Machine Learning Engineer Nanodegree
Capstone Proposal

Jean-Christophe Quillet
22/11/2016

1. Domain Background

The project proposed is within the healthcare domain, and more specifically address the prediction of epilepsy seizures.
Epilepsy afflicts nearly 1% of the world's population, and is characterized by the occurrence of spontaneous seizures which can result in physical injuries. Anticonvulsant medications can be given at sufficiently high doses to prevent seizures, but patients frequently suffer side effects. For 20-40% of patients with epilepsy, medications are not effective. Epilepsy surgery may be an option when treatments are ineffective, but many patients continue to experience spontaneous seizures.
Despite the fact that seizures occur infrequently, patients with epilepsy experience persistent anxiety due to the possibility of a seizure occurring. Therefore, seizure forecasting systems have the potential to help patients with epilepsy to live more normal lives.

A possibility is to develop an electrical brain activity (EEG) based seizure forecasting systems able to identify periods of increased probability of seizure occurrence.
If seizure-permissive brain states can be identified, devices designed to warn patients of impeding seizures could be developed. Patients could punctually avoid potentially dangerous activities like driving or swimming, and medications could be administered only when needed to prevent impending seizures, reducing overall side effects.

This project is the subject of a Kaggle competition proposed by the Melbourne University AES (link: https://www.kaggle.com/c/melbourne-university-seizure-prediction). There is currently no technology available that can reliably predict imminent seizures. A first limited trial, by the university research team, using a small device implanted in the brain has showed possibilities to predict seizures. The dataset currently available is the result of a larger study led by this research group.

I am personally focused on finding ways to utilize Machine Learning methods for innovative healthcare purposes. The study of EEG signals in particular holds a lot of promising future healthcare applications. This topic is also for me the opportunity to work in depth on signal processing methods which have applications in many other domains.

2. Problem Statement

The problem is to demonstrate the existence and accurate classification of a pre-seizure brain state in humans suffering from epilepsy from existing intracranial EEG recordings.
The performance of the model will be assessed by the prediction accuracy on the available test sets.

3. Datasets and Inputs

The temporal dynamics of brain activity can be classified into 4 states:

- Interictal (between seizures, or baseline),
- Preictal (prior to seizure),
- Ictal (seizure)
- Post-ictal (after seizures).

The goal of this project is to differentiate the preictal and interictal states.

Human brain activity was recorded in the form of intracranial EEG (iEEG) which are measures of electrical voltage fluctuation. For all the patients, iEEG was sampled simultaneously from 16 electrodes at 400 Hz, and recorded voltages were referenced to the electrode group average.
Each file in the dataset corresponds to ten-minute-long data segments from measures covering an hour prior to a seizure (preictal states), and ten-minute-long data segments of interictal states. Each data segment includes 16 channels.

Preictal data segments are provided covering one hour prior to seizure with a five-minute seizure horizon. (i.e. from 1:05 to 0:05 before seizure onset.)
The interictal data segments were chosen randomly from the full data record, with the restriction that interictal segments be at least 4 hours away from any seizure.
It has to be noted that some data segments contain 100% data drop-out and the data may also contain artifacts such as large amplitude rapid signal transitions which need to be sorted.
The dataset is pre-divided between a train and a test set.

The full train set provided is composed of recordings for 3 patients.
The present analysis is carried out on a subset corresponding to the patient 3 and composed of 150 preictal samples and 759 interictal samples, of 10 minutes each.

I use this training set as a train/test dataset for cross-validation of my model. The test set is composed of 20% of the subset samples with the same proportion of preictal and interictal samples as the train set.

In this project the ten-minute-long data segments will be first divided over each of the 16 channels, each electrode recordings will be considered independently, and furthermore into one-minute signal samples. For prediction purpose, the classification of a 10-minutes samples will be the result of the prediction of the 16 related channels.

