PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

CD100

Métodos Exploratorios en Ciencia de Datos

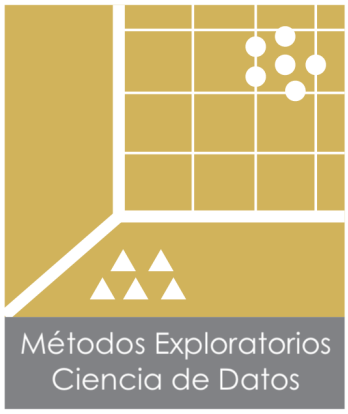
(506) 4030.1205 - (506) 4030.1114 info@promidat.com facebook.com/promidat Twitter.com/promidat www.promidat.com

Índice

Duración

Cuatro semanas.

Descripción



En este curso se presentarán los principales conceptos y métodos en Ciencia 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 Ciencia de Datos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán los programas especializados en Ciencia de Datos como discoveR 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 Ciencia 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", Ciencia 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 discoveR 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:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.

3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Ciencia de Datos que involucren segmetación de carteras de clientes.

Contenido

1. **Conceptos de la Ciencia de Datos.**
    a. Definiciones básicas en Ciencia de Datos.
    b. Instalación de la Plataforma R y RStudio.
    c. Instalación del paquete discoveR.
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 discoveR sobre la plataforma R.
3. **Métodos de Reducción de la Dimensión.**
    a. Análisis en Componentes Principales – ACP (PCA, Karhunen-Loeve o K-L Method).
    b. Plano principal.
    c. Círculo de correlaciones.
    d. Dualidad y sobre-posición de gráficos.
    e. Análisis Factorial de Correspondencias Múltiples.
    f. Aplicaciones en casos reales con el paquete discoveR
4. **Clustering Jerárquico Aglomerativo.**
    a. ¿Qué es “cluster analysis”?
    b. Clustering Jerárquica Aglomerativa.
    c. Distancias y matrices de distancias.
    d. Agregaciones.
    e. Jerarquías binarias.

f. Jerarquías Binarias sobre las Componentes Principales.

g. Aplicaciones en casos reales con discoveR.

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 discoveR.

Evaluación

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

Bibliografía

[1] *Analyses factorielles simples - Xavier Bry - Librairie Eyrolles*. URL: https://www.eyrolles.com/Sciences/Livre/analyses-factorielles-simples-9782717828597/ (visited on 10/17/2022).

[2] W. John Braun and Duncan J. Murdoch. *A First Course in Statistical Programming with R*. Google-Books-ID: NzorEAAAQBAJ. Cambridge University Press, May 20, 2021. 281 pp. ISBN: 978-1-108-99514-6.

[3] *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd Edition | Wiley*. Wiley.com. URL: https://www.wiley.com/en-us/Data+Mining+Techniques%3A+For+Marketing%2C+Sales%2C+and+Customer+Relationship+Management%2C+3rd+Edition-p-9781118087459 (visited on 10/17/2022).

[4] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer, 2001. ISBN: 978-1-4899-0519-2 978-0-387-21606-5. DOI: 10.1007/978-0-387-21606-5. URL: http://link.springer.com/10.1007/978-0-387-21606-5 (visited on 10/17/2022).

[5] Michel Jambu. "Introduction au data mining". In: (1999).

[6] Boris Mirkin and Boris Mirkin. *Clustering for Data Mining: A Data Recovery Approach*. New York: Chapman and Hall/CRC, Apr. 28, 2005. 296 pp. ISBN: 978-0-429-13754-9. DOI: 10.1201/9781420034912.

[7] *R: The R Project for Statistical Computing*. URL: https://www.r-project.org/.

[8] O Rodríguez, PJF Groenen S Winsberg, and E Diday. "I-Scal: Symbolic Multidimensional Scaling of Interval Dissimilarities". In: *COMPUTATIONAL STATISTICS & DATA ANALYSIS the Official Journal of the International Association for Statistical Computing, London* (2006).