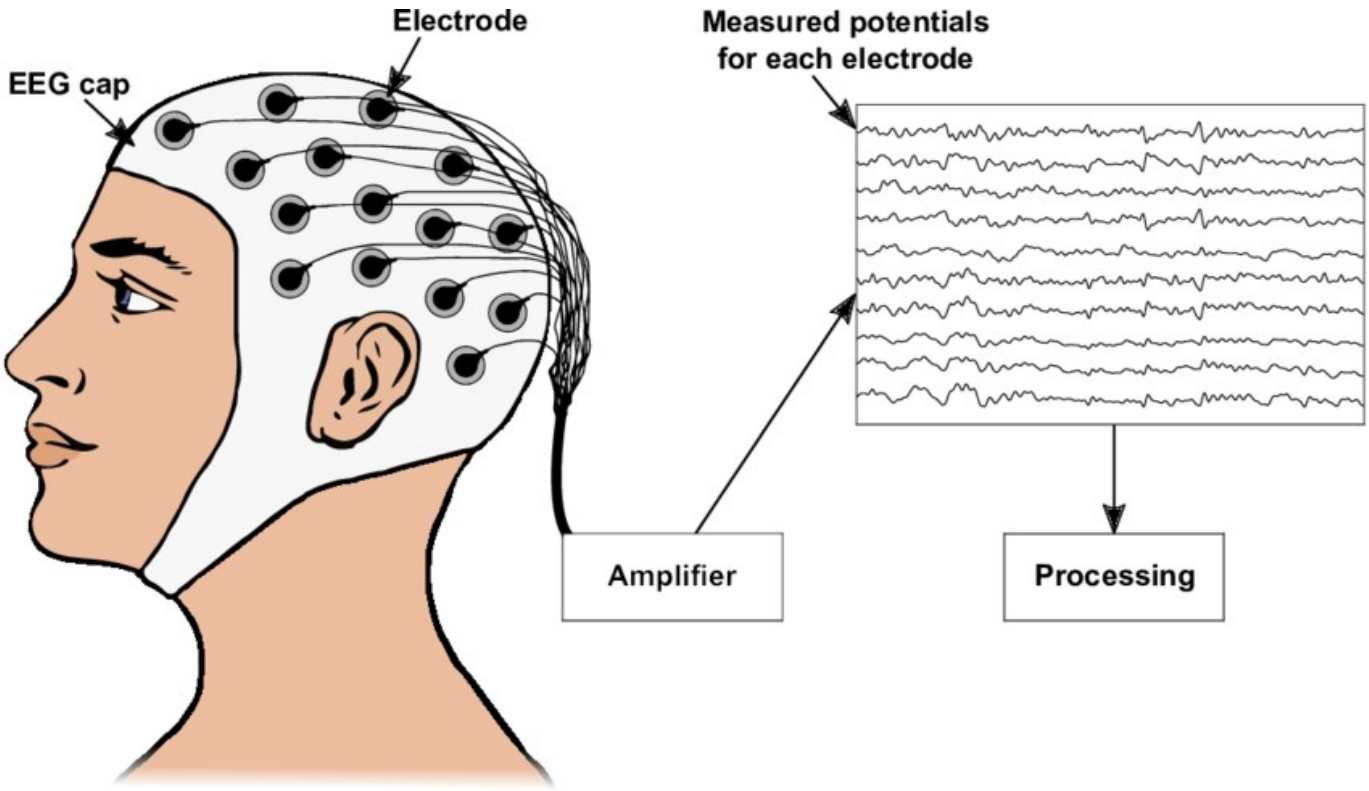
Aspiring Engineer: Catto Alex

Capstone Project:

Select and develop a project on your own design

Confused Person EGG Brainwave



Using encephalogram (EGG) brainwaves to classify a person state of confusion

Domain Background

Our field of research is the EEG (electroencephalography) analysis. This methodology measures the brain activity by detecting ‘”brainwaves “, or rather, electrical signals emitted continuously by the brain itself.

References

Ariely, Dan, and Gregory S. Berns. "Neuromarketing: the hope and hype of neuroimaging in business." *Nature reviews neuroscience* 11.4 (2010): 284.

Barnett, Samuel B., and Moran Cerf. "A ticket for your thoughts: Method for predicting content recall and sales using neural similarity of moviegoers." *Journal of Consumer Research* 44.1 (2017): 160-181.

Boksem, Maarten AS, and Ale Smidts. "Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success." *Journal of Marketing Research* 52.4 (2015): 482-492.

Christoforou, Christoforos, et al. "Your Brain on the Movies: A Computational Approach for Predicting Box-office Performance from Viewer’s Brain Responses to Movie Trailers." *Frontiers in neuroinformatics* 11 (2017): 72.