The data samples are then being processed into a set of different type of features using multiple signal treatment process. This allows to efficiently compare the signal samples.
Temporal and frequency based features
First the signals are processed through a discrete Fourier transform, using a Fast Fourier transform algorithm, to allow their spectral analysis
The signals are also processed through a discrete wavelet transform that in addition to spectral information give some temporal information.
The signal dynamic signature is studied through the analysis of its fractal dimension using Katz and Higuchi.
The cross-correlation is calculated between successive one-minute signal samples of the same channel and averaged over a 10-minutes sample.
Kurtosis and skewness of the data samples are also computed as features.

This selection of features is inspired by the multiple approaches described in the current published researches. Studies have been carried out over limited dataset and no single solution has shown a decisive superiority on epilepsy seizure detection.

4. Solution statement

I propose to use a supervised learning approach to infer a classifier from the dataset that will be able to distinguish iEEG signal segment in a pre-seizure period from a baseline period. The prediction performance of such classifier can first be assessed against the test dataset available. It can be use on future studies.

5. Benchmark Model

The model prediction performance is assessed by comparison with a naive bayes algorithm. The test set used for submission evaluation by Kaggle in the competition is composed of samples without label. As a first approach focusing on one patient, I use a labeled train/test dataset. The test set is composed of 20% of the subset samples with the same proportion of preictal and interictal samples as the train set. This method allows more control over the test dataset for improving the model. The model performance will be compared over the test set with the benchmark model performance.

6. Evaluation Metrics

The model prediction performance on the test set is assessed by the F1 score obtained on the test set.
The equation for the F1 score is the following:

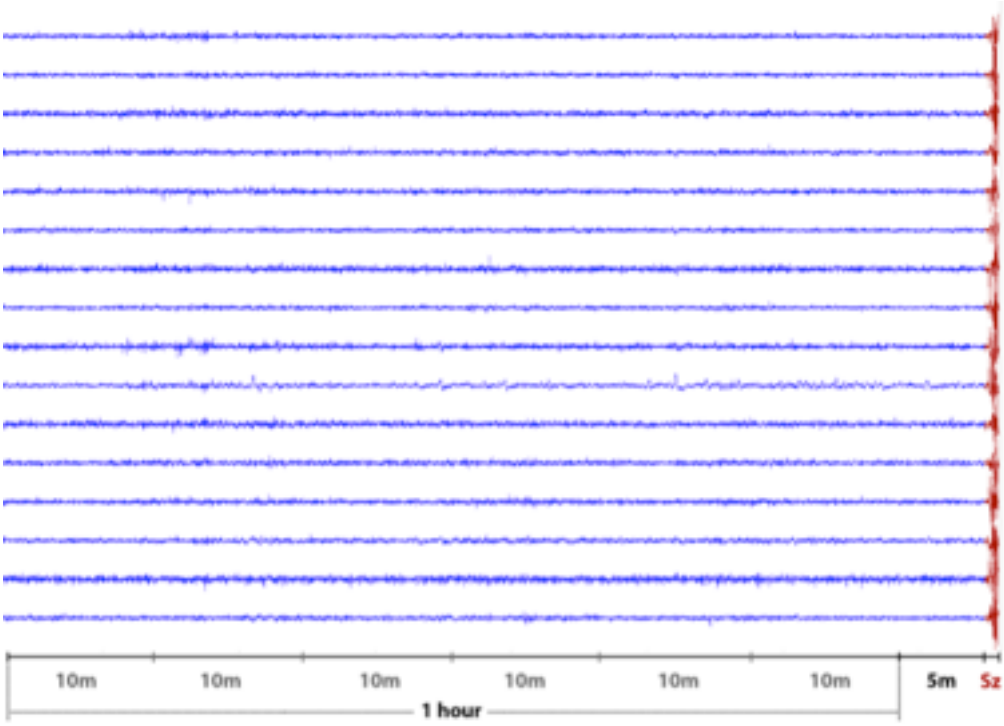
F1 = 2 * (precision * recall) / (precision + recall)

Considering the importance of predicting seizures correctly for the patient comfort, it is necessary to calculate the recall on preictal (pre-seizure) periods prediction, which represent the ability of the model to predict the positive samples.
The equation for the recall is the following:

Recall = true pos predictions / (true pos predictions + false neg predictions)

7. Project Design

For this project, I have at my disposal a set of labelled data composed of iEEG signals from baseline and prior to seizure periods, and a table listing the files that can be safely used as well as the label of data.



