MAC0425 - PE04 Report

Matheus T. de Laurentys, 9793714

July 19, 2020

1 Summary

The purpose of the following project is to exercise important topics discussed in the latter parts of the class. Together with the always expected research on the object of interest, the project relies on selection of features, on data treatment, on a general understanding of machine learning and training and on the specific topics of k-folding method and loss functions. This report, that describes the programming part of the assignment, is also given great importance.

You will find that a large amount of the effort is designated to the treatment of data and preparing it for the training portion portion, which is brief and more focused on displaying and understanding results. This report will follow the steps taken during the project and can be used as a check-list for most things done in the programming parts of the project.

2 Introduction

The treatment of a few hospital's public data and the creation of a neural network that, from the results of other ordinary exams, can predict those specific of COVID-19 is a multipurpose project purpose from professor Marcelo Finger. The extensive work with the hospital's data, prepares the student to handle data-sets or databases one can easily find working outside the academy: they can be elusive with non-standard column names or explanation, can keep unexpected or non-standardized entries, can keep "garbage" data, can be much larger than samples sets and can be poorly planned out.

Another most important motivation for the project is dealing with something that is very current and achieving results that are meaningful outside the learning experience. Working with a small and invented problem prepares for the work with a much more complex and large problem but, unlike this project, might not give the confidence to do so, because they may seem very far apart. Although those two reasons might seem more important, the mechanical ability to treat data-sets, select a learning model and build neural network, exercised in the project, is paramount for any student.

The rest of this text is composed of the description of each step taken in order to achieve the stated results. Methodology is a step-by-step description following the code, while the remaining sections are

3 Metodology

The process to achieve the objectives introduced above started with the preprocessing of data. As the nature of such activity is in itself composed of several parallel tasks, its report will be formatted as a description (or merely naming) of its collection of tasks:

Merge/Separation of the datasets:

The data-sets from the different hospitals will remain separated.
This is mainly due to different notation used is the values inserted in the rows of their tables. Its merge, while attempted, required further knowledge of the different exams. Because the same exam might be written differently (and therefore should be merged), it would require someone of great knowledge to pinpoint when that happens. Not only that, but going through a long list (over 300 exams) and looking for this information would require a lot of work that does not really contribute with the motivation of this work.

One/Several datasets:

The analysis and pre-processing will only use one of the datasets.
This is unfortunate but is mainly due to my lack of preparation. Since I am not merging the datasets, it would be more interesting to analyze all of them separately and check if the results differ (finding, possibly, more meaningful exams). Since I took too much time preparing only one dataset, I will only use one, as to reach the end-goal.

Selecting Dataset:

I selected the dataset with the most rows. That is the data/raw/added/fleury tables.

Removing Duplicates

Mapping values.

As we want to remove non-numerical values, we want to keep the negative/positive results. In order to accomplish that, I will map each value in 0/1.

Removing non-numerical results.

I cannot say for sure why is it important to do so, but this follows the class recommendation.

Dropping Useless Columns:

I chose to remove geographical location from patient dataframe, as that is irrelevant to the contamination measurements. I also removed the DE_ORIGEM columns from the exam table.

Dropping some exams:

It makes sense that some of the exams that very few people did, can be removed as they definitely will not correlate well to the much larger dataset. For this part, with exception to exams I know are important (based on class lecture), I removed all exams that less than 98% (or 2592) of patients did. This is an arbitrary number, but it is a fine one based on my results. As a side note, everything up to this points was done in the pre-process jupyter notebook, and the resulting table is saved as 'SELECTED_EXAMS.csv'. Alternatively, a python script called select_data.py is available.

Analysis of the pre-process results:

At the stage I was planning my neural network I realized my data was still not ready, in fact, that is what inspired the SELECTED_EXAMS.csv file generation. There are two things that are incorrect. The first is that, since we want to predict IGM and IGG values, there is need to merge theirs results, that is spread through different exams. However, these exams, unlike the PCR ones, do are not in 0/1 format, but are shown as float values. Although I do not know how to interpret these values, I was able to create a limiar value based on the PCR results. I estimated that the percentage of people whose PCR exams were negative was the same as those who took the other two exams.

For this part I renamed all of the IGG and IGM exams, in order to merge them, and then, I mapped the values to 0/1 using the calculated limiar.

Join Exams with Patient data:

Important to get the birthday, which may be relevant.

Remove exams with no reference values:

Fore each of PCR, IGG, IGM exams, remove the rows correspondent to people/dates that the exams did not include them.

4 Discussion

Ii is very sad that I did not go further. Although I spent a lot of time doing and researching to to this project, it was not enough. My stopping point code-wise was at the end of the pre-processing, right before the building of the neural network. Because of this the remaining portions of this report were omitted.