M2 NLP - Signal processing

TD 1 : Speech biomarkers of depression

This practical class aims at reproducing (a part of) the results from this research article : [Tao et al. 2023, Interspeech]. In this article, the authors introduce a new corpus for depression detection using speech. The originality of their work is that the corpus is annotated using robust diagnosis : the dataset is labeled using binary targets (depressed vs. non-depressed). Using statistical features (a custom openSMILE features subset), our aim is to replicate the *BS1* approach described in the article.

The example audio file, as well as the features used for the classification, are available in the following online archive : [link]

1 Extraction of features

In order to get familiar with features extraction, we will extract some features from an example audio file `audio_example.wav`.

→ **What do you think of the acoustic quality of recording ?**

Using `audio_example.wav`, extract the following feature sets :
— eGeMAPS
— ComParE_2016

→ **How many features contain each feature set ? Are they all interpretable ?**

2 Classification pipeline

2.1 Features

Since I'm not allowed to share the entire database with you, I've already extracted the features in the same way as you did in the previous question. In the downloaded archive, you will find two folders, 'reading task' and 'interview', containing different sets of features. In the following questions, we will focus on the features used in the article (named "original" in the archive).

→ **Could you describe these features ? Are they all interpretable ?**

2.2 Dataset & Performance metric

2.2.1 Dataset

→ **Why the label of this dataset could be considered as more robust that the previous one (see article) ?**

2.2.2 Performance measurement

→ **Is the dataset balanced ? What will be the most relevant metric ? How is it implemented in `sklearn` ?**

→ **How is this cross validation implemented in `sklearn` ?**

2.3 Scaling

→ **Why do we scale features ?**

2.4 Classifier

The classifier used is an SVC, with a Gaussian kernel (rbf), and with the default parameter C = 1,

2.5 Evaluation

In order to avoid overlearning, we will use a *cross-validation* procedure. In the same way as the original article, we will use a `5-fold` cross-validation, and compute the average of the performances across the folds.

3 To go further ...

3.1 Contribution of the features

How could you analyze the relative contribution of the different features in order to identify the most important ones ?

3.2 Other speech features

What are the performances obtained with different feature sets ? Are the most important features in the classification the same for each group of features ? Are any of them identical ? Different ?

3.3 Other classifiers

Could you obtain better results using different classifiers ? Or by optimizing the hyperparameters of the classifiers ? (you may need double cross-validation !)