Dmochowski, Jacek P., et al. "Audience preferences are predicted by temporal reliability of neural processing." *Nature communications* 5 (2014): 4567.

Halchenko, Yaroslav O., and Michael Hanke. "Advancing Neuroimaging Research with Predictive Multivariate Pattern Analysis (MVPA)." (2010).

Harmon-Jones, Eddie, et al. "The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update." *Biological psychology* 84.3 (2010): 451-462.

Krugman, Herbert E. "Brain wave measures of media involvement." *Journal of Advertising Research* 11.1 (1971): 3-9.

Luck, Steven J., and Emily S. Kappenman, eds. *The Oxford handbook of event-related potential components*. Oxford university press, 2011.

Ma, Qingguo, et al. "P300 and categorization in brand extension." *Neuroscience letters* 431.1 (2008): 57-61.

Ramsøy, Thomas Z., et al. "Frontal Brain Asymmetry and Willingness to Pay." *Frontiers in neuroscience* 12 (2018): 138.

Ravaja, Niklas, et al. "Predicting purchase decision: The role of hemispheric asymmetry over the frontal cortex." *Journal of Neuroscience, Psychology, and Economics* 6.1 (2013): 1.

Telpaz, Ariel, et al. "Using EEG to predict consumers' future choices." *Journal of Marketing Research* 52.4 (2015): 511-529.

Neuromarketers scientists are investing a lot of time and resources in this field of research. This analysis is a way to understand the human behavior, and, what happens in our brain when it communicates and synchronizes activities across different anatomical regions. The cognitive process is defined by the variation of these signals, both in humans and animals. This knowledge may be applied for many technological solutions able to detect the will of a person and to process it in order to transmit that information. In recent years, EEG equipment has become more inexpensive, portable, and wireless, opening up new possibilities for mobile, in-store, and virtual reality studies. New statistical and machine-learning techniques are starting to be used to decode and interpret the EEG signal at the level of the full brain, portending many new and original findings.

Brainwaves can be measured with electrodes placed on the scalp and connected to a signal amplifier. There are three types of EGG measurement:

- Brainwave Frequency Analysis
- Hemispheric Asymmetry Analysis (an application of Frequency Analysis)
- Event-Related Potential Analysis.

Brainwave signals emitted by the brain have frequency characteristics. Electrical frequency is measured in hertz (cycles per second) . The most common frequency is 10 Hz (10 cycles per second). Frequencies change due to different mental states, and also vary over time and across different parts of the brain.

Brainwave frequencies have been classified in these groups:

- *"Delta"* less than 4 Hz, dominant frequency in dreamless sleep;
- *"Theta"* 4-8 Hz, associated with internally focused processing, such as memory activation and conscious concentration;
- *"Alpha"* 8-12 Hz, the brain's "default" frequency, dominant when the brain is in a relaxed state and suppressed under the influence of attention;
- *"Beta"* 13-30 Hz, associated with alertness, active attention, and reward expectation,
- *"Gamma"* greater than 30 Hz, associated with information processing, learning, and emotional processing.

In order to measure brainwave frequencies, two metrics are commonly used :

- *"power"* measures the activity at a particular frequency, over a period of time;
- *"coherence"* measures the correlation of frequencies across different parts of the brain.

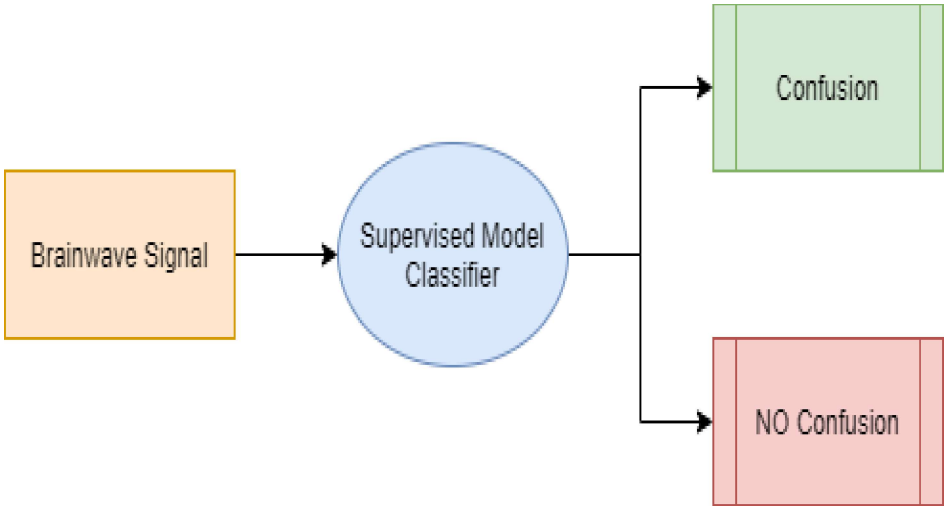
Problem Statement

The purpose of this project is the trial of developing some automation able to analyze a brainwave and to classify it, in order to understand if the person on which the signal has been extracted was in a mental state of confusion at the moment.

