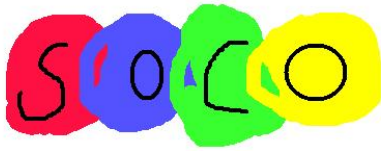
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development



Soft Computing Research Group

Practical work

Lluís A. Belanche
Computer Science Department
belanche@cs.upc.edu

Computación y Sistemas Inteligentes (MEI - UPC)

December 4, 2017

Project 1: the Bank Marketing dataset

Direct marketing campaign (by means of phone calls) from a Portuguese banking institution

- Often, more than one contact to the same client was required, in order to assess if the product would be subscribed
- Number of examples: 45,211; 16 predictors, of very different nature and type, including factors, '999' and 'unknown'
- The **target** variable is whether a term deposit was subscribed ('yes') or not ('no')

More information can be found in https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

The **input** (predictive) variables are:

1. bank client data:
   1. age (numeric)
   2. job : type of job
      ("admin.","unknown","unemployed","management","housemaid","entrepreneur",
      "student", "blue-collar","self-employed","retired","technician","services")
   3. marital : marital status ("married","divorced","single")
   4. education ("unknown","secondary","primary","tertiary")
   5. default: has credit in default? ("yes","no")
   6. balance: average yearly balance, in euros (numeric)
   7. housing: has housing loan? ("yes","no")
   8. loan: has personal loan? ("yes","no")

2. related with the last contact of the current campaign:
   9. contact: contact communication type ("unknown","telephone","cellular")
   10. day: last contact day of the month (numeric)
   11. month: last contact month of year ("jan", "feb", "mar", ..., "nov", "dec")
   12. duration: last contact duration, in seconds (numeric)

3. other variables:
   13. campaign: number of contacts performed during this campaign and for this client
       (numeric, includes last contact)
   14. pdays: number of days passed after the client was last contacted from a previous campaign
       (numeric, -1 means client was not previously contacted)
   15. previous: number of contacts performed before this campaign and for this client (numeric)
   16. poutcome: outcome of the previous marketing campaign ("unknown","other","failure","success")

Project 2: Yacht Hydrodynamics Data Set

The task is to predict the hydrodynamic performance of sailing yachts from basic hull dimensions and boat velocity

- Prediction of residuary resistance of sailing yachts at the initial design stage is of a great value for evaluating the ship's performance and for estimating the required propulsive power
- The data set comprises 308 full-scale experiments, which were performed at the Delft Ship Hydromechanics Laboratory
- The **input** (predictive) variables are:
  1. Longitudinal position of the center of buoyancy, adimensional.
  2. Prismatic coefficient, adimensional.
  3. Length-displacement ratio, adimensional.
  4. Beam-draught ratio, adimensional.
  5. Length-beam ratio, adimensional.
  6. Froude number, adimensional.
- The **target** variable is the residuary resistance per unit weight of displacement

Project 2: Yacht Hydrodynamics Data Set

Some issues to consider when pre-processing:

1. Should we take the log of the target?
2. Should we scale and center the data?
3. Should we consider using ordered variables?
4. Compare against ridge regression

More information can be found in https://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics