M1 Info Data Science
Spring 2022

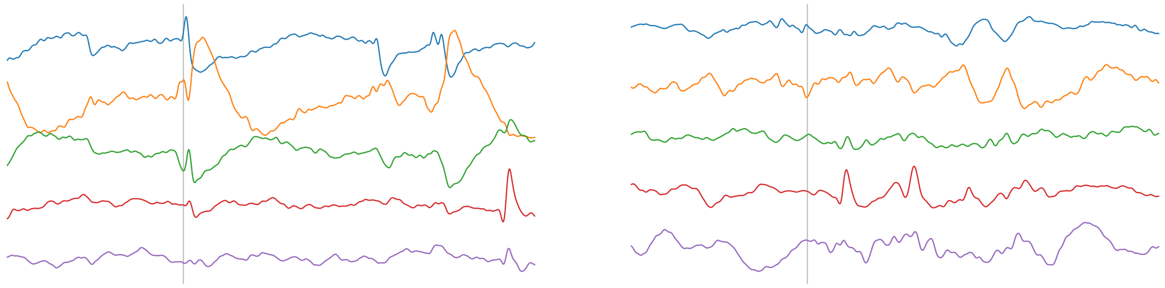
Course Project
Themis Palpanas, Qitong Wang

Topic:

EEG epileptic spike detection is a crucial and challenging problem in both data science and neural science. The goal is to develop an algorithm to automatically detect whether an instance (such as EEG data series) captures an epileptic spike or not. Two common challenges are a). epileptic spike behaviors are minorities compared to normal behaviors; b). epileptic spike morphologies across different patients are quite different. In this project, you will tackle both challenges in the two tasks.

Context:

The use case that will be tackled during this project is the analysis of multi-channel EEG data series. Our datasets were extracted from two different experiments and preprocessed for desensitization.
Specifically, our dataset consists of multiple multivariate data series with a constant frequency, taken from two different experiments and split into fixed-length subsequences. Each series has 5 channels, and each channel has 768 sample points (aligned by time). It represents either an epileptic spike (i.e., a positive instance, to be labeled as 1), or a normal EEG series (i.e., a negative instance, to be labeled as 0). The two types of data series are depicted in Figure 1, where the vertical line in the positive example indicates when the spike happens.



(a) positive example: data series of epileptic spike that appears in the orange channel (in other examples, the spike may appear in a different channel, in 1 or more channels)

(b) negative example: normal data series

Figure 1: Two types of data series of the dataset

Step 1:

You will have at your disposal 400 epileptic spike data series (positive instances) and 2000 normal data series (negative instances) from experiment 1. Each data series contains 5x768 sample points. Your goal is to develop a method that is able to recognize correctly if a series represents a series of epileptic spikes or a normal EEG series.

You will provide the predicted label (0: normal, or 1: epilepsy spike) on a test set of another 100 epileptic spike data series and 500 normal data series from both experiment 1 and 2, i.e., 200 positive and 1000 negative series in total. We will use Precision/Recall/F1 regarding epilepsy spikes to evaluate your method.

You should also measure and report the execution/training time: between two methods that have the same accuracy, the faster method is better.

Deadline for Step1: Mar27 11:59pm

Step 2:

In addition to the 400 positives and 2000 negatives you already have in step 1, you will have at your disposal extra 400 epileptic spike data series (positive instances) and 2000 normal data series (negative instances) from experiment 2. Each data series also contains 5x768 sample points. Your goal is to develop a method that is able to recognize correctly if a series represents a series of epileptic spikes or a normal EEG series.

You will provide the predicted label (0: normal, or 1: epilepsy spike) on a test set of another 100 epileptic spike data series and 500 normal data series from both experiment 1 and 2, i.e., 200 positive and 1000 negative series totally. We will use Precision/Recall/F1 regarding epileptic spikes to evaluate your method.

You should also measure and report the execution/training time: between two methods that have the same accuracy, the faster method is better.

Deadline for Step2: Apr24 11:59pm

Datasets:

Input:

- Step1:
    - Train: 400 positive + 2000 negative instances from experiment 1, each contains 5x768 points
    - Test: 100 positive + 500 negative instances from both experiment 1 and experiment 2 respectively, i.e., 200 positives and 1000 negatives totally, each contains 5x768 points
- Step2:
    - Train: 400 positive + 2000 negative instances from both experiment 1 (as in step 1 train) and experiment 2 respectively, i.e., 800 positives and 4000 negatives totally, each contains 5x768 points
    - Test: 100 positive + 500 negative instances from both experiment 1 and experiment 2 respectively, i.e., 200 positives and 1000 negatives totally, each contains 5x768 points

Output:

- Labels for the series in the test set (remember: the test dataset contains 1200 series, both normal and epilepsy spike), expressed as integer values: 0 means normal, 1 means epileptic spike (in a csv file with one value per row, for a total of 1200 rows).

Implementation details:

- The input data is stored in binary files.
- Suggested way to read these input files:
    - import numpy as np
    - np.fromfile(‘exp1-train-400pos.bin’, dtype=np.float32).reshape([400, 5, 768])

Bonus Task:

Once the classification is performed and the accuracy is satisfactory, a bonus task is to find and explain the classification decision of your algorithm. Based on your trained algorithm, provide a method that highlights the subsequence of the series your algorithm labeled as epileptic spike that contributes the most to the algorithm's choice to classify it as abnormal. In other words, which part of the series that your algorithm labeled as epileptic spike played the most significant role for this classification decision.