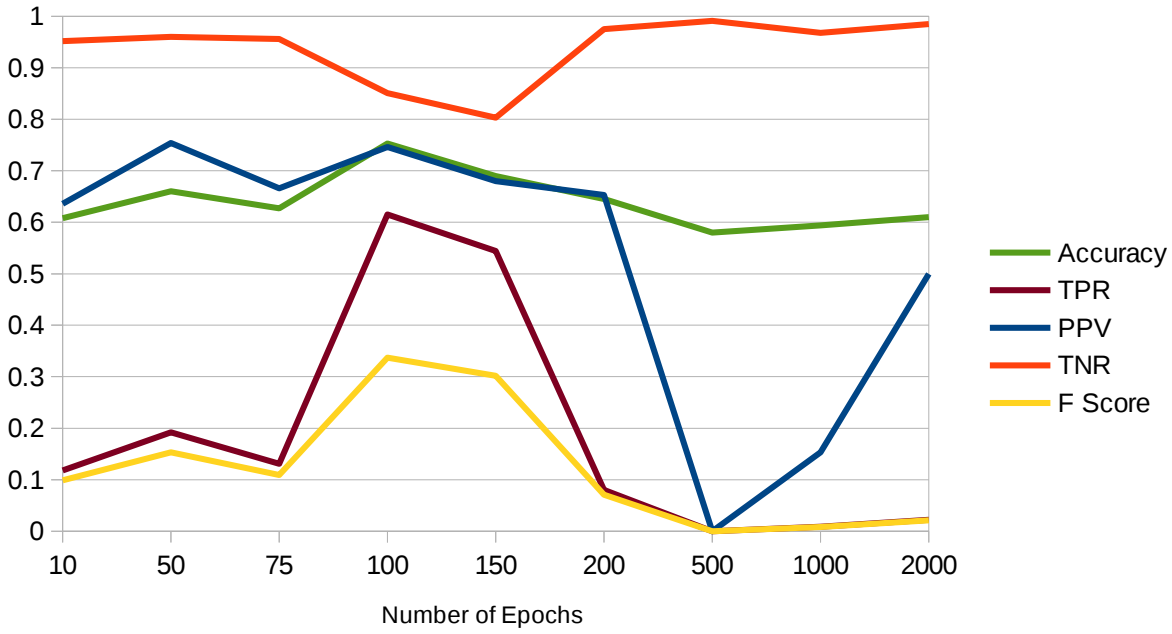
Jacob Vargo

CS 425

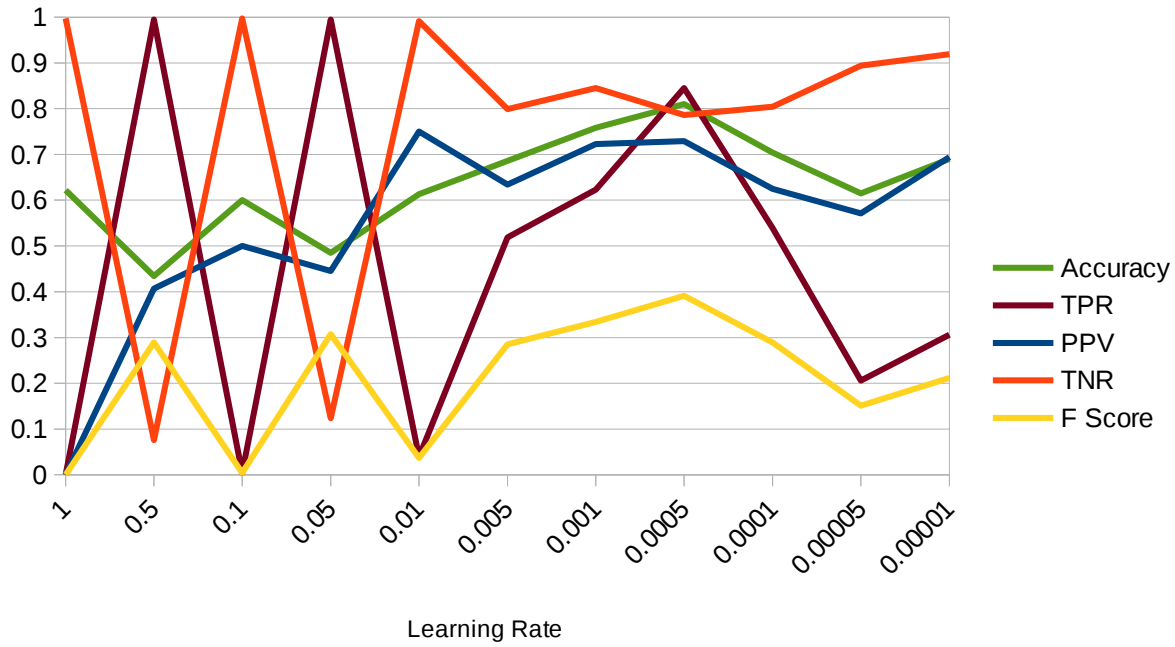
Project 4

When first viewing the raw data, I noticed that it appeared to be sorted with spam classified e-mails first. This meant that I needed to shuffle the data before splitting it into the, test, validate, and training sets and then saving the sets for repeated use later. I split the data such that half of the data is in training, one eighth for validation, and three eighths for testing. There are no missing values in the original data; so I did not need to handle these.

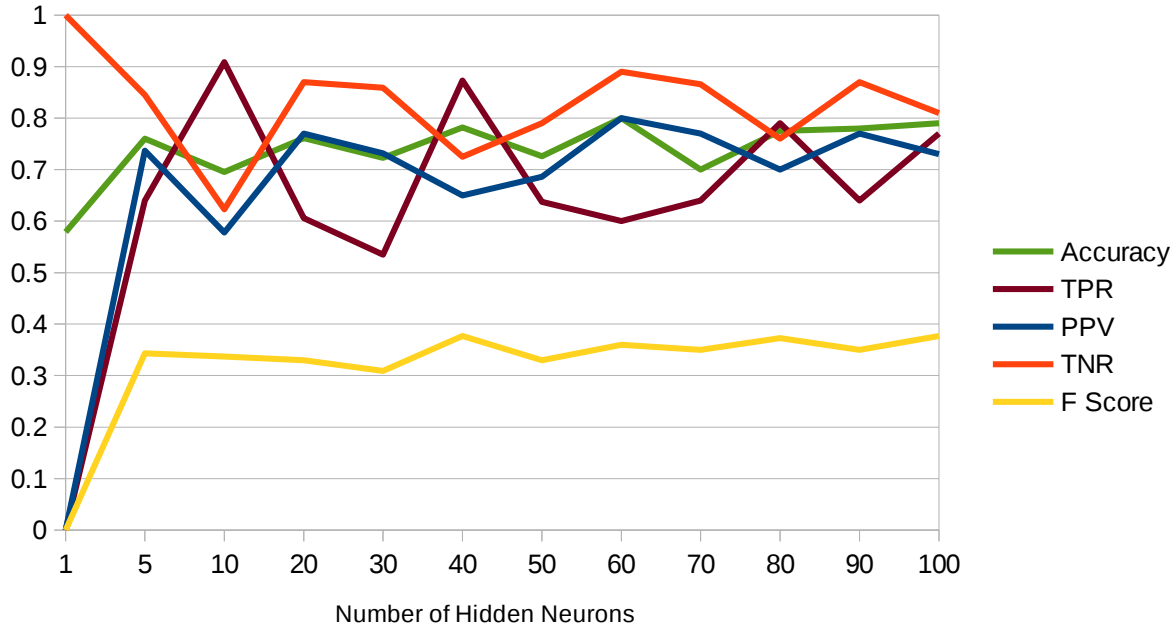
I initialized the weights and biases with a normal distribution and a standard deviation of 0.5. To begin testing, I started with a learning rate of 0.01 and 40 hidden neurons in 1 hidden layer. I then tested values for the epoch to find which value had the best results.




With the results of my epoch testing, I decided to continue forward using 100 epochs. I chose 100 epochs because it had the greatest accuracy and F Score. I then tested the learning rate.




With the results of my learning rate testing, I decided to continue forward using a learning rate of 0.0005. I chose 0.0005 because it had the best accuracy and F Score. I then test the number of neurons to use.



After testing the number of neurons to use, I decided to continue forward using 40. I chose 40 neurons because it had comparable results to the best results seen at 100 neurons. Also, I decided it would likely make for a more robust network to use a number of neurons that is less than the number of features in the data set.

When I did a final tested of the neural network with the test data, my results were: Accuracy = 0.787, TPR = 0.741, PPV = 0.718, TNR = 0.816, F Score = 0.364.