EEG signal Based Eye State Classification using ML algorithms

Anshuman Pillai, CB.EN.P2DSC21003

Class (I-M.Tech-DS) Batch (2021-23)

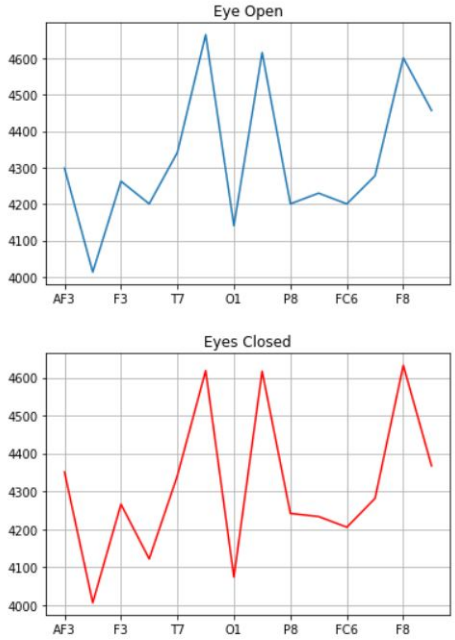
Course: 21DS602 (21-22(Odd)) Date : 31st January 2022

Abstract

Whenever a human body performs any actions, like moving hands or opening of eyes or closing it, the neurons inside the brain are sparked which generates current, with the help of electrodes on the scalp and then measuring the current and amplifying the current, an EEG signal will be created. For every action, there is a different EEG signal. So if we could classify the signals with the help of Machine Learning models, we can be able to create machines that could tell whether the person is awake or sleeping or what the person is doing. But for this particular case, we will only be focusing on the status of the eye whether is it open or closed.

I. Introduction

The EEG signals which contain the electrodes in a time series format tell us how the brain reacts to different conditions, so we will try to classify the status of eyes using various Machine Learning models. The given dataset contains 14 features as electrodes and the last feature as the class label or the target set. There is a total of 14980 instances in the data. The given data is :



II. Literature

EEG (Electro Encephalo Gram) signals are the signals which are used to check the sensitivity of brain or to check how different the neurons inside the person's brain will be reacting when a particular part of the body like eyes or hands are working. It is done by taking electrodes from the scalp which is further amplified and could tell us how the brain is reacting to different actions.

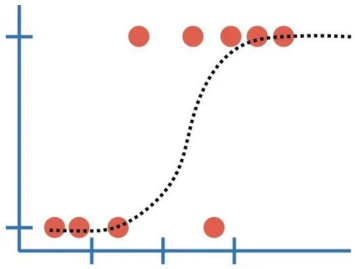
III. Objectives

The main aim behind this is to do a comparative analysis of different Machine Learning Models to correctly classify the status of the eyes from the given EEG Signals

IV. Theoretical Background

Here I will be using 4 models:

1) Logistic Regression: It is a supervised Machine learning Algorithm that is used for classification problems. It uses a logistic function for binary classification.



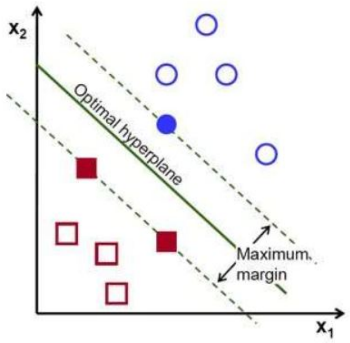
Its mathematical formulation is:

$$S(z) = \frac{1}{1 - e^{-z}} where \tag{1}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2}$$

2) Support Vector Machine: It is also a machine learning algorithm that is used for classification. In this algorithm we try to find a hyper plane where the data points will be linearly separable, there are different kinds of kernels like linear, poly. I went with the polynomial kernel with degree 5 as this was giving a better accuracy as compared to other kernels.The main formulae of SVM where it separates the two different classes is :

$$w^T x - \gamma = 0 \tag{3}$$



It can be further classified for two parts ,one is with hard margin(With No misclassification) and the other with soft margin(with minimal misclassification).

Mathematical Formulation for hard margin is :

$$\min_{w,\gamma} \frac{1}{2} w^T w \tag{4}$$

s.t

$$d_i[w^T - \gamma] \geq 1 \tag{5}$$

Its mathematical formulation for soft margin will be :

$$\min_{w,\gamma,\xi} \frac{1}{2} w^t w + c \sum_i \xi_i \tag{6}$$

s.t

$$d_i[w^T \phi(x) - \gamma] + \xi_i - 1 \geq 0 \tag{7}$$

$$\xi \geq 0$$

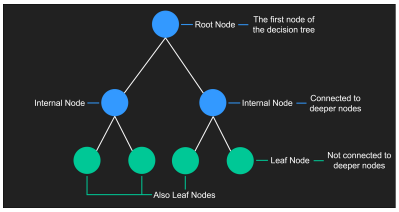
ϕ will be used for mapping the features to higher dimensions.

