



CD106

Métodos de Regresión en Ciencia de Datos

Descripción

En este curso se presentarán los principales métodos en Ciencia de Datos, especialmente enfocados en métodos de Regresión. 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 regressoR.

Objetivos

En este curso el estudiante será capaz de:

1. Comprender la diferencia entre modelos de clasificación y modelos de regresión.
2. Comprender la diferencia y necesidad de usar 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 de regresión, técnicas y algoritmos utilizados para predecir conductas a partir de grandes volúmenes de datos históricos.
5. 5. Utilizar la plataforma R y el paquete regressoR 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 video 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 la predicción de una variable numérica utilizando modelos de Regresión.

Contenido

1. Conceptos Básicos de Regresión.

- a. Conceptos y diferencias entre clasificación y regresión
- b. Diseño de bases de datos de aprendizaje.
- c. Diseño de bases de datos de testing.
- d. Manipulación de variables cuantitativas y variables cualitativas.
- e. ¿Cómo evaluar la calidad de un modelo un modelo de regresión?
- f. Índices de error. Error cuadrático medio, error absoluto medio, correlación y otros.

2. Regresión Lineal Clásica.

- a. Regresión Lineal simple y múltiple.
- b. Interpretación de coeficientes.
- c. Aplicaciones en casos reales con el paquete regressorR.

3. Método de Regresión Lineal Penalizada.

- a. Regresión LASSO.
- b. Regresión RIDGE.
- c. Regresión combinando métodos de reducción de la dimensión ACP y MCP.
- d. Aplicaciones en casos reales con el paquete regressorR.

4. Árboles de Decisión para Regresión – Método CART.

- a. Árboles de Regresión.
- b. Bosques de Regresión.
- c. Boosting para Regresión.
- d. Aplicaciones en casos reales con el paquete regressoR.

5. Máquinas de Soporte Vectorial para Regresión.

- a. Máquinas de Soporte Vectorial en problemas regresión.
- b. Truco del Kernel.
- c. Aplicaciones en casos reales con el paquete regressoR.

6. Deep Learning en Problemas de Regresión.

- a. Redes Neuronales y su estructura. Neuronas, capas, funciones de activación.
- b. Redes Neuronales para problemas predictivos.
- c. Redes Neuronales y Deep Learning para problemas de regresión.
- d. Aplicaciones en casos reales con el paquete regressoR.

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] 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).
- [2] Brad Boehmke and Brandon M. Greenwell. Hands-On Machine Learning with R. New York: Chapman and Hall/CRC, Nov. 14, 2019. 488 pp. isbn: 978-0-367-81637-7. doi:10.1201/9780367816377.
- [3] John Fox and Sanford Weisberg. An R Companion to Applied Regression. Google-Books-ID: uPNrDwAAQBAJ. SAGE Publications, Sept. 27, 2018. 608 pp. isbn: 978-1-5443-3648-0.
- [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 10/19/2022).
- [5] 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>.
- [6] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: 2 (2002), p. 5.
- [7] 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>.
- [8] R: The R Project for Statistical Computing. url: <https://www.r-project.org/>.
- [9] 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.