Pre-ictal one-hour iEEG signal recording (16 channels)

The first task is to prepare this data for the analysis, since sequences of iEEG signals cannot be easily compared without some pre-processing.

I consider a data pre-processing program that implement the following workflow:

- Open and run through the .csv file listing the safe files from the dataset and their label,
- Successively open the safe .mat (Matlab) files and convert them into pandas dataframes that can be handled more easily,
- Merge those dataframes,
- Remove from the dataframes the rows containing no or only null values,
- Remove the samples containing less than 2 minutes of recording without drop-off, since it is not sufficient for the estimating the features defined,
- Subdivide the iEEG signals into 1-minute-long samples (24 000 datapoints with the 400Hz sampling rate),
- Process a Fast Fourrier Transform (numpy function) of the samples,
- Process a Discrete Wavelet Tranform,
- Calculate features from the sample DFT and DWT that are characteristics of the signal:
    - Entropy
    - Power Spectral Density
    - Spectral Edge Frequency,
    - Hjorth parameters (Activity, Mobility, Complexity)
    - Kurtosis

    - Skewness
    - Fractal Dimension (Katz, Higuchi, Hurst)

  Those characteristics are used to differentiate EEG signals in neurological studies, the project will determine which are useful for seizure-risk identification.
- Calculate feature scaling using the MinMax method,
- Store the dataframes created in .csv files for future use.

The next step is to select a supervised learning method and to train the model with the dataset. I plan on trying multiple supervised learning solutions for this classification problem and use the performance metrics to select the method giving the best prediction performance.
I consider a program that implement the following workflow:

- Open the.csv files into pandas dataframes,
- Shuffle and split the dataset into training and testing samples,
- Successively train and make predictions using the following classifiers:
    - SVM
    - Adaboost (Ensemble method)
    - Gaussian Naïve Bayes (the benchmark solution)

The last step, after selecting a supervised learning model according to the performance metrics, is to optimize this classifier.

This process is completed by running a feature preprocessing to find which compound combinations of features best describe the class of a signal and therefore possibly reduce the dimensionality of the dataset.:

- Run a PCA analysis using the Sklearn function and display the explained variance for each feature. An analysis is necessary to choose how many features to use after PCA. To prioritize the prediction performance, I decide conservatively to keep a number of component that retains 99 % of variance.
- Train the selected supervised learning model,
- Compare the result obtained without this preprocessing.

References:

- Cook MJ, O'Brien TJ, Berkovic SF, Murphy M, Morokoff A, Fabinyi G, D'Souza W, Yerra R, Archer J, Litewka L, Hosking S, Lightfoot P, Ruedebusch V, Sheffield WD, Snyder D, Leyde K, Himes D (2013) Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. LANCET NEUROL 12:563-571.
- Brinkmann, B. H., Wagenaar, J., Abbot, D., Adkins, P., Bosshard, S. C., Chen, M., ... & Pardo, J. (2016). Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. Brain, 139(6), 1713-1722.
- Gadhoumi, K., Lina, J. M., Mormann, F., & Gotman, J. (2016). Seizure prediction for therapeutic devices: A review. Journal of neuroscience methods, 260, 270-282.
- Karoly, P. J., Freestone, D. R., Boston, R., Grayden, D. B., Himes, D., Leyde, K., ... & Cook, M. J. (2016). Interictal spikes and epileptic seizures: their relationship and underlying rhythmicity. Brain, aww019.
- Andrzejak RG, Chicharro D, Elger CE, Mormann F (2009) Seizure prediction: Any better than chance? Clin Neurophysiol.
- Snyder DE, Echauz J, Grimes DB, Litt B (2008) The statistics of a practical seizure warning system. J Neural Eng 5: 392–401.