Data Mining: Second project

Mario Martin

November 2, 2020

This document describes what should be done for the second project of the Data Mining course. The goal of this project is to show that you are able to develop a data mining process for a predictive task in a realistic dataset.

1 What should be done in the project?

In summary, in this second project you will:

1. Select a *non-trivial* dataset for a **classification task** (more than 20 columns, more than 1.000 examples and, preferable, with missing, noisy or heterogeneous data)
2. Do the necessary pre-processing for you data. Also, describe and justify it in the documentation.
3. Apply the methods explained in theory lectures to your dataset. If the algorithm requires it, find the best parameters for the algorithm. Tuning of parameters has to be done in a meaningful way. An explanation of the procedure followed and parameters tested should appear in the documentation.
4. Try to interpret the models returned by the algorithms, specially in the case of decision trees.
5. Evaluate and compare different methods applied.
6. Discuss why you think that one method works better than another in *your* dataset.

Of course, for the second project you can apply, if necessary, any technique learned in the first part of the course.

2 What should be specially considered?

Accuracy values are not the most important outcome of your work. At some extent, it depends on the dataset selected. Even when you do everything right, if the dataset is hard, you may obtain low accuracies for your models.

The thing I will evaluate is how you understand the different steps of the data mining process and how do you reason, take decisions, and critically evaluate results you obtain. You should not mechanically execute a script in python and show results. You should be able to explain each step and decision you take on that script.

3 Contents of the written document

The written document to be delivered should contain at least the following sections. For each section there is a short description of what should appear there.

3.1 Description of the original data

Where was it obtained (web page link), description of the problem on hands: target of classification, original number of examples, original number of columns meaning and kind of values, impact of missing values, etc.

3.2 Description of pre-processing of data

Kind of preprocessing done to the original data: Have you simplified the data set? Removed some examples? Feature selection done? Did you enrich your dataset with other columns or more information? Imputing/removing missing values? Simplification of values? Normalization? Remember that you should describe the all procedures performed on your raw dataset.

3.3 Evaluation criteria of data mining models

Description of the procedure followed in order to obtain a representative validation data set and description of the method that will be used to evaluate the different data mining models. This includes description of parameters used in the evaluation (Cross-validation? K-fold cross-validation? How many folders? Why that number?) and discussion of the metric used for evaluation (accuracy, f1, etc).

Which is the splitting procedure of data set into train and validation data set? Description of the procedure followed in order to obtain a representative validation data set.

3.4 Execution of different machine learning methods

For each method you should describe how you adjusted parameters for the algorithm and how did you perform an evaluation of the method. You have to report also the performance of the model on the validation dataset. In addition, for each method you should describe and discus some particular issues:

- **Naïve Bayes**: Think about hypothesis of independence of variables. Do you have enough number of elements to obtain reliable probabilities? Keep that information for the discussion section.
- **K-NN**: Description of procedure followed for choosing the best k-parameter. Show a graph with varying k. Have you adjusted other parameters as distance measure? Have you considered removal of irrelevant features if accuracy is poor compared with other approaches? (remember that k-nn is sensible to irrelevant features when computing distance to closest examples)
- **Decision Trees**: Discussion of choice of parameters used. Try to interpret the obtained DT using some examples of the validation set. Show some of the most relevant rules. Discuss how + and − examples are mixed in leaves in order to estimate the reliability of the tree.
- **Support Vector Machines**: Discuss choice of kernel and parameters used. Did you run any method to speed the building of the model? Report number of supports of the selected machine and try to interpret why the kernel selected and parameters selected for the final run give you the best results for your dataset. Try also to inspect main supports of your machine.
- **Meta-learning algorithms**: Performance Majority Voting, Bagging, RandomForest and Adaboost. Explain parameters selected for each algorithm.

3.5 Comparison and conclusions.

Comparison and discussion of results of the different data mining methods on the validation data-set. Which is the best method when testing on the validation data set? Write a comparative table. Try to compare different methods using McNemar test or at least show the interval of confidence for each method. Is there an explanation for those reults (some hypothesis applicable, etc.)? Are in general accuracy on validation data set similar to the obtained with cross-validation? Are there for some methods huge differences? If that's the case, why do you think that happens? Final personal evaluation of which is the best method you consider and why.

4 Delivery and defense of the project

The presentation of the project will consist in an oral presentation by all members of the group plus the delivery of the documentation. Presentation of the project will be on **16th of December** on time scheduled for the laboratory.

Documentation of the project should be delivered to the *practicals* section *racó* before **15th of December at midnight**. The delivery consists in a zip

file that must contain (1) a pdf file with the documentation of the project, (2) code developed for the project ready to run and (3) the preprocessed data (or a link to the preprocessed data if the dataset is too large to be sent by the racó).

Oral presentation will follow the same rules that were followed for the first project. That is:

- All members should participate in the oral presentation.
- Presentation will be done in front of your colleagues.
- Presentation will not last more that 20-25 minutes, including time for questions.

Members of the group could obtain different marks depending on the individual implication, work done and presentation of the project.