Josh Brown

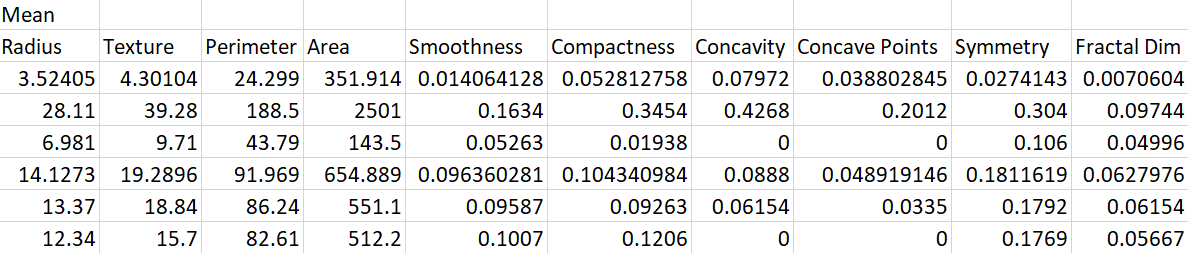
Dr. Duan

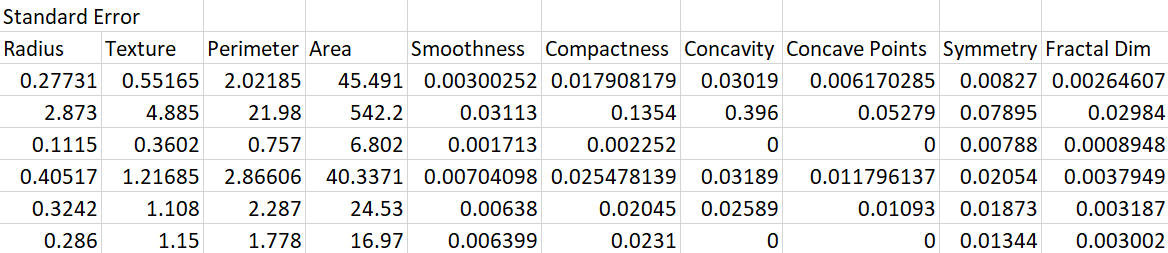
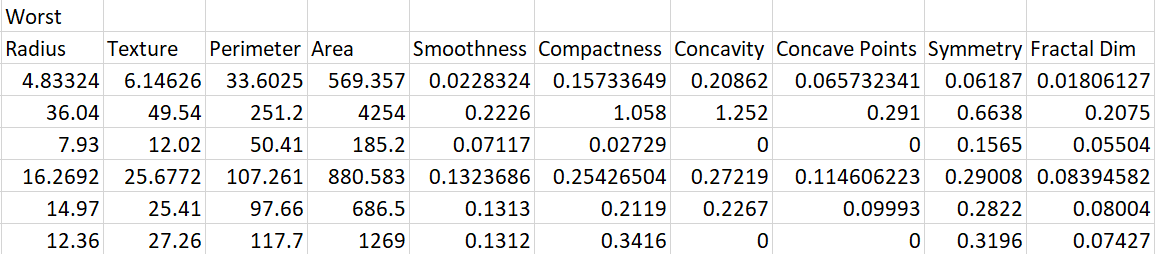
Applied ML

