Solim LeGris
November 7, 2020

MAIS 202 - PROJECT DELIVERABLE 2

Problem Statement

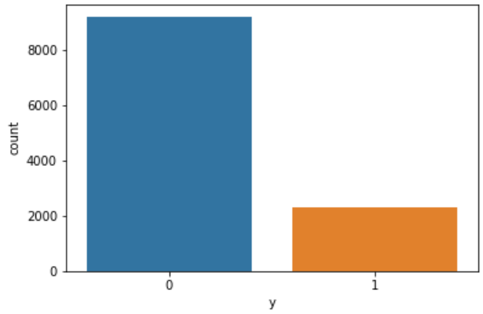
My project consists in training a binary classification model for EEG data on epileptic patients. This can be useful in the medical and research settings. The model will be able to predict from a given patient's EEG times-series whether that patient is experiencing an epileptic seizure. There are different kinds of epileptic seizures ranging from generalized petit-mal seizures to more debilitating grand-mal seizures.

Data Preprocessing

As I decided in D1, I will be using the Epileptic Seizure Recognition Data Set from the UCI Machine Learning repository[1]. The dataset contains a wealth of samples collected from 500 individuals. Each individual has 4097 data points for 23.5 seconds which were divided in 23 chunks of 1s with 178 data points for each subject. The dataset has 11500x178 data points divided among 5 possible labels. Given that the data is mostly preprocessed, not much more preprocessing was needed. I removed the first column because the information in it was not useful and I checked for NaN data points. Fortunately, there were none. Lastly, I standardized the dataset to facilitate model training using Scikit-Learn's StandardScaler. Finally, I split my dataset in training set and test set using an 80%/20% proportion.

Machine Learning

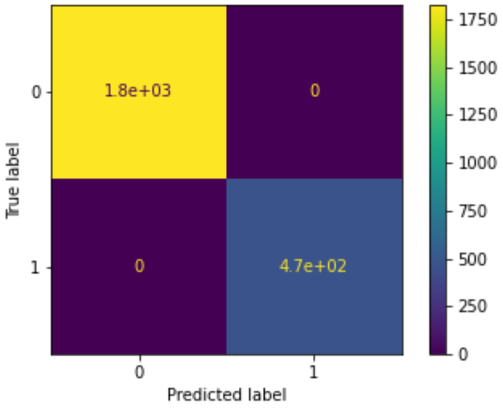
For my project, I chose a LogisticRegression model. I used Scikit-Learn's implementation of LogisticRegression. I first trained my model with the default hyper-parameters and did not use any regularization technique. When testing my model on the test set, I obtained a perfect score of 1.0. I am not sure if the model is overfitting the data given that this is the test set. I also considered that my dataset has proportionally much more datapoints that are categorized as non-epileptic which may have influenced the results of my model's predictions.



I am still looking into how and why I am getting perfect results and considering finding additional data to test the model on.

Preliminary Results

Given my results, the confusion matrix was not very useful. Nonetheless, I plotted it and included it here.



Next Steps

Even though I am getting extremely good results, I want to test a SVM and potentially a neural network on this dataset to compare results.

[1] Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., & Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, *64*(6), 061907.