Proposal of Machine Learning Project

Experimental comparison of ML algorithms

1. *Group's member*:

- Nguyễn Hoàng Anh – 20194417 – anh.nh194417@sis.hust.edu.vn
- Nguyễn Minh Châu – 20194420 – chau.nm194420@sis.hust.edu.vn
- Chu Hoàng Dương – 20194429 -duong.ch194429@sis.hust.edu.vn

2. *Description of the problem*:

The problem is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

- Input. The representation of customer's information (e.g., a vector of attribute values).
- Output. Predict if the client will subscribe (yes/no) a term deposit (variable y).

3. *Machine learning algorithms*: Nearest neighbor learning, Decision tree classification
4. *Dataset*: Bank Marketing Dataset

❖ General description:

- Data Set Characteristics: Multivariate
- Number of Instances: 45211 for bank-full.csv (4521 for bank.csv)
- Area: Business
- Attribute Characteristics: Real
- Number of Attributes: 16 + output attribute

❖ Missing Attribute Values: None

❖ Source: The full dataset (bank-full.csv) was described and analyzed in:
UCI Machine Learning Repository: Bank Marketing Data Set

❖ The zip file includes two datasets:

1) bank-full.csv with all examples, ordered by date (from May 2008 to November 2010).

2) bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv.

The smallest dataset is provided to test more computationally demanding machine learning algorithms.

- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).
- Attribute information:
  - Input variables: # bank client data:

1 - age (numeric)

2 - job: type of job (categorical:"admin.","unknown","unemployed","management","housemaid","entrepreneur", "student", "blue-collar","self-employed","retired","technician","services")

3 - marital: marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by 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 (categorical:"unknown","other","failure","success")

- Output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: "yes","no")