Team members:
- Martí Cardoso i Sabé
- Meysam Zamani Forooshani

ML - Machine Learning
9th of April, 2019

Project proposal

In this document, we are going to explain the problem that we want to tackle for the ML course project and the motivations that made us choose this dataset.

Problem description

First of all, the dataset that we have chosen is the **Bank Marketing Data Set** [7]. It can be found in the *UCI Machine Learning Repository*, and it is available online on the following URL:

`https://archive.ics.uci.edu/ml/datasets/bank+marketing`

We have data related with direct marketing campaigns of a Portuguese banking institution, that were based on phone calls. The main goal of the problem is to predict if the client will subscribe a term deposit or not, so it is a binary **classification** problem.

The Bank Marketing dataset has **20 explanatory** variables (10 of which are continuous, 7 categorical and 3 binary) and the binary response variable. These explanatory variables are:

1. Age (cont.)
2. Type of job (cat.)
3. Marital status (cat.)
4. education (cat.)
5. Has credit in default (binary)
6. Has housing loan? (binary)
7. Has personal loan? (binary)
8. Contact communication type (cat.)
9. Last contact month of year (cat.)
10. Last contact day of the week (cat.)
11. Last contact duration, in seconds (cont.)
12. Number of contacts performed during this campaign and for this client (cont.)
13. Number of days that passed by after the client was last contacted from a previous campaign (cont.)
14. Number of contacts performed before this campaign and for this client (cont.)
15. Outcome of the previous marketing campaign (cat.)
16. Employment variation rate (cont.)
17. Consumer price index (cont.)
18. Consumer confidence index (cont.)
19. Euribor 3 month rate (cont.)
20. Number of employees (cont.)

We have **41.188 observations** and there are **no missing** values (although some of the categorical variables have, it has been created an *unknown* modality for them).

Motivations

We have chosen this dataset because it is an interesting classification problem related to business. The problem is not an easy one neither a very complex, so we think that with it, we will apply some of the machine learning techniques explained in class and learn how to develop a ML

project. Also, the dataset comes from the UCI repository, so it is a well-studied problem and it is useful for academic purposes.

The dataset has enough explanatory variables (20) and the number of observations is not low for this type of problem (more than 40.000), so we think that we will not have any problem with this dataset.

Fundamental references

As said before, this dataset is a well-studied one and many papers have used it.

For example, [2] compares multilayer perception neural network (MLPNN), tree augmented Naïve Bayes (TAN), logistic regression (LR), and Ross Quinlan new decision tree model (C5.0) in order to predict whether a client will subscribe a term deposit, it obtains testing accuracies near to 90%. [3] makes a comparison between J48-graft and LAD decision trees, radial basis function network and support vector machine, the best accuracy is obtained with SVM (87%). [4] uses the Naïve Bayes and the C4.5 decision tree algorithms and gets 85% and 93% of accuracy respectively.

As well, we can mention to papers referenced in the UCI Repository ([7]): [5] performs a semi-automatic feature selection and a comparison between several models (logistic regression, decision trees, neural network and support vector machine), and the classification accuracy achieved is 81%. Additionally [6] uses the CRISP-DM methodology and the best model is obtained using SVM.

Also, the OpenML webpage [1] has a task associated with this dataset and our classification purposes. In it, we can see a comparison between several methods and the results that they get. The best accuracy is near 90% and it is obtained with random forest.

Preliminary title: Classification on bank marketing data

References

[1] OpenML Supervised Classification on bank-marketing. `https://www.openml.org/t/9899`. [Online; accessed 03-04-2019].

[2] Hany A Elsalamony. Bank Direct Marketing Analysis of Data Mining Techniques. Technical Report 7, 2014.

[3] K. Wisaeng. A Comparison of Different Classification Techniques for Bank Direct Marketing. *International Journal of Soft Computing and Engineering (IJSCE)*, 2013.

[4] Masud Karim and Rashedur M Rahman. Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications*, (6):196–206, 2013.

[5] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31. June 2014.

[6] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip].

[7] UCI Machine Learning Repository. Bank Marketing Data Set. `https://archive.ics.uci.edu/ml/datasets/bank+marketing`. [Online; accessed 03-04-2019].