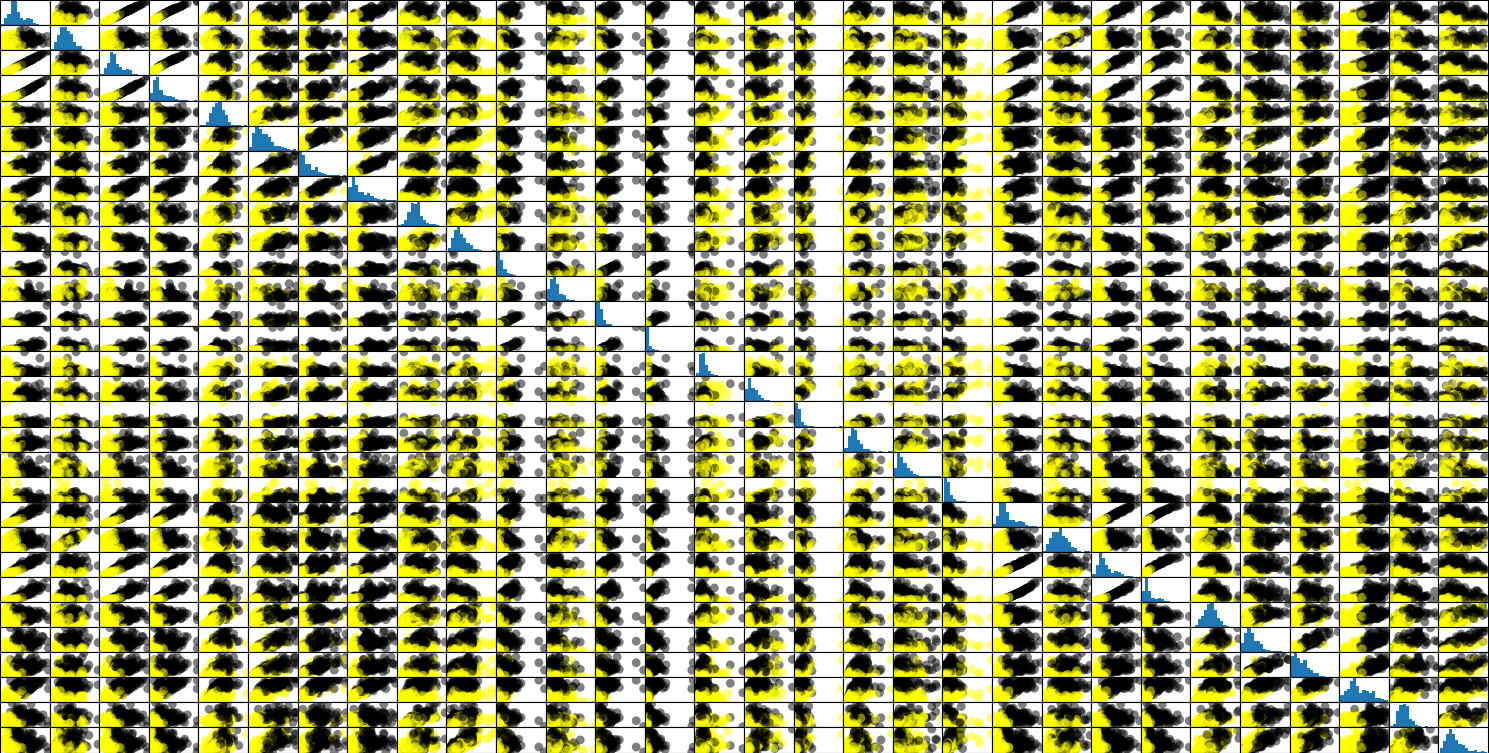
4/14/21

Neural Network for Breast Cancer

In this paper I will explain the methods used to achieve an extremely accurate neural network model that can predict breast cancer with very high efficiency. We will explore the different parameters tested with the model, as well as what models and model trends fit the data the best. We will also see how scaling the data affected the results, along with a general statistical overview of the raw data.

First, a summary of the data:

