PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

CD103

Métodos Predictivos en Ciencia de Datos

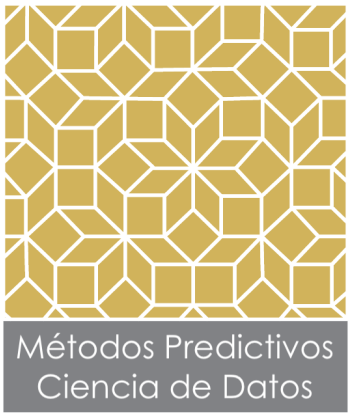
(506) 4030.1205 - (506) 4030.1114 info@promidat.com facebook.com/promidat Twitter.com/promidat www.promidat.com

Índice

Duración

Cuatro semanas.

Descripción



En este curso se presentarán los principales métodos en Ciencia de Datos, especialmente enfocados en métodos predictivos, conocidos también como métodos de aprendizaje supervisado. El énfasis principal del curso será examinar dichos métodos desde un punto de vista algorítmico y de sus aplicaciones en casos reales. 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 son la plataforma de desarrollo R y el paquete predictoR.

Objetivos

En este curso el estudiante será capaz de:

1. Comprender la diferencia entre modelos de aprendizaje supervisado (minería predictiva) y modelos de aprendizaje no supervisado (minería descriptiva).
2. Comprender la diferencia entre bases de datos de aprendizaje y bases de datos de "testing".
3. Comprender la necesidad de la utilización de modelos, algoritmos, software para predecir el comportamiento futuro.
4. Conocer los principales modelos predictivos, técnicas y algoritmos utilizados para predecir conductas a partir de grandes volúmenes de datos históricos.
5. Utilizar la plataforma R y el paquete predictoR para analizar y desarrollar 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 predicción utilizando modelos predictivos.

Contenido

1. **Conceptos de la Analítica Predictiva.**
    a. Conceptos y diferencias entre aprendizaje supervisado y aprendizaje no supervisado.
    b. Diseño de bases de datos de aprendizaje.
    c. Diseño de bases de datos de testing.
    d. Variables cuantitativas y variables cualitativas.
    e. ¿Cómo evaluar la calidad de un modelo predictivos?
    f. Cálculo de la Matriz de confusión e índices de calidad.
    g. Curvas ROC.
    h. Aplicación con datos reales con predictoR.
2. **Método de los K vecinos más cercanos.**
    a. Estructura General del método.
    b. El mejor valor de K.
    c. Algoritmo de Aprendizaje.
    d. Aplicación con datos reales con predictoR.
3. **Máquinas Vectoriales de Soporte.**
    a. Hiperplano de separación de las clases.
    b. Vectores de soporte.
    c. Función discriminante lineal.
    d. ¿Cómo resolver un Problema Optimización?
    e. MVS no linealmente separables.
    f. Núcleos en Máquinas Vectoriales de Soporte.
    g. Aplicación con datos reales con predictoR.
4. **Árboles de Decisión (Método CART).**

a. Algoritmos ID3, C4.5, C5.0 y CART.

b. Árboles de auto-regresión.

c. Aplicación con datos reales con predictoR.

5. **Métodos de consenso y de Potenciación.**

a. Métodos de Consenso (Bagging).

b. Bosques Aleatorios (Random forests).

c. Métodos de impulso (Boosting).

d. Métodos de Potenciación (ADA Boosting).

e. Aplicación con datos reales con predictoR.

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] Esteban Alfaro, Matias Gamez, and Noelia García. "adabag: An R Package for Classification with Boosting and Bagging". In: *Journal of Statistical Software* 54 (Sept. 3, 2013), pp. 1–35. ISSN: 1548-7660. DOI: 10.18637/jss.v054.i02. URL: https://doi.org/10.18637/jss.v054.i02 (visited on 10/19/2022).

[2] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* 25.2 (June 1, 2016), pp. 197–227. ISSN: 1863-8260. DOI: 10.1007/s11749-016-0481-7. URL: https://doi.org/10.1007/s11749-016-0481-7 (visited on 10/19/2022).

[3] 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 10/19/2022).

[4] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. "Support Vector Machines in R". In: *Journal of Statistical Software* 15 (Apr. 6, 2006), pp. 1–28. ISSN: 1548-7660. DOI: 10.18637/jss.v015.i09. URL: https://doi.org/10.18637/jss.v015.i09.

[5] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: 2 (2002), p. 5.

[6] Wei-Yin Loh. "Classification and regression trees". In: *WIREs Data Mining and Knowledge Discovery* 1.1 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8, pp. 14–23. ISSN: 1942-4795. DOI: 10.1002/widm.8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.8.

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