Tailer Nguyen

1792311

4/1/2022

Problem Set 2 Task 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type of classification algorithm | variation | Accuracy rate Train Set | Accuracy Rate Test Set | Cross Eval / MAE  Train Set | Cross Eval MAE / Test Set |
| SVM | Linear | 70.2970297% | 70.2970297% | 0.0384615386 | 0.20000004 |
| SVM | Radial | 86.13861386% | 79.5454545% | 0.24818181818 | 0.205 |
| MLP | Relu | 80.15019% | 85.9920% | 0.025106 | 0.2041345 |
| MLP | Sigmoid | 82.059391% | 77.198385% | 0.0384615 | 0.20000 |

From the experimental results, we can see that radial performed the best in terms of accuracy rate. However, this could be because some grades have more observations than other grades, and when datasets are imbalanced, SVMs struggle with having best accuracy.

It could have easily just have been a lot lower than normal for radial accuracy. However, with neural network, a convolutional method which transforms the summed weight input from nodes into activations for output for that input, the accuracies reflect, on average, to be more consistent and more efficient than SVM’s.

Some methods I used to try to increase the accuracy of the neural networks (because they were higher accuracy in general, too) was to play around with the batch size and the number of epochs, change the activation function, and normalize the data even more.

When I am learning about the batch size and number of epochs, I learned that the batch size is the gradient descent that initializes training samples until to work through before the internal model is finalized, so it is constantly being updated and it cannot be more than the number of samples in the network. When I test with 130 samples, I get increasingly more consistent accuracies around the mean average of the set, but the accuracies do not get more higher, so the batch size is more noisy which could actually lead to more inaccuracy. When I change the epoch size, I found that the perfect epoch number is 11, and at 12 will lead to overfitting, which is an inaccuracy from the line of best fit.

When I changing the activation function, I tested RelU, sigmoid, and tanh, but the other relu and sigmoid gives me best values even when testing with epoch as 11.

At first, I was getting very low values for svm because I was using svm functions and numpy. However, when I start to use preprocessing to remove student id and not include course id to not include the categorical variables, then I noticed that my accuracy got higher.

In my findings of this task, I learned that data must be collected, processed, trained, evaluated, tested. However, this is not the last case. Scientists see good accuracies with machine learning but always strive to improve the data. They do this by parameter tuning, making more predictions, getting rid of unnecessary variables, and starting back at step 1 just to get their data to look cleaner.

I learned the difference between SVM and MLP. Neural network is faster because there is a clear decision factor between each node in neural network rather than in SVM, it take all.