Presentación de **Minería de datos**

Máster de Bioestadística, Universitat de València. Curso 2021-2022

Paloma Botella Rocamora
Paloma.Botella@gmail.com

Temario/Sesiones/Evaluación.

Introducción a la Minería de datos.

Bibliografía/Recursos

Temario/Sesiones/Evaluación.

Temario/Sesiones

- 16/12/2021 (J) : Análisis exploratorio multivariante (teoría y ejercicio preparatorio banco de datos).
  - 21/12/2021* (M) : *Prácticas [trabajo individual - presencia opcional]*.
- 23/12/2021* (J) : Análisis de componentes principales (teoría y ejercicio práctico).
  - 11/01/2022 (M) : *Prácticas [trabajo individual - presencia opcional]*.
- 18/01/2022 (M) : Análisis cluster (teoría y ejercicio práctico).
  - 25/01/2022 (M) : *Prácticas [trabajo individual - presencia opcional]*.
- 1/02/2022 (M): Escalado multidimensional + Análisis discriminante (teoría y ejercicio practico).
- 8/02/2022 (J): Examen final.

* ¿Intercambiamos contenido clases 21/12 (presencial) y 23/12 (online)?

Evaluación

- Práctica 1: *Análisis exploratorio multivariante.* 10%
  - *Fecha máxima entrega: Domingo 16/01/2022*
- Práctica 2: *Análisis de Componentes Principales.* 10%
  - *Fecha máxima entrega: Domingo 23/01/2022*
- Práctica 3: *Análisis de agrupamiento.* 10%
  - *Fecha máxima entrega: Domingo 30/01/2022*
- Práctica 4: *Escalado multidimensional+Análisis discrim.* 10%
  - *Fecha máxima entrega: Domingo 13/02/2022*
- Examen final: *Prueba de evaluación final.*60%

Entrega de las prácticas

- Para cada práctica propuesta el alumno deberá entregar una memoria que contenga tanto el **código R** empleado para resolver la práctica como **los resultados** y **la interpretación** que el alumno hace de los mismos.
- El formato *preferido* para este documento es un pdf generado, a ser posible, a partir de un documento *R Markdown*. Esta posibilidad que ofrece RStudio, además de ser una herramienta muy potente que interesa conocer, permite combinar de forma muy sencilla en un documento el código y los resultados obtenidos en la ejecución del mismo.

Tutorías

El alumno podrá solicitar **tutorías** presenciales o virtuales (*aula virtual*, *skype*, *telegram*,. . .) contactando a través de correo electrónico (*Paloma.Botella@gmail.com*).

Introducción a la Minería de datos.

Minería de datos (Data Mining)

- No existe una única defínción del término **Data Mining**.

Podemos referirnos a este término como:

Conjunto de técnicas estadísticas que permiten explorar **grandes bases de datos** y que tienen como objetivo el encontrar *patrones repetitivos*, *tendencias* o *relaciones* que expliquen el comportamiento de los datos.

Minería de datos (Data Mining)

Para estudiar cualquier fenómeno real se dispone de información de **varias variables** simultáneamente (situación económica de un país, compras de una gran superficie, características médicas de una persona,...).

La **Minería de Datos** comprende el estudio estadístico de varias variables medidas en elementos de una población.

Minería de datos (Data Mining)

- Involucra gran variedad de técnicas estadísticas.
- Visión práctica: métodos aproximados
- Carácter descriptivo.
- Conocido en otro tiempo como *Análisis Multivariante*

Minería de datos: *otros nombres*

Muchas definiciones alternativas y visiones desde otros campos, como la Informática y la Inteligencia Artificial

Extracción no trivial de información que reside de manera implícita en los datos (Knowledge Discovery in Databases, KDD)

- Statistical Learning
- Inteligencia Artificial
- Sistemas Expertos
- Sistemas Inteligentes
- Redes Neuronales
- ...

Tipos de métodos

Según el **sujeto de estudio**:

- **Supervisadas**

  hay una variable que debe ser explicada por otras $>$ *Conlleva un* ***modelo***

  (pueden ser utilizadas para hacer predicciones)

- **No Supervisadas**

  no hay una variable preferente que explicar $>$ *Conlleva un* ***criterio***

  (extracción de información útil a partir de los datos)

Tipos de métodos

- **Supervisadas**

  Una variable de interés observada en todos los individuos, variable **respuesta**, quiere ser explicada mediante variables explicativas o **predictores** que también son observadas en todos los individuos.

