ADVANCED MACHINE LEARNING

ASSIGNMENT2 – NEURAL NETWORKS

SUMMARY REPORT

Overview:

The IMDb review dataset consists of 50,000 movie reviews, of which 25,000 are categorized as "positive" or "good," and the remaining 25,000 as "negative." This study looks at various ways to improve the performance of a neural network model using the IMDb dataset.

An existing neural network model may be modified in several ways, such as by changing the activation function, loss function, units, number of hidden layers, and regularization methods like dropout. The results are then analyzed.

Data Processing:

To convert the IMDb reviewer dataset's raw text data into a format suitable for neural network training, a few preparatory tasks had to be completed. We selected only the top 10,000 phrases from the dataset because analyzing every word would yield a high-dimensional input space. After that, we transformed the text reviews using integer representations by using a dictionary to shift the meanings in the top 10,000-word list to the appropriate indices. Prior to using neural networks, the integer approximations must be converted into tensors. We cut the longer reviews down to zero and added zeros to the shorter ones to ensure that all the assessments had the same length. To that end, every review was a fixed-size vector with an index of a word for each element.

Methodologies:

When the data is integrated, we establish a maximum word count and review duration for each review. We then built a basic neural network model with a single 16-unit hidden layer. The optimizer was Adam; the triggering rates were relu and tanh; the loss parameter was binary Cross entropy; the optimization was MSE; and the hidden layer parameters were dropout and hyper tuned. Subsequently, we investigated the previously suggested methods to raise the model's effectiveness. Subsequently, we altered the overall quantity of concealed layers to generate prototypes featuring one, two, and three invisible hidden layers. We compared, evaluated, and improved the models using the test and instruction datasets. Our findings demonstrate that, in comparison to using only one hidden layer, the addition of three enhanced test validity and accuracy.

Accuracy Percentage and Hidden Layers:

1. 1-hidden layer, 16-units Accuracy = 88.6%
2. 3-hidden layer, 16-units Accuracy = 88.4%
3. 3-hidden layer, 32-units Accuracy = 86.3%
4. 2-hidden layer, 64-units Accuracy = 86.1%
5. 3-hidden layer, 128-units Accuracy = 87.3%
6. 3-hidden layer, 16-units Accuracy = 88.1%
7. 1-hidden layer, 16-units Accuracy = 88.4%
8. 3-hidden layer, 16-units Accuracy = 88.1%
9. 2-hidden layer, 16-units Accuracy = 88.7%
10. 3-hidden layer, 16-units Accuracy = 88.1%
11. 3-hidden layer, 32-units Accuracy = 86.3%

In summary:

To avoid overfitting, we then attempted dropout regularization. We created a unique model with dropout layers utilizing training and test datasets. Our findings demonstrated that regularized dropouts decreased validation precision in contrast to the baseline model. Variable neural network modeling alterations are thought to have different accuracy values and loss functions as a result. Three thick layers with a drop-out rate of 0.5 can yield better results for the IMDB information set than the Model Hyper, which had the best accuracy and loss. When binary cross entropy was considered, the mean square error (MSE) loss function displayed the lowest loss value. Ignoring the vanishing gradient problem due to the tanh activator function's decreasing accuracy. It has been demonstrated that the Adam optimized function is efficiently utilized in the model construction process. Compared to Model Hyper, Model MSE is a little bit more incorrect due to its small loss value. The Model Regularization consistently exhibits low accuracy when compared to other models. It follows that, of all the models examined, the Model Hyper performs the best.