Mais 202 Project Deliverable #3

Cole Killian

November 09 2019

Problem Statement

Since the last deliverable I have been working on training a neural network to predict someone's age based on their epigenome.

Data preprocessing

In this experiment I was working with 732 blood samples from a study titled "Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan". Each of these samples has 485512 features. Initially, I replaced null feature values with 0.5 (because that is the average value), and then removed samples that didn't have their ages marked (there were three such cases which brought the number of samples I was working with down to 729). I then used a chi function to bring the number of features down to 50 from 485512. These remaining 50 are the statistically most significant features with respect to classifying a samples age. After importing the age annotations, I was ready to proceed.

Machine Learning Model

My chosen model is dimension reduction via a chi significance function, and regression via a neural network.

I chose to perform a chi significance function before hand because working with 485512 features makes it too easy for the neural network to overfit. I then used a neural network to predict the samples age because I figured that this deep learning technique would prove more effective than other more elementary machine learning methods.

Results

After training the neural network PCA, it was able to beat random guessing. It was able to predict the validation samples ages within 25 years with 69% accuracy. This may not seem very impressive, but there are some who doubt that it is even possible to detect age from blood methylation data. These findings suggest that it is indeed possible. I was limited to 729 samples for training. Revisiting this process with more samples could yield a very powerful regression algorithm.

See graphs on following pages. Red points correspond to model predictions while blue points correspond to real data.

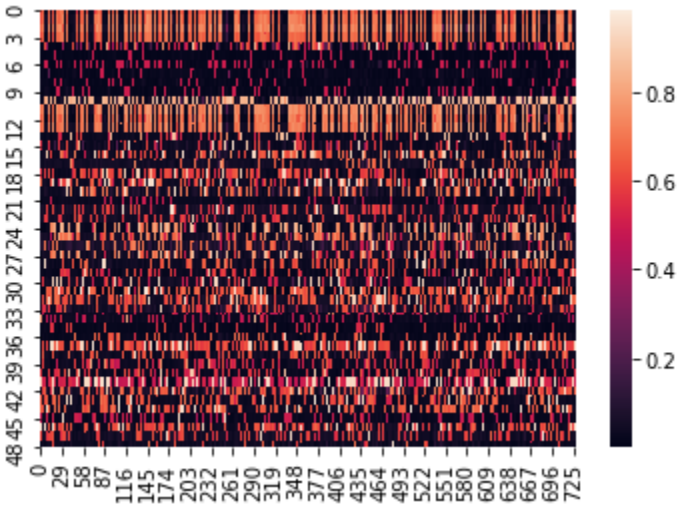

Figure 1: Heat map that helps a human to understand what the neural network is looking at.

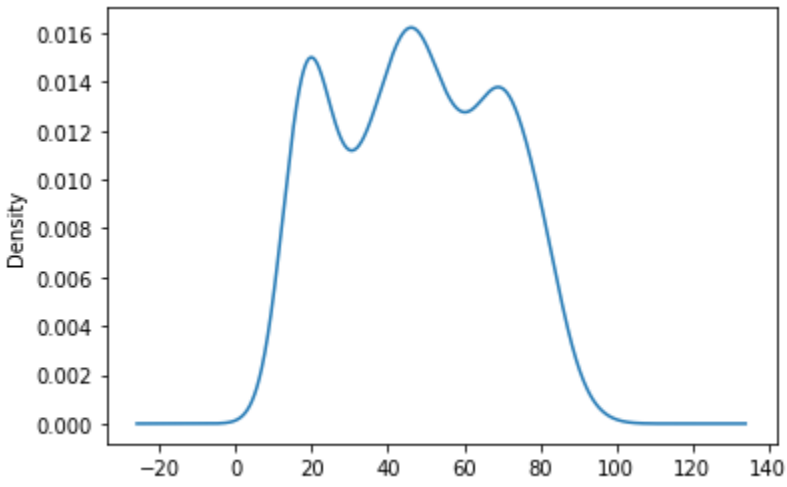

Figure 2: Kernel density estimate plot for visualizing the age distribution.

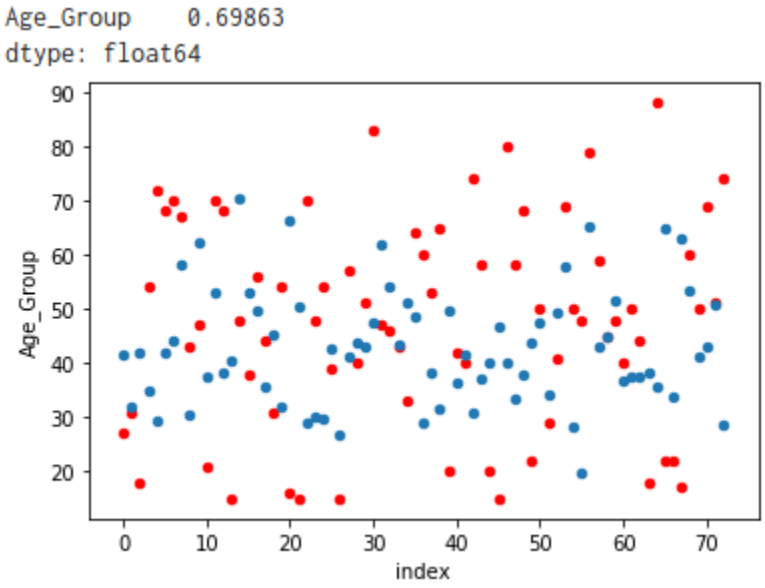

Figure 3: Scatter plot showing predicted vs. real ages. Red points correspond to model predictions while blue points correspond to real data.

Final Demonstration Proposal

I will build a landing page type website to demo the model and results. It is difficult to make this project interactive because there isn't anything to visualize that an untrained human could internalize easily. In order to try and overcome this I will display heat maps that allow a human to see what the machine learning model is looking at. Then I will allow the user to select various samples, view it's heatmap, view the samples age, and view the predicted age.

As far as technologies go, I will be making the page using the materialize css framework as I have used it in the past. For the interactive section I will use a Flask backend. My current idea is to use plotly to display a scatter plot. The user can hover over different points to get their id numbers. Then there will be a text box where the user can enter one of the id numbers. After doing so, the user will be shown that samples heatmap column, the predicted age of the sample, and the real age of the sample. As an additional visualization, I will show a 3d PCA plot with points colored on a spectrum based on age.