Advanced Topics in SMC – Lab 02

Quim Marcé

For this project, I developed a feature selector and an MLP classifier to perform the goal of choosing the best features to predict the mood class of the songs present in the provided dataset. The code has been implemented in a python notebook which can be found in the following link:

GitHub Repo

So, one of the relevant things I did to make the data preprocessing easier was to delete all the empty rows in the both train and test sets, to avoid code issues dealing with the NaN values. This is important to be taken into account for the code to be tested, because otherwise it will crash. I spent a lot of time dealing with this kind of errors and the most simple but effective way to solve them was modifying the csv files directly.

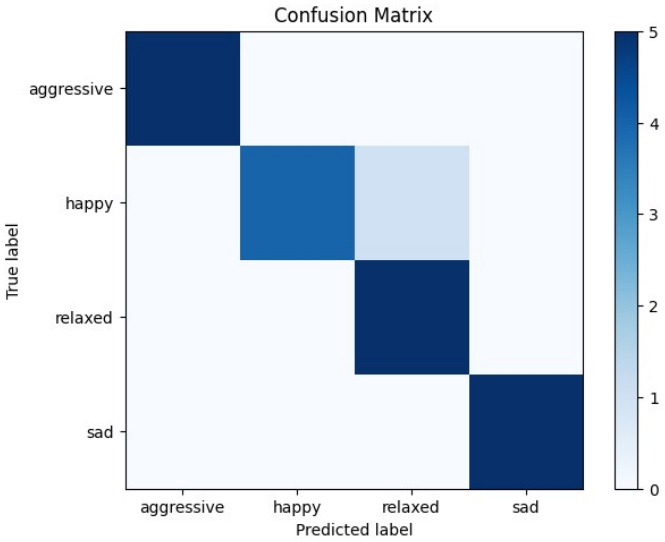
Other than this, the methodology used for the project is pretty straightforward. I used pandas to work with the csv files in a proper way. Then, I concatenated both train and test sets to ensure a consistent label encoding for the file names, as many *sklearn* models have issues dealing with variables which are not encoded as floats or numbers. Both sets were afterwards separated and also X and Y subsets were created for an adequate future prediction process. After this, I still had to deal with the NaN values present for some features, so I used the *imputer* dependencies included in the *sklearn.* Being honest, my programming skills are not the best so forums such as stackoverflow and the use of AI tools such as ChatGPT helped me solving some doubts in the data preprocessing tasks.

Once the data was ready for the task to be performed properly, I used the SelectKBest function from *sklearn* to choose the best 5 features for the given task. The obtained features were the following:

EDm – Event Density
SEm – Spectral Entropy
SFm – Spectral Flux
Rsm, RDm – Repetition Spectrum

At first what jumps out is the unbalanced distribution of the features category. For example, there is just one feature regarding the rhythm of the songs (event density), while the rest of them refer to the spectral properties of sound. I would not say that the importance of the spectrum properties is surprising, but at least I expected the appearance of some features corresponding to harmony and dynamics of music, as to my perception, they are fundamental for determining the mood of a song (ie, there are many chord progressions directly associated to song moods).

A part from this, the chosen model for the prediction task was an MLP Classifier due to its flexibility, robustness and versatility. It was trained with the corresponding set and tested with the selected features, and the results were really satisfactory for this case. Over the 20 song test set, it got an accuracy of the 95% when classifying songs by mood according the previously exposed features. In absolute numbers, this means that it only missclassified one song which is the *dream-world* one. The following plot of the confusion matrix shows the results:



These results suggest that, besides the model performs obtains really accurate results for this set, it is probably overfitted, which means that its generalisation becomes ambiguous. The mood classification task I think it would be better performed by a human, as in the end, it can easily become a subjective task. To avoid this kind of overfitting I would suggest reducing the number of features to take into account to get something more realistic in terms of mood classification and the system's accuracy.