



CD200

Métodos Avanzados en Ciencia de Datos

Descripción

En este curso se estudiarán en detalle las técnicas de Validación Cruzada (Cross-Validation) y Remuestreo (bootstrapping) con el objetivo de calibrar y seleccionar el mejor método de Ciencia de Datos para un problema y juego de datos dado. Se estudiará como programar y automatizar la Validación Cruzada (Cross-Validation) y Remuestreo (bootstrapping) 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. Usar el paquete predictoR para Validación Cruzada (Cross-Validation) y Remuestreo (bootstrapping).
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. Instalar y utilizar paquetes avanzados en 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 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 calibración y selección de métodos exploratorios y predictivos en código R.

Contenido

1. **Métodos Bayesianos**
 - a. Método de Bayes.
 - b. Análisis Discriminante lineal y cuadrático.
2. **Probabilidad de Corte y Curva ROC.**
 - a. Probabilidad de corte.
 - b. Curva ROC.
3. **Validación Cruzada (cross-validation).**
 - 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).
4. **Calibración y Selección de Métodos.**
 - a. Calibración de Métodos Exploratorios (descriptivos).
 - b. Calibración de Métodos Predictivos.
 - c. Seleccionando el mejor método predictivo.

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] Hervé Abdi and Dominique Valentin. "Multiple Correspondence Analysis". In: Multiple Correspondence Analysis (), p. 13.
- [2] Michael Greenacre and Jorg Blasius, eds. Multiple Correspondence Analysis and Related Methods. New York: Chapman and Hall/CRC, June 22, 2006. 608 pp. isbn: 978-0-429-14196-6. doi: 10.1201/9781420011319.
- [3] H. Teil. "Correspondence factor analysis: An outline of its method". In: Journal of the International Association for Mathematical Geology 7.1 (Feb. 1, 1975), pp. 3-12. issn:1573-8868. doi: 10.1007/BF02080630. url: <https://doi.org/10.1007/BF02080630> (visited on 10/19/2022).
- [4] The 'K' in K-fold Cross Validation. url: <https://arpi.unipi.it/handle/11568/962587> (visited on 10/19/2022).
- [5] R. Vidal, Yi Ma, and S. Sastry. "Generalized principal component analysis (GPCA)". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 27.12 (Dec. 2005). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1945-1959. issn: 1939-3539. doi: 10.1109/TPAMI.2005.244.
- [6] 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).