**Optimal Neural Network Configurations for IMDB Sentiment Analysis: A Comprehensive Performance Evaluation**

**Assignