Ryan Burden
Chloe Easterbrook
Abigail Han
Jaikishan Mali
Tatum Maston
Jacob Schenberg

COGS 189 Report

Introduction and Motivation

We were interested in predicting whether a subject is an alcoholic or not based off of this dataset: https://archive.ics.uci.edu/ml/datasets/eeg+database. The main problem is that alcoholism is currently and has been a prevalent issue for people due to not only detrimental health effects but also interfering with a person's ability to work, form and maintain relationships, and live an overall healthy and happy life. Naturally, we want to do more research into how and why people are more predisposed to becoming alcoholics so that we can mitigate this problem and create beneficial solutions for future generations. This data arises from a large study to examine EEG correlates of genetic predisposition to alcoholism which we are going to use to create a model which predicts whether or not a subject is an alcoholic.

Related work

1. Mention any related work from our class readings or from other literature.

In the Brain Computer Interfaces review we looked at throughout the course, there were a myriad of topics explored in the paper that correspond with techniques or methods utilized in our project such as EEG's and their employment in a study, principal component analyses and variance, and regression algorithms. The data we sourced was produced from a larger EEG study where 64 electrodes were situated on a subject's scalp which were sampled at 256 Hz for 1 second and were positioned using the Standard Electrode Position Nomenclature. We applied a Principal Component Analysis (PCA), an extraction method that creates principal components by converting possibly correlated observations into a set of uncorrelated variables. We also generated a plot to visualize variance by percentage which is elucidated by the principal components. A logistic regression was also performed where EEG signals are extracted as independent variables in order to predict user intentions (Fernando et. al, 1239).

In our search to understand, in depth, the role of our data, it was imperative to apply research from outside class. Firstly, it was important to our group to see if there was any literature that supported our project idea. Since we wanted to explore if there was any correlation between EEG readings and a genetic predisposition to alcoholism, these ideas defined our literature research. We were able to read a paper from Taiyuan University of Technology titled *Disturbed Connectivity of EEG Functional Networks in Alcoholism: A Graph-Theoretic Analysis*, by Cao, et al. This study, much like from our dataset, applied scalp electrodes to 28 alcoholic and 28 healthy participants, and then presented them with a stimulus to record and compare brain function. They found that in comparison to the healthy volunteers, the alcoholics had relatively smaller **C** values, which indicated less linkages in the neural networks of the alcoholics. Whereas the paper we referenced wanted to see the differences in working memory between alcoholic and healthy participants, our aim was to use EEG data to detect a genetic inclination towards alcoholism.

In addition, the research article "Alcoholic EEG signal classification with Correlation Dimension based distance metrics approach and Modified Adaboost classification," is related to our project because of their motive for conducting research regarding alcoholism in relation to EEGs. The researchers in this study identify Computer Aided Diagnosis (CAD) Techniques to identify the normal versus alcoholic EEG signals has been of great help to the medical field. The paper goes over the importance of conducting the study and cites reasons of importance: "A major reason for worldwide mortality rates every year is due to alcoholism and its related disorders… Alcoholism even contributes to deaths caused by accident, homicide, suicide, depression etc [5]. The social and personal relationships at colleges, offices, social gatherings, meetings etc is fully spoilt by alcoholism."

Methods:

Our dataset has 256 sensors for each channel in a 64 channel EEG. We are trying to predict the presence of alcoholism in a subject on a trial-by-trial basis based on the EEG data from the 256 sensors.

1. First, the data was split into a training and testing set, with 33% of the dataset being reserved for testing.
2. We then cleaned and manipulated the dataset into a workable dataframe.
    a. The data was given in RAR files, which had to be unzipped through unarchiver, and were placed into R files. Each five rows of the file was information about the trial that was run, so we had to run a loop through all the files to delete the first five rows. We needed to create a new column in the dataframe that stored important data about the participants and stimulus from the first five rows.
    b. The sensor numbers were all in one column, so 264 sensors total per trial. We transformed the sensor number column into up to 264 columns, and each of those columns had a corresponding sensor value.
3. The first method we implemented was a logistic regression.

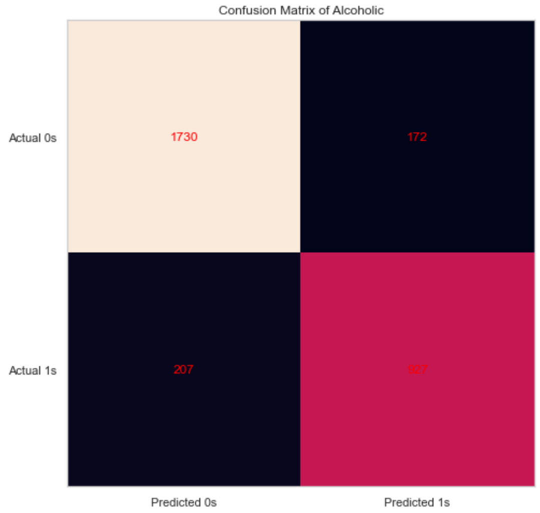
   a. The features that we trained on include: the sensor position and the sensors that were recorded in that position, and which stimuli were shown. The alcoholic column was what we tried to predict.
   b. This initially gave us a pretty high score of 0.87.
4. We then used principal component analysis in order to reduce the dimensionality of the data.
   a. We found plotted the proportion of variance explained by various principal components and found that variance reaches a minimum asymptote around 10 principal components. We experimented with 1-10 components.
5. We then trained a ridge regression classifier to our PCA-reduced training dataset with two principal components to find our highest accuracy with this set.