First, we want to understand if there are some patterns in the brainwaves signals able to correlate different frequencies to our specific mental state research. What are the frequencies that affect the confusion state the most? Is it possible to cluster data in order to get a clear separable groups of frequency activities assignable to different mental state labels?

Furthermore, we want to build a classifier able to define if a brainwave signal sample owns to a confused person or not. Our question is "Is this signal a confusion mental state?"

In order to do this, we are going to use unsupervised and supervised Machine Learning techniques.



This application is able to detect if a person is in front of something new or difficult or unknown. This detection should be used as part of a modern "Truth Machine" or for computerized auxiliary systems able to understand when you need for more explanations (such as the fanta-scientific IronMan intelligent armor or StarWars protocol robot) or for user experience improvement automation.

Datasets and Inputs

In order to solve our problem, we need for samples of brainwaves labeled with a state indication of confusion. We will use a dataset of samples collected by a group of students and released for public access on Kaggle at the URL https://www.kaggle.com/wanghaohan/confused-eeg.

This data has been collected by submitting a test to 10 college students. The students was asked to wear the EGG signal detector while watching some MOOC video clips (Massive Open Online Courses). These videos have been selected according to students knowledge. Everyone had to watch for both a simple subject video and an unknown/difficult one. Furthermore, the videos expected to be muddler have been chopped with the purpose of showing the middle of the topic explanation: by this way, the subject should have been much more confusing.

The EGG cap was made of a single-channel wireless MindSet able to measure the activity over the frontal lobe: the voltage between an electrode resting on the forehead and two electrodes (one ground and one reference) each in contact with an ear. Every recording was two minutes long.

Every student, at the end of the video clip, rated his confusion level. According to this, a label (confused: yes or not) has been created and attached to the signal recording. Every signal data has been enriched with the indication of the student ownership (with his age and nationality data) and the video reference.

The data have been structured in this way:

- Every signal has been reduced to a one minute shot;
- Every minute of recording has been sampled every 0.5 seconds, thus every signal is represented in more than 120 rows.

The data representation is structured in different files:

- EEG_data.csv: Contains the EEG data recorded from 10 students, with features
    - SubjectID, the student identifier code
    - VideoID, the watched videoclip identifier code
    - Attention, proprietary measure of mental focus
    - Mediation, Proprietary measure of calmness
    - Raw, Raw EEG signal
    - Delta, 1-3 Hz of power spectrum
    - Theta, 4-7 Hz of power spectrum
    - Alpha1, Lower 8-11 Hz of power spectrum
    - Alpha2, Higher 8-11 Hz of power spectrum
    - Beta1, Lower 12-29 Hz of power spectrum
    - Beta2, Higher 12-29 Hz of power spectrum)
    - Gamma1, Lower 30-100 Hz of power spectrum
    - Gamma2, Higher 30-100 Hz of power spectrum
    - Predefinedlabel, whether the subject is expected to be confused
    - user-definedlabeln, whether the subject is actually confused

  This file contains 12811 rows of samples for 15 features.

- demographic.csv: Contains demographic information for each student
  - SubjectID, the student identifier code
  - Age, the student Age
  - Ethnicity, the student Ethnicity (TYPE string)
  - Gender, the student gender (M/F)

  This file contains 10 rows, as the 10 students submitted to the test
- video data : Each video lasts roughly two-minute long, we remove the first 30 seconds and last 30 seconds, only collect the EEG data during the middle 1 minute.

Solution Statement

In order to solve this problem, it is necessary to apply some unsupervised and supervised Machine Learning techniques.

Exploration is needed both for data and models, thus a Jupyter notebook will be used. In this notebook data will be explored in order to find the best pre-processing solution. Moreover, different models will be tuned with different hyperparameters and will be compared one to each other. The best solution will be detected in the notebook.

After that, a Python application will be designed with the pre-processing data steps, the chosen model training process and the prediction module:

- DataManagement.py is a module containing data processing functions
- Train.py is an application able to train the model and to save it on a dump file
- Predict.py is an application able to classify a brainwave.

Benchmark Model

The author of the dataset collection tried to apply Machine Learning binary classification on the dataset.