When we will be considering for Linear Kernel, then the formulation will be same as given above but when we will be

$$K(X_1, X_2) = (a + X_1^T X_2)^b \tag{8}$$

Here b will be the degree that it will be mapping which for our case will be 5.

3) Decision Tree: It is a supervised Machine Learning algorithm that predicts the class labels based on simple decision rules. It consists of two parts main parts one is the root and the other is the node. This works on finding the impure nodes and splitting them in parts to get a pure node.



The main criteria we will be using for splitting will be Entropy which can be formulated as :

$$Entropy = -\sum_{j} P(j|t) log_2 P(j|t) \tag{9}$$

Where P(j—t) is the relative frequency of class j at node t

4) Random Forest: Random Forest is another supervised machine algorithm that uses a base of decision tree and it uses the concept of bagging which splits the training dataset into random samples of k folds and then does apply decision trees to each sample and take its score and since this is a classification problem after this is done, by maximum voting we choose the score for Random Forest model

V. Methodology

A. Pre-Processing Data Set

In this step, will be differentiating the class (target variable) from the dataset and the dataset without the target will be one dataset and the target one will be another one.Then We will be splitting out the dataset in 75-25 split and storing it as xtrain, xtest, ytrain and ytest

B. Training on Models

I will now first normalize the dataset and then would be training it on different Machine Learning models which are Logistic Regression, Support Vector Machine(with C=1.5, kernel as polynomial and its degree of 5), Decision Tree(with criteria as Entropy), Random Forest

C. Comparing the Results

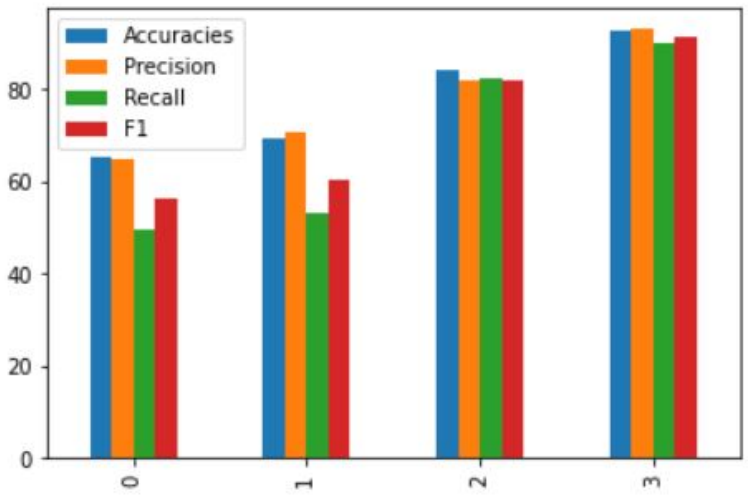
Once the dataset is trained on the respective models, I will be predicting and will be finding out the accuracy, precision and F1 scores for the respective models

VI. Experimental Results and Discussion

The Results obtained are:

Serial No	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.6553	0.6466	0.4973	0.5672
1	Support Vector Machine	0.6921	0.7049	0.5303	0.6053
2	Decision Trees	0.8395	0.8184	0.8218	0.8201
3	Random Forest	0.9255	0.9300	0.9004	0.91.55

The data can be represented through a graph as:(0,1,2,3 are the model Numbers)



For different models, we had calculated the confusion matrix and calculated their respective accuracy, F1 Score, Precision and recall

VII. Conclusion

For the given dataset, I firstly normalized the data and then split it in the 75-25 split and then trained it on Logistic Regression, Support Vector Machines, Decision Trees and Random Forest. When the trained models were evaluated, it showed that the Random Forest model was the most efficient model which gave an accuracy of 92.55 per cent

VIII. References

1) Sanei, Saeid, and Jonathon A. Chambers. EEG signal processing. John Wiley Sons, 2013.
2) Tagliaferri L., Morales M., Birkbeck E., Wan A., (2019). Python Machine Learning Projects. : DigitalOcean.
3) Analytics Vidhya - Learn Machine learning, artificial intelligence, business analytics, data science, big data, data visualizations tools and techniques. — Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog
4) GeeksforGeeks — A computer science portal for geeks. Retrieved from https://www.geeksforgeeks.org/