ML Assignment: Work Plan

Data Mining, University of Aveiro
2019

Filipe Pires, 85122
DETI, MSc. Informatics Engineering

João Alegria, 85048
DETI, MSc. Informatics Engineering

I. Introduction

The quest for better understanding how our brains work is always an intriguing subject, so both assignment datasets were found interesting by us. However, the Raven test, presented on the assignment instructions [1], caught our eyes as it aimed at analyzing intelligent measures in the general population.

For this reason, and taking in consideration that the participants had background knowledge similar to ours and to people we deal with on a daily basis, we chose the second dataset to analyse using a supervised machine learning algorithm.

In this document we present our work plan for the assignment, including the chosen algorithm, the question to be addressed and the way we are going to organize our time.

II. Dataset

According to the assignment's description, the dataset to be used corresponds to the application of the Advanced Progressive Matrices test with 48 problems to 2 distinct populations: the 1st consisting of 21 students of Design and Multimedia (DM); the 2nd of 24 students of Informatics Engineering (IE).

The data was collected from EEG signal receptors while the participants performed the test tasks. The registered signals were from 8 channels: F3, F4, T7, C3, Cz, C4, T8 and Pz. The relevant marks for the signal analysis are presented in Table 1, along with the respective signal processing windows.

Time Marks	Signal Windows (ms)
Problem Display	[-75 500]
Possible Solutions Display	[-75 500]
Student Answer	[-500 500]

TABLE I

The time features extracted from the raw data were **P100** and **P300**, due to their significant characteristics for the study. Also, through the energy (E) of the characteristic bands estimated in all defined windows, some frequency features defined as energy ratios are included: **Stress Index; Mental Fatigue; Alpha Lateralization; Immersion Index**.

Having gathered this knowledge and analysed the available dataset and its structure and division between train and test sets or average and non-average results, we were able to discuss how could the analysis be done and what algorithms might be useful for the proposed machine learning tasks.

III. Goals & Strategies

The first goal of our work is to understand in greater depth the dataset, its quality and normalization, its features, and search for outliers and the percentage of non-successful measurements on the trials. Once a refined dataset is achieved, we intend to address the more demanding question of whether there is a significant difference between the 2 groups of participants or not, and, if so, how might this difference be characterized.

With this in mind, and considering the fact that the solution should address supervised algorithms only, we decided to apply the following machine learning algorithms to the dataset to seek for answers to our posed questions: **Neural Network** and **Support Vector Machine**. We chose these algorithms since we are facing a classification problem, where the possible classes are the study field of the participant whose features are used as input. SVMs and NNs are famous for their high performance in issues such as these, but we are considering exploring additional promising alternative worth testing.

IV. Work Plan

The first phase of development will be to analyse the available features (extracted in different ways) and use only the data considered relevant. The resulting dataset is to be divided into 3 subsets: train, validation and test.

Then, we intend to develop the algorithms in Python, with the help of libraries dedicated to statistical analysis and data science, and train the models with the proper subset. A study is to be conducted on hyperparameters variation and the models are to be refined with the knowledge gained from this.

At this phase we will compare the algorithms in terms of performance, using several appropriate metrics. Simultaneously we will also attempt to answer the question "are the two groups different?" based on the results.

The work is to be fairly distributed between team elements and we intend to schedule dedicated time for the project in compatible hours so that more important issues can be addressed by the whole team.

References

[1] A. Tomé, "Data Mining Assignment", https://elearning.ua.pt/pluginfile.php/1496406/mod_resource/content/3/ED_HCT_Raven.pdf, accessed in November 2019.