Statistical Learning: Evaluation

Javier Nogales
Full Professor, Department of Statistics, UC3M
fcojavier.nogales@uc3m.es

MS in Statistics for Data Science

Deadline 1: Feb 28, 2024 at 14:00 (midterm project)
Deadline 2: Mar 18, 2023 at 14:00 (final project)

Final Project (including midterm)

- Apply classification tools to two real datasets, one for each project
- The project is divided in two parts (midterm plus final):
  - The objective of the first part is to explain the main predictors affecting the output (using the complete data set): statistical learning tools
  - The objective of the second part is to predict the output (using different training/testing sets): machine learning tools
- Choose an application of your interest: engineering, finance and banks, insurance companies, health care, marketing, research centers, etc.
- Use the internet to make your own survey and data collection (open data, APIs, etc.)
- The larger the dataset (in terms of rows or columns) the better
- It is mandatory to draw some conclusions for each of the two parts

Final Project: Statistical Learning

The two parts:

- Part 1: statistical tools
  - Prepare the input (Download data, pre-process, EDA, etc.): 2 points
  - Classification modeling using statistical tools (emphasis on probabilities): 3 points

    Upload to Aula Global before deadline: a notebook (.rmd and .html) and the dataset

- Part 2: machine-learning tools
  - Prepare the input (Download data, pre-process, EDA, etc.): 2 points
  - Classification modeling using machine-learning tools (emphasis on performance): 3 points

    Upload to Aula Global before deadline: a notebook (.rmd and .html) and the dataset

Open Data

Some interesting links to get data (but you can use any other):

http://datos.madrid.es/
http://datos.gob.es/
http://open-data.europa.eu/es/data/
http://data.gov/
http://quandl.com/
http://datacatalog.worldbank.org/
https://research.stlouisfed.org/fred2/
https://archive.ics.uci.edu/ml/index.html
http://www.statsci.org/datasets.html
http://lib.stat.cmu.edu/DASL
http://www.umass.edu/statdata/statdata/
http://www.philender.com/courses/multivariate/data.html
http://biostatistics.iop.kcl.ac.uk/publications/everitt/
http://www.oecd.org/statistics/

Open Data

Alternatively, you can use R-libraries to connect open data:

- World Health Organization https://www.who.int/gho/en/
  using the rgho R-library
- Air Quality https://openaq.org/ using the ropenaq R-library
- World Bank https://data.worldbank.org
  using the WDI or wbstats R-library
- Organization for Economic Cooperation and Development https://data.oecd.org using the OECD R-library
- Financial and economic data using the quantmod R-library or the Quandl
- Weather data using the NOAA R-library
- Etc.