I attach here their Benchmark, as written on Kaggle:

"*This dataset is an extremely challenging data set to perform binary classification. Here are some recent classification results for reference:*

- *SVM: 67.2%*
- *Bi-LSTM (Ni et al 2017): 73.3%*
- *CF-Bi-LSTM (Wang et al 2018): 75.0%*"

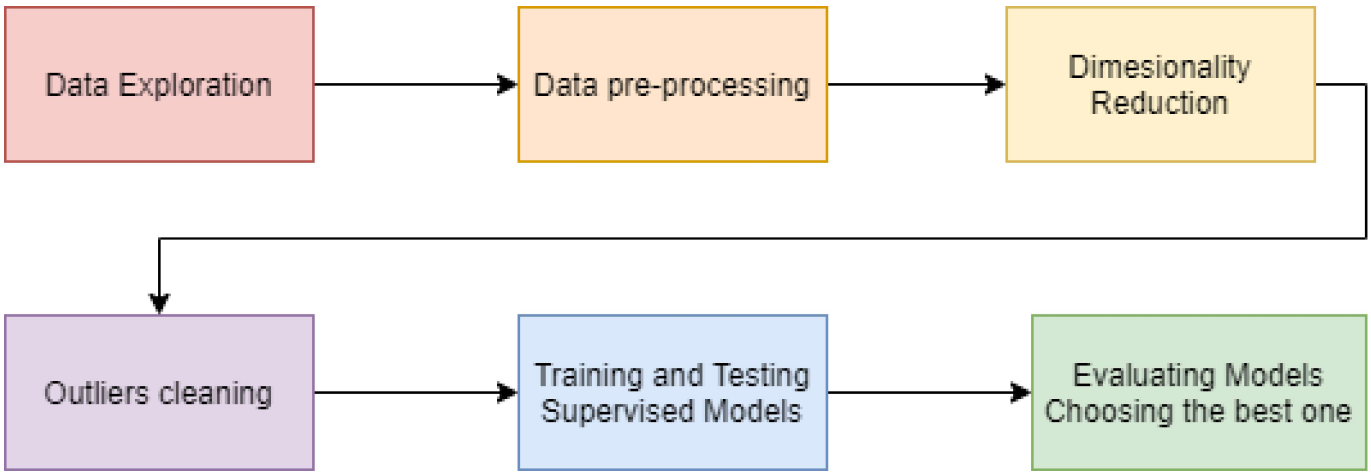
These values of model accuracy may be used for benchmark comparsion between the model of this project and the already existing one.

Evaluation Metrics

The Benchmark model is evaluated with accuracy score. For this reason, the model of this project will be evaluated in the same way. Moreover, Precision and Recall metrics will be evaluated, in order to understand the False Positives and False Negatives detection. It should be useful evaluate the model in its computation speed: time for the training process and for the prediction.

Project Design

In order to solve this problem, it is necessary to apply some unsupervised and supervised Machine Learning techniques.



First, data will be studied, in order to explore features variance and patterns on the data. By this way, it may be possible to identify useless or outliers features, and then, as a consequence, applying feature reduction. While conducting this data exploration, it is possible to understand if features such as gender and age are relevant or not. Moreover, a hierarchical clustering should be able to find eventual outlier students. If the clustering process detects a distance from one or a few students respect to the others: they should be considered as outliers, it is necessary to delete them. In order to do this, data will be scaled, and eventually imputed. Furthermore, categorical features will be one-hot encoded.

After data exploration and pre-processing, a supervised learning model has to be trained and tested. In order to do this, three models will be tuned with different hyperparameters (using grid search):

- Decision Trees

  Decision Trees model is used for classification problems. This algorithm is able to analyse data features and to understand how much information each of these fetaures can give to us, in order to make our prediction. The model can make really accurate predictions, because of each feature can be combined with every other feature according to its specific information gain. This type of model has also a bad behavior. It tends to overfit a lot, thus, we have to use it with high attention or use it with any ensemble method.

- Ensemble Methods

  Ensemble methods are able to solve classification problems combining different classification models. It can be used for real world applications, especially when data is not linearly separable. We can make a prediction using different models combined with an ensemble method, such as Ada Boost or Bagging or Random Forest. Ensemble methods perform very well when data is not linearly separable. Furthermore, they are able to create models fitting data well because of their ability of finding a good compromise between bias and variance.

- Support Vector Machine
  - Support Vector Machine is able to make good classification predictions, defining models with a well-fitted boundary. This boundary is so good because of the use of the margin. This

model has a good flexibility because of its parameters. This alorithm is able to work on different types of data, in fact, it has three kernels: linear, polynomial and Rbf. Moreover, an SVM model can give us the ability of tuning it in order to decide how much the model has to be precise: with the C parameter we can define how much weight we want to assign to the Classification error respect to the margin error.

The best model will be chosen, evaluating:

- Testing accuracy;
- Training and Testing Loss, Learning Curves, Fscore (we don't want to overfit or underfit);
- Execution time.