NTDS: Project Proposal

Epileptic seizures prediction

Sophie du Bois

This project is based on a closed competition that took place on Kaggle in 2014. The goal is to predict the onset of epileptic seizures from electroencephalogram (EEG) signals. In the end, this could allow the patients to avoid dangerous activities when a seizure is probable to happen, or to take their medications only at these time points. To do so, one hour long EEG signals have to be classified into *interictal* (between seizures) or *preictal* (before seizure, Fig.1).

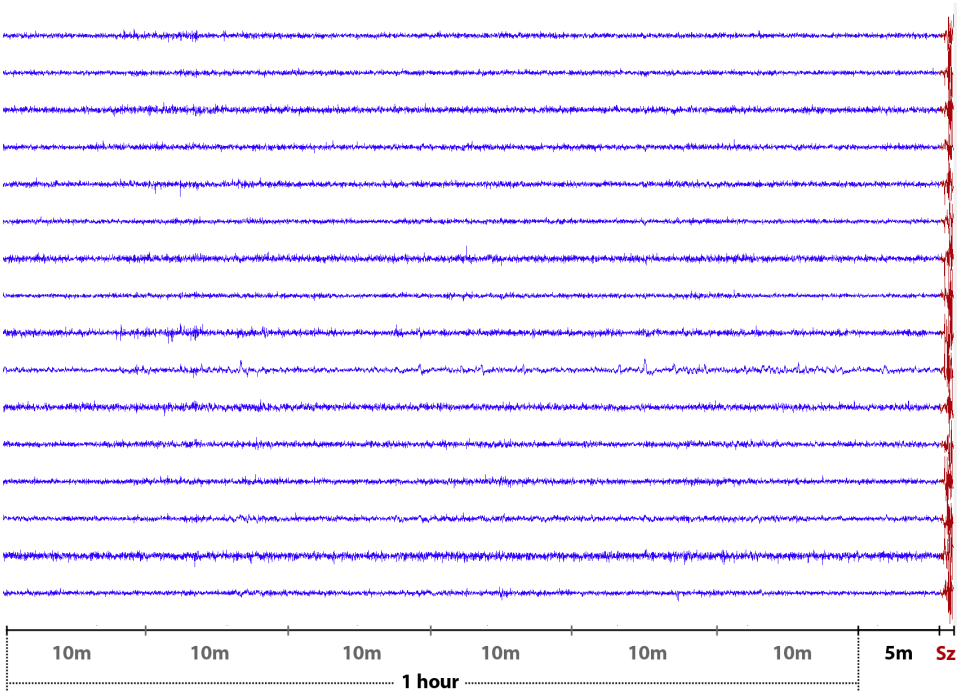

Figure 1: One datapoint consist of multiple EEG channels recorded during one hour.

1 Data acquisition

EEG sequences of 1 hour are available on Kaggle, as `.mat` files. Data is already separated into train (labelled) and test (unlabelled) sets. Coming from an online contest, data should already be quite clean. However, it is still possible that some EEG channels are missing, or take aberrant values. Furthermore, if the sampling frequency of the different sequences is not the same, it could lead to interpretation problems, and the data should thus be resampled.

2 Data exploration

Data is coming from both animal subjects (dogs, for most of the datapoints), and human patients (less recordings performed). The anatomical positions of the electrodes, and the recording devices may differ. In addition, as the inputs are temporal signals, frequency analysis could eventually give some useful information for classification. Field knowledge could also be required to preprocess the signals.

3 Data exploitation

Each datapoint that has to be classified consists of a one hour long recording. If enough datapoints are available, it could be possible to directly use the time samples as inputs of the model. However, it may also be necessary to reduce the input dimensions, either by down-sampling the signals, or by extracting features using field knowledge. Indeed, some EEG frequencies seems more related to epileptic seizures than other, and frequency (or time-frequency) analysis may be useful in this problem, as well as filtering and prepocessing.
The size of the datasets will dictate the kind of classifier to be used. Indeed, neural networks may be too complicated to implement

4 Evaluation

For the Kaggle contest, submissions were judged on the area under the ROC curve (AUC). This measure takes both sensitivity and selectivity into account. In total, there are 3,935 unlabelled datapoints. As the competition is over, the true labels were actually published.
A model should be developed on the training set, probably using cross-validation to assess the different parameters. The result could then be tested on the "unlabelled" test set, in order to stay in the spirit of the competition, during which the number of test submissions was limited.