Université de St-Etienne.

Data Mining for Big Data: Project

Assignment: Due date January 31st 2019

All material is available on claroline
Project groups of at most 4 students

1 Project Objectives

The objective of this project is to achieve a large scientific analysis of a dataset by means of both data mining, machine learning and data analysis techniques (that you have learned so far in any course!).

You must take this project according to the following context: Suppose you work for an Information Technology company specialized in data science. Your boss is in touch with a financial customer company. This customer company want to take into account news provided by a news agency to help its decision making process. The goal is thus to do an automatic analysis of a news feed to detect business opportunities and to solve their prediction task (see below).

You must present your study in a report which is supposed to be delivered to your boss. The quality of the presentation, of the writing and of your scientific conclusions are very important. Another important feature for you: this report must be sufficiently well written and structured to be shown to any prospective employer in order to give an overview of your skills in data mining, machine learning and data analysis.

You are free to define the structure of your report, but you can base your work on the following guidelines:

- Provide a cover page with title, name of the contributors, logos (e.g., from the university); do not forget to provide a table of contents.
- Provide an introduction that will present your work and the structure of the report.
- Provide a section describing the dataset.
- Provide a section describing your objectives and the different studies made.
- Provide a section for each study made: you must clearly state the objective of the section, present clearly the experimental setup (preparation of the data, algorithm(s) used, any relevant information), gives the results in a neat way (table of results, plots, curves, graphs, . . . , do not forget to comment them) and give your conclusions for this section.
- Provide a final section summarizing what you have done, developing your conclusions with personal remarks and/or suggestions in order to help your boss to choose which services can be proposed to the city.

You can use anything you think relevant for doing this analysis. You must use both data mining and machine learning methods, and possible some data analysis approaches (depending on your background). Your study must then be based on different methods, on the one hand you can choose methods that can be easy to interpret (for the end-user) and on the other hand try more complex ones that may give better result. Your study must address the prediction task and at least some other problems: information / knowledge prediction, data categorization, clustering, feature analysis, ...

You are free to use any existing software, implementation, library, platforms... free of use, but you have to mention all the tools used in the report.

2 Context of the Study

News agency provide news feed to the public. These feeds consist of generally short articles which focus on facts. These feeds can be general or more specialized (sports, politics, business...). In this project, the customer company is interested in an automatic analysis of financial news and in particular filtering news related to the earnings of companies (this could be useful for their investment strategy for instance).

3 Data

The data is an extract from a news feed from the Reuters news agency. There are 5000 short news article (each in a separate file) about financial news. The articles are in an xml format (but the format is very simple, you will not need a real xml parser to read the files). For instance, this is file `7.xml`:

```
<?xml version='1.0' encoding='utf-8'?><BODY>Red Lion Inns Limited Partnership
said it filed a registration statement with the Securities and
Exchange Commission covering a proposed offering of 4,790,000
units of limited partnership interests.
    The company said it expects the offering to be priced at 20
dlrs per unit.
    It said proceeds from the offering, along with a 102.5 mln
dlr mortgage loan, will be used to finance its planned
acquisition of 10 Red Lion hotels.
 Reuter
    </BODY>
```

For the prediction task, the set of the 5000 articles have been randomly partitioned into a training set (with 4800 articles) and a test set of 200 articles. You have access only to the label for the training set and you will have to submit (with your report) your prediction for the label of the 200 test set articles.

There are two files `train.csv` and `test.csv` in csv format (comma separated values) to describe the training and test set.

The first file contains the training set of 4800 articles with their labels and you will have to fill the second file of 200 articles with the results of your predicted labels. Each of this file contains one row for each news article. Each row contains 3 values (for the training set) or 2 values (for the test set):

- The id of the article (a number in the interval 1-5000);
- the name of the xml file containing the article;
- the label of the article (0 or 1), only for the training set.

4 Prediction Task

The goal of this task is to predict the label for each article in the test set. The label is 1 if the article is related to the earnings of a company and 0 otherwise. The articles were labeled by hand by (expensive) specialists of the customer company. The goal for the customer company is to have an automatic tool be able to automatically predict the label of articles without using human specialists. To prove that your prediction tool is efficient, you will have to fill the label for the test set. Then, the customer company will compare your predicted label on this set with the correct labels.

To conduct this prediction task, you need to describe in your report the technique(s) that you used and the results you obtained (e.g., by cross validation) on the available data (the train set). You will also have to complete the file containing your predictions for the test set. Your predictions will be evaluated by comparing your results on the test set with the real labels.

5 Data Analysis

You are free to conduct a general analysis of the data using any tool / technique that you know. Your aim is to find new and interesting knowledge in the data. For instance, are there redundant variables, correlations between some of the variables, ...

6 Reproducibility of Your Results

Your report must be a scientific report. This means that every result presented in your report (in the text or in a graphic, table, etc.) can be replicated by anyone having the datasets. You must therefore describe what you have done with sufficient details such that anyone reading your documents (for instance, your professors) can do the same and get the same results as you. You must, for instance, include the code you have written, explain which software you used (with the version number), give all the parameters that you used each time you did an experiment and so on.

This information can be put in the report itself or in an additional document. You can include a log file of every command you launch as we did in the practical sessions.

Do not wait until the end to write this document! Do it each time you do an experiment, otherwise you will forget some details.

7 Assignment

Your work (zip file) must be uploaded on claroline connect for **January the 31st, 23:00**. Your work must contains:

- the report;
- the file `test.csv` with your predictions for the label (in the correct format);
- anything (code, documentation) necessary for the reproducibility of your results.

The work can be done by groups up to 4 students.

8 Evaluation

Your work will be evaluated based on:

- the report;
- the quality of your prediction on the test dataset;
- the reproducibility of your results.

9 Plagiarism

Plagiarism is considered a fraud and can lead to the exclusion from the master. We use an automatic tool to detect it. Plagiarism is the fact to *"copy texts or take ideas from someone else's work and use them as if they were one's own"*[1]. It means that everything in your report, document, code, ... that is not your own work must be properly cited: for instance, use a different typeface and/or indentation and give the precise reference of the source (web site, book, paper, ...).

[1] definition from Cambridge Dictionary, `http://dictionary.cambridge.org`