PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

CD400

Implementación de Proyectos 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 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

En este curso el estudiante será capaz de:

1. Implementar programas en R para Validación Cruzada (Cross-Validation) y Remuestreo (bootsrapping) tanto en Clasificación como en Regresión.
2. Aplicar adecuadamente una Validación Cruzada y un Remuestreo.
3. Calibrar adecuadamente tanto métodos exploratorios como métodos predictivos en Ciencia de Datos.
4. Seleccionar el mejor modelo predictivo dado un conjunto de datos.
5. Implementar proyectos avanzados en Ciencia de Datos usando R.

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 Minería de Datos que involucren métodos exploratorios y predictivos avanzados usando código en R.

Contenido

1. **Implementación de métodos Métodos Exploratorios (Clustering o Aprendizaje no Supervisado).**
    a. Análisis en Componentes Principales.
    b. Clustering Jerárquico.
    c. El método de K-medias.
2. **Implementación de Métodos de Predictivos y Métodos de Regresión.**
    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 todos los métodos anteriores para Regresión.
    k. Implementación de la Regresión Clásica.
3. **Implementación de la Validación Cruzada (cross-validation) para regresión y para clasificación, Curvas ROC.**
    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. Versiones en paralelo.
    e. Implementación de la Calibración y Selección de Modelos.
    f. Implementación de las Curvas ROC.
4. **Implementación de Métodos en Series de Tiempo.**

a. Trabajando con fechas.

b. Implementación de los métodos de Holt-Winters.

c. Implementación de los Métodos ARIMA.

d. Evaluación del error.

e. Seleccionando el mejor método.

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] 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.

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

[4] 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.

[5] 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.

[6] Simon Sheather. *A Modern Approach to Regression with R*. Google-Books-ID: IZ6Ra0efSdsC. Springer Science & Business Media, Feb. 27, 2009. 398 pp. ISBN: 978-0-387-09608-7.

[7] *The 'K' in K-fold Cross Validation*. URL: https://arpi.unipi.it/handle/11568/962587 (visited on 10/19/2022).

[8] Tzu-Tsung Wong. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation". In: *Pattern Recognition* 48.9 (Sept. 1, 2015), pp. 2839–2846. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2015.03.009. URL: https://www.sciencedirect.com/science/article/pii/S0031320315000989 (visited on 10/19/2022).

[9] Zhongheng Zhang. "Naïve Bayes classification in R". In: *Annals of Translational Medicine* 4.12 (June 2016), p. 241. ISSN: 2305-5839. DOI: 10.21037/atm.2016.03.38. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930525/ (visited on 10/19/2022).