Machine Learning - Binary Classification

by
QUINTANA Gonzalo
MARRA Tales
WANG Lei
AGUDELO Santiago

Dans le cadre du cours
Nom du cours (Code du cours)

Travail présenté à
Nom du professeur

TAF Mathematical Computing Engenieering
IMT Atlantique

Tuesday 3rd December, 2019

Introduction

The following report summarizes our work on Machine Learning course project. This project consisted in binary classification of two datasets. The first dataset we took into account was the "Banknote Authentication Dataset", in which data was extracted from images that were taken for evaluating the authentication procedure for banknotes. The second dataset we considered was the "Chronic Kidney disease Dataset", which consisted in 25 features which may allow predicting a patient with chronic kidney disease.

The main goal of the project was to apply on practical datasets some of the different machine learning algorithms that were introduced during course sessions, all this by adopting a teamwork-based collaborative approach. It was also searched to raise consciousness about good programming practices, which are necessary for all developed codes to be organized, understandable and properly documented.

First, the datasets will be described as well as the pre-processing they require. Then, the binary classification problem will be solved for both datasets by using machine learning algorithms such as support vector machines, neural networks, decision trees and clustering techniques. The obtained results will be presented and discussed; this will allow evaluating the pertinence of the different algorithms in terms of their performance and complexity. In the end, a conclusion section will summarize the main points of discussion as well as the principal difficulties we encountered during the project's realization.

1 Datasets: description and pre-processing

1.1 Banknote Authentication Dataset

These data were extracted from images that were taken from genuine and forged banknote-like specimens. These images have a size of 400x400 pixels with a resolution of about 660 dpi. In addition, wavelet transform was used to extract different features from images.

The variance, skewness and curtosis of the wavelet transformed image, the entropy of the image and its class are included in the dataset. This last information sets whether the image corresponds to a genuine of a forged specimen; it is therefore our objective to develop machine learning algorithms allowing to predict the image's class from information on the dataset.

1.2 Chronic Kidney Disease Dataset

The data was taken over a 2 months period in India. It includes 25 features that allow predicting the presence of chronic kidney disease. Our goal is therefore to use the information contained in the dataset to determine through machine learning algorithms whether a patient suffers from this disease (we will call this case ckd) or not (we will call this case notckd).

1.3 Preprocessing

The datasets we worked on encountered problems such as missing values, out-of-range values, invalid characters, non-centered data, etc. Also, some of the data corresponded to text that cannot be directly processed; therefore, a conversion to a numerical format is necessary.

For pre-processing data, we developed a module that can be applied to a general dataset.This

way, we could use the same procedure to pre-process both of the datasets concerned by this project. In such module, the user must indicate the repertory where data are localized and whether the header arrow and the index column are present or not. The user is also requested if text data should be transformed into numerical data.

The following steps allow obtaining a exploitable dataset:

1. The files (.text or .csv) are read through pandas in Python.

2. Features and targets are separated.

3. For each feature and target, the strings they contain are identified. In this same step, all special characters are left behind.

4. Data are cleansed. If there happens to be a missing value in a numerical column (feature), it is replaced by the average value of the column. For modal columns (i.e, the columns that contain strings), strings are converted into a numerical format. For example, the words yes/no are converted to a binary format 1/0. If these columns happen to have missing values, they are filled with random data in a way the proportion of the values in the column doesn't change.

5. Features are normalized (the mean is subtracted and the result is scaled in order to have unit variance).

2 SVM

The main goal of this algorithm is to separate binary data through an hyperplane. Since data may not be linearly separable, the kernel trick may be used to map data into a higher dimensional space, in which it would become linearly separable.

SVM was implemented through scikit-learn library. This function solves the following optimization problem, given training vectors x_i and $y_i \in 0{,}1$:

$$\begin{aligned} \min_{w,\zeta} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \\ st \quad & y_i (w^T \varphi(x_i) + b \geq 1 - \zeta_i \\ & \zeta_i \geq 0 \end{aligned}$$

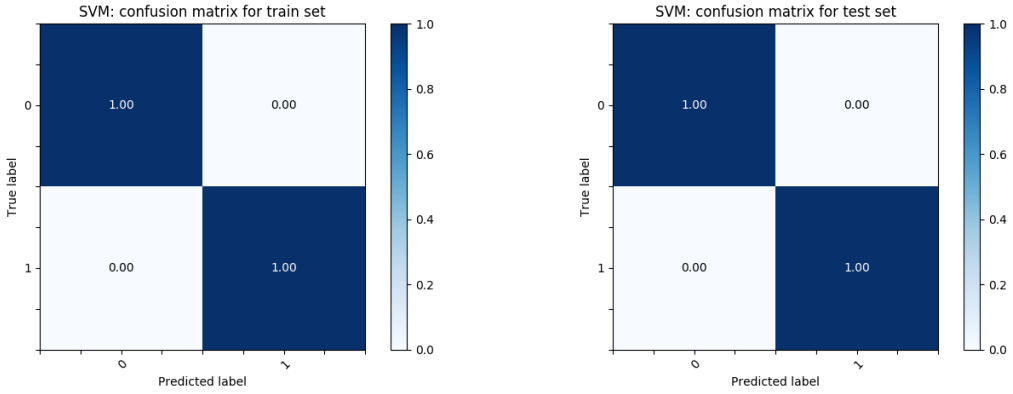
Its dual problem:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ st \quad & y^T \alpha = 0 \end{aligned}$$

$$0 \leq \alpha_i \leq C$$

e is an only ones vector and the matrix Q is defined as $Q_{ij} = y_i y_j K(x_i,x_j)$, where $K(x_i,x_j) = \varphi(x_i)^T \varphi(x_j)$ is the kernel that implicitly maps our data into a higher dimensional space.

For our implementation, we defined $K(x,x') = exp(-\gamma||x-x'||^2)$ with $\gamma = \frac{1}{N \cdot var(x)}$. The following figure shows the confusion matrices for training and test data that were obtained when running the algorithm on the Kidney Disease Dataset:



(a) Confusion matrix of training data

(b) Confusion matrix test data

Figure 1: Confusion matrices SVM