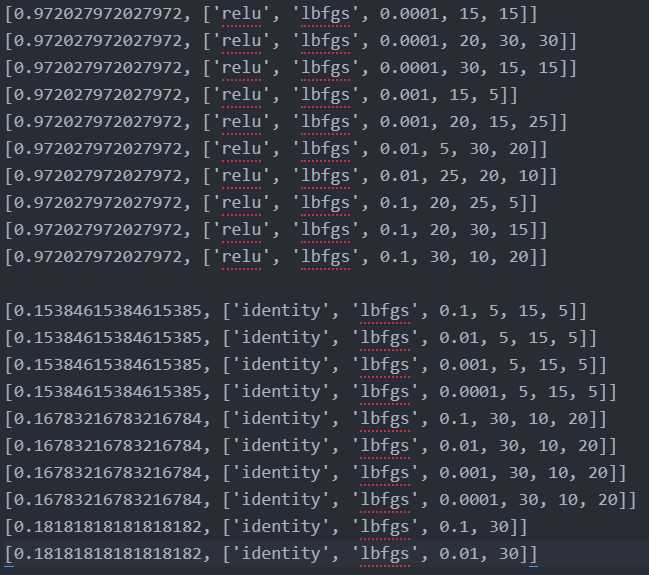


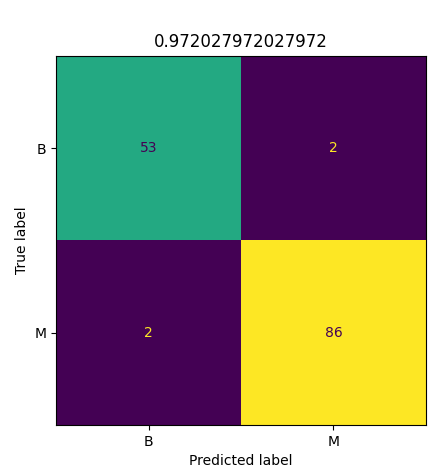
As we can see, even without scaling, the data is very close in size. The only outlier being Perimeter, having values at 100. Other than that however, the data is very stable. Later on, we will see this lines up with our results, as we will find that scaling the data does not have that large of an impact on the overall result of the fit.

Above, we can see that the raw data is very clumped. There are distinct yellow clumps and distinct black clumps in the majority of the data. We can also see from the histograms that a lot of the data fits some sort of bell curve, this is not that important but an interesting observation. With the data being so clumped, this means we can expect our results of the neural network to be fairly accurate, which they indeed are.

Next, we will talk about the different architectures of neural networks used. In my code, I use deeply nested for loops, combined with list appends to iterate over and keep track of each neural net. Since my program examines over 20k different combinations of parameters, it takes some time to run. When it is done, it outputs the top 10 for scaled and unscaled fits into the file Scores.txt. The different combinations used were as follows: four different activations (Logistic, relu, tanh, identity), three different descents (lbfgs, sgd, adam), four different alphas (.0001, .001, .01, .1) each with 6 different combinations of neurons, in up to 3 layers. This makes for a total of 10k combinations tested. This is then repeated on unscaled data, doubling the amount of combinations to over 20k.

After collecting all the data, we can now examine the worst and best fits for the data. The best neural networks used relu activation exclusively, for scaled and non scaled, as well as all using LBFGS for gradient, scaled and non scaled. Alpha values varied, with the smaller alpha values being preferred. Surprisingly, the top ten results for scaled and unscaled data were the exact same. The only difference was hidden layers and alpha values. Also interesting, is that all 10 had the same accuracy at 97.2%, which is extremely promising. This tells us that the network is extremely confident in predicting the data. The worst models all used identity activation, and LBFGS for gradient. The 10 worst fits also were the same regardless of scale. Here is a snippet of some of the best and worst fits; First is the accuracy, followed by a list of the parameters in the order of activation, decent solver, alpha, and the last 2 - 3 are j neurons in i-th layer.

Here, we can see that the accuracy is extremely high for the top 10, and very low for the bottom 10. However, we also see that we are much more accurate than inaccurate on a worst / best case basis. In other words, we are only 3% away from perfect prediction, but more than 15% away from complete failure.

The following is the confusion matrix for each of the fits; since each model was so accurate, it seems they were each tripped up by the same 4 outliers. This makes a lot of sense in terms of how the algorithms work. An outlier will most likely trip up a neural network no matter the parameters. Again, This confusion matrix is the matrix for ALL 10, They are the exact same!

In conclusion, these neural networks were able to fit the data extremely well with very high accuracy. Our confusion matrices tell us that each neural net was tripped by the same outliers, and that overall they are extremely accurate. I think with more data, these models could be extremely useful in the medical field because of how precise they are. If it is this accurate on breast cancer, maybe we could use this on Covid symptoms, and achieve similar results. Being able to predict these kinds of diseases and illnesses, with this level of accuracy has the power to save many people's lives. In the end, just by testing different neural architectures, it's possible to find extremely well fitting results.