Solim LeGris
November 6, 2020

MAIS 202 - PROJECT DELIVERABLE 1

Dataset Selection

For my dataset, I have decided to use the Epileptic Seizure Recognition Data Set from UCI. I chose this dataset because of my interest in neuroscience and the fact that it is pre-processed which will allow me to focus on finding the optimal model for this problem. Each subject's brain activity represents 23.6 seconds of EEG recording sampled from 4097 data points. The data is structured in such a way that one instance represents 178 data points for 1 second with a corresponding label.

Methodology

Data Preprocessing

Although the dataset is preprocessed, I am choosing to reduce it to a binary classification. Each instance is labelled with some $y \in \{1,2,3,4,5\}$ where only $y = 5$ corresponds to subjects which showed epileptic brain activity on the EEG recording. Therefore, I will classify $y \in \{1,2,3,4\}$ as non-epileptic brain activity.

Machine Learning Model

For this data, I wish to predict non-epileptic activity and epileptic activity for the EEG time series of the subject recordings. I will use a logistic regression model. The spiking activity of the EEG is what defines whether or not the activity is considered epileptic or non-epileptic. Linear regression is a simple model to implement and may yield good results. If that does not work, I am considering using a Random Forest classifier.

Evaluation Metric

For my evaluation metrics, I will be using a confusion matrix and logistic loss.

Final Conceptualization

I will be presenting my project in a simple landing webpage as required.