Aprovechar las ventajas del computador para trabajar con una matemática más experimental en grandes volúmenes de datos y lograr así una mejor aproximación a lo concreto en matemática.
3. Estudiar los fundamentos teóricos de los métodos factoriales y de la clasificación simbólica.
4. Reconocer en el análisis de datos simbólico una herramienta que con frecuencia se utilizará en la Minería de Datos aplicada.

5. Entender la necesidad de la utilización de modelos simbólicos para el descubrimiento de conocimiento en grandes volúmenes de datos.
6. Conocer los principales modelos, técnicas y algoritmos simbólicos utilizados para descubrir el conocimiento en grandes volúmenes de datos.

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 el Aula Virtual del curso, 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 utilizando datos masivos, es decir, con amplio volumen, velocidad y variadad. Además podrá desarrollar proyecto de Minería de Datos utilizado computación distribuida.

Contenido:

1. ¿Qué es "Big Data"?

a. Historia de Big Data.
b. Los datos (la vida) en la nube: Big data y cloudcomputing.
c. Volumen, Variedad y Velocidad (las3V's).
d. Big Data - Big Analytics.
e. Plataforma de código abierto "Hadoop"

2. Big Data en R

a. Paqutes biglm, party, ff, bigmemory, bigtabulate, snow.
b. R Hadoop.
c. Computación paralela.
d. Paralelizando el proceso de calibración de las K-Medias.

e. Paralelizando los procesos de Validación Cruzada y selección de modelos predictivos

3. Introducción a los datos simbólicos

a. Uso de los paquetes de R: RSDA,
b. Tablas simbólicas.
c. Definición de objeto simbólico.
d. De las bases de datos relacionales a los datos simbólicos.
e. Estadísticas básicas sobre datos simbólicos, Media, varianza, mediana, Covarianza, correlación, entre otros.
f. Métodos de regresión para datos de tipo intervalo
g. Análisis en Componentes Principales para datos de tipo intervalo, Método de las esquinas y Método de los centros.

Bibliografía:

1. Billard, L. and Diday E. *Symbolic data analysis: Conceptual statistics and data mining.* Wiley, New York, 2006.
2. Bock H-H. and Diday E. (eds.) *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg, 425 pages, ISBN 3-540-66619-2, 2000.
3. Cazes P., Chouakria A., Diday E. et Schektman Y. *Extension de l'analyse en composantes principales à des données de type intervalle.* Rev. Statistique Appliquée, Vol. XLV Num. 3., pag. 5-24, Francia, 1997.
4. Chouakria A. *Extension des méthodes d'analyse factorielle à des données de type intervalle.* Thése de doctorat, Université Paris IX Dauphine.
5. Diday E. and Rodríguez, O. (eds.) "*Workshop on Symbolic Data Analysis*". PKDD–Lyon-France, 2000.
6. Groenen P.J.F., Rodríguez O., Winsberg S. and Diday E. IScal: *Symbolic Multidimensional Scaling of Interval Dissimilarities*. In COMPUTATIONAL STATISTICS & DATA ANALYSIS the Official Journal of the International Association for Statistical Computing, Vol. 51, Nov. 2006.
7. Meneses E. and Rodríguez O. *Using symbolic objects to cluster web documents*. 15th World Wide Web Conference, 2006.

8. Rodríguez O. *Classification et Modèles Linéaires en Analyse des Données Symboliques.* Thèse de doctorat, Université Paris IX Dauphine, France, 2000.
9. Rodríguez O. *The Knowledge Mining Suite (KMS).* Publicado en ECML/PKDD 2004 The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Pisa Italia, 2004.
10. Rodríguez O., Diday E. and Winsberg S. *Generalization of the Principal Components Analysis to Histogram Data.* PKDD2000, Lyon-France, 2000.
11. Rodríguez O., Castillo W., Diday E. and González J. *Correspondence Factorial Análisis for Symbolic Multi–Valued Variables.* Subjected for publication in Journal of Symbolic Data Analysis, 2003.
12. Rodríguez O. and Pacheco A. *Applications of Histogram Principal Components Analysis.* Publicado en ECML/ PKDD 2004 The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Pisa Italia, 2004.