Results:

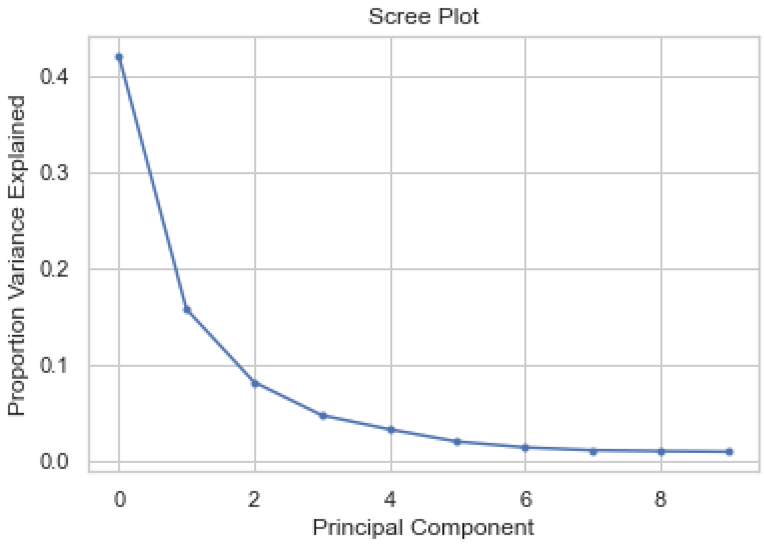
By first doing a logistic regression to our cleaned data frame, we were able to produce a classification accuracy of ~87.5%. . We found that the first two principal components account for about half of the total variance in the data. This reduced the dimensionality of our dataset from 256-d to 2-d. When fit and evaluated using the PCA-reduced test set, the ridge classifier had an accuracy of 63.97%.

We discovered that logistic regression worked very well initially on this data. Because of the high dimensionality of our dataset, we assumed that using PCA to focus on the most important information / trends in our dataset would have helped us get to a higher accuracy. However this was not the case. After experimenting with various amounts of components, we found that using two components returned the highest accuracy, which is surprising as this is a very, very low dimension as opposed to what we began with. However, this accuracy did not even compare to the accuracy we got from just the original, high dimensional dataset.

Here is the confusion matrix for the logistic regression:



6.



Discussion:

This project provided valuable insight on issues that can potentially arise while working with data. Most notably, our team had to do a complete pivot from our intended project due to the fact that the data that we first encountered was extremely difficult to work with. Because of this, it took a while for us to find data that we could utilize with our skills and available tools. Thankfully, the UCI Machine Learning Repository provided us with lots of data to choose from. One of the valuable lessons we learned was to overestimate the time needed to clean & manipulate data into a usable data frame for analysis. At first glance, it may seem that it will go by quickly, but there are a lot of issues that can arise, namely the fact that our dataset only came in RAR files that can only be used on windows, and our team only has Macbooks. So, after some research we found an extension that allowed us to convert the data into files we could use, which took up more time than anticipated. Considering the fact that we were able to train the model to classify the two groups indicates that there is an underlying pattern between alcoholics and healthy participants that shows they are different. Though we know that the EEG data between the two are different, we haven't done enough experimentation or analysis to draw any specific conclusions on genetic predispositions to alcoholism.

Improvement 1: If we had the luxury of more time, our team would like to collect our own data to compare with this data frame. We would intentionally recreate the study in order to further validify the results and analysis. This way, we would be able to again be able to show the differences in EEG activity between the experimental and control groups.

Justification 2: Another improvement we could make is to use the larger dataset provided from the UCI Lab rather than the smaller one. We used the smaller one for this project, however a larger dataset would provide more accurate results based on the fact that there are more observations to draw a conclusion from.

References

Cao, Rui et al. 'Disturbed Connectivity of EEG Functional Networks in Alcoholism: A Graph-Theoretic Analysis'. 1 Jan. 2014 : 2927 – 2936.

Prabhakar, Sunil Kumar, and Harikumar Rajaguru. "Alcoholic EEG Signal Classification with Correlation Dimension Based Distance Metrics Approach and Modified Adaboost Classification." *Heliyon*, vol. 6, no. 12, 2020, https://doi.org/10.1016/j.heliyon.2020.e05689.

Nicolas-Alonso, L. F., & Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors (Basel, Switzerland)*, *12*(2), 1211–1279. https://doi.org/10.3390/s120201211