Assignment 1
Advanced Machine Learning
University of Milano-Bicocca
M.Sc. Data Science

Alessandro Riboni
October 20, 2019
ID number: 847160
a.riboni2@campus.unimib.it

Prediction of Default Payments using a Neural Network

The aim of this assignment is to perform a binary classification problem using an Artificial Neural Network. The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. After a first phase of Data Preparation, the network was developed. It is composed by 17 input neurons (one for each independent variable), a hidden layer and an output neuron.

1 Data Exploration

The dataset was presented already divided into two parts: training and test data. The only difference between the two is that the test set does not contain the variable to predict (*default.payment.next.month*). This variable represents the default payment of a client and can assume the values of 1 or 0. The other features describe the client and his financial situation. The following are briefly presented:

- *AGE*: age of client (in years).
- *MARRIAGE*: Marital status (0=unknown, 1=married, 2=single, 3=others)
- *EDUCATION*: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- *SEX*: gender (1=male, 2=female)
- *PAY_k*: repayment status. It takes discrete values $p \in [-2, 8]$. If $p \leq 0$ than the client paid in advance and if $p \geq 0$ than the client paid late. Moreover, $k = 0, 2, 3, 4, 5, 6$ and it represents the month associated to the repayment status.
- *BILL_AMT_j*: amount of bill statement, where $j = 1, 2, 3, 4, 5, 6$ and it represents the month associated to the amount of bill statement.
- *PAY_AMT_j*: amount of previous payment, where $j = 1, 2, 3, 4, 5, 6$ and it represents the month associated to the amount of previous payment.

An important consideration must be made about the target variable. The training set contains 21027 rows with target = 0 (78%) and 5973 with target = 1 (22%). There is a class imbalance problem and it is necessary to evaluate the model with the appropriate metrics.

2 Data Preparation

After exploring the dataset, it was necessary to make some changes to some attributes. In particular, the variables *SEX* and *MARRIAGE* have been converted to boolean attributes. *BILL_AMT_j*, *PAY_AMT_k* and *AGE* have been normalized so that their distribution is mapped between 0 and 1. Finally, it was decided to not consider the dependent variables correlated with each other to avoid problems of multicollinearity.

3 Neural Network

Through an evaluation of the importance of features with *ExtraTreesClassifier()* and the analysis of plots of variable distributions with respect to the target, it was decided to use 17 features for the classification problem. The dataset was splitted into training (75%) and validation set (25%). Afterwards, the neural network was developed. Several tests have been performed to choose the number of neurons, loss function, activation function and network parameters. As the complexity of the network increased, the improvements were not evident and, consequently, it was decided to develop a network that was easier to interpret and with less computation time. The only hidden layer is composed of 16 neurons with *relu* activation function (Rectified Linear Unit). The output neuron has a *sigmoid* activation function, which is suitable for a problem of binary classification. The *sigmoid* is also called logistic function. It is an S-shaped function that maps each value of x into the interval [0,1].
The optimizer used is *Adam* (Adaptive moment estimation), which combines momentum and RMSprop (root mean squared prop). The loss function used is the *binary_crossentropy* and it was necessary to associate weights to the classes so that the model could learn better how to predict the rare class.

4 Results and Consideration

The network was validated on a small sample of data. As there was a class imbalance problem, the chosen reference metric was F-measure.
In the following figure are reported the results obtained with a batch_size = 32, epochs = 50 and class_weights ={1:2, 0:1}.

	precision	recall	f1-score	support
0.0	0.86	0.90	0.88	5240
1.0	0.58	0.51	0.54	1510
micro avg	0.81	0.81	0.81	6750
macro avg	0.72	0.70	0.71	6750
weighted avg	0.80	0.81	0.80	6750

Figure 1: *Results on validation set.*

As can be seen from the *Figure 1*, the f-measure of class 1 is 0.54, the f-measure of class 0 is 0.88 and the weighted avg is 0.80. By testing the parameters it was possible to notice that, by increasing the weight associated with the rare class, it would be possible to increase the relative f-measure.This increase involved a decrease in the exactness in classifying the largest class. In these cases it is necessary to find a trade-off according to the objective of the analysis that is wanted to carry out.