  El objetivo es ajustar un modelo que relacione la variable respuesta con los predictores con el objetivo de *predecir futuras observaciones* (**predicción**) o *entender mejor la relación entre la variable respuesta y los predictores* (**inferencia**).

  Ejemplos: *regresión lineal, regresión logística (o glm), gam, métodos de clasificación,. . .*

Tipos de métodos

- **No Supervisadas**

  Un conjunto de variables son observadas en todos los individuos.

  Entre ellas no hay una que sea la que más nos interesa, por lo que no hay un modelo que ajustar.

  Podemos tratar de entender la **relación entre las variables** o la **relación entre los individuos**.

  Ejemplos: *clustering, análisis de componentes principales, análisis factorial,...*

Tipos de métodos

- **No Supervisadas**

  En este bloque de la asignatura la mayoría de técnicas que forman parte de los contenidos se enmarcan dentro de las "No Supervisadas".

  El *aprendizaje no supervisado* con frecuencia forma parte de un *análisis exploratorio de los datos*. La validación de los resultados obtenidos en estas técnicas no es tan sencillo como en el caso de las técnicas de análisis supervisado. Por este motivo los resultados proporcionados por estas técnicas son, en general, considerados como **exploratorios**.

Tipos de métodos

Según el **objetivo de estudio** podemos encontrar diferentes tipos de técnicas:

- Regresión
- Clasificación
- Agrupamiento
- Asociación
- Patrones

Tipos de métodos

Según el **objetivo de estudio** podemos encontrar diferentes tipos de técnicas:

- **Regresión**
  - **Explicar una variable** (normalmente cuantitativa) a partir del resto de variables explicativas medidas. *Hay una variable que explicar*.
- Clasificación
- Agrupamiento
- Asociación
- Patrones

Tipos de métodos

Según el **objetivo de estudio** podemos encontrar diferentes tipos de técnicas:

- Regresión
- **Clasificación**
  - **Asignar cada individuo de un conjunto de datos a grupos fijados** de manera que se minimice la probabilidad de una clasificación errónea. *Hay una variable que explicar, cualitativa, la que define la clasificación.*
- Agrupamiento
- Asociación
- Patrones

Tipos de métodos

Según el **objetivo de estudio** podemos encontrar diferentes tipos de técnicas:

- Regresión
- Clasificación
- **Agrupamiento**
  - El objetivo es **agrupar los elementos** de un conjunto de datos en grupos mutuamente excluyentes de tal manera que:
    - Los elementos de un mismo grupo están lo más cerca posible entre sí.
    - Los elementos de grupos diferentes están lo más lejos posible entre sí
    - La distancia está medida respecto a todas las variables disponibles

  No hay una variable a explicar.
- Asociación
- Patrones

Tipos de métodos

Según el **objetivo de estudio** podemos encontrar diferentes tipos de técnicas:

- Regresión
- Clasificación
- Agrupamiento
- **Asociación**
  - Las observaciones son usadas para identificar **asociaciones** entre variables

  Si hay asociaciones no tiene por qué haber causalidad
- Patrones

Tipos de métodos

Según el **objetivo de estudio** podemos encontrar diferentes tipos de técnicas:

- Regresión
- Clasificación
- Agrupamiento
- Asociación
- **Patrones**
  - Identificar **tendencias** y **patrones** de comportamiento (puede contener alguno de los puntos anteriores, identificación datos anómalos,...).

Bibliografía/Recursos

Bibliografía

- J.Aldas, E.Uriel (2017) Análisis multivariante aplicado con R (2ª edición). Paraninfo.
- D. Peña (2002) Análisis de datos multivariantes. McGraw-Hill
- G.James, D.Witten, T.Hastie and R.Tibshirani (2013) An Introduction to Statistical Learning (with application in R). Springer. (disponible en www.StatLearning.com)
- T. Hastie, R. Tibshirani and J. Friedman (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer

Recursos

- Data Visualization with R: https://rkabacoff.github.io/datavis/
- CRAN task view: Machine Learning & Statistical Learning http://cran.r-project.org/web/views/MachineLearning.html
- CRAN task view: Multivariate Statistics http://cran.r-project.org/web/views/Multivariate.html
- R and Data Mining: http://www.rdatamining.com/docs/introduction-to-data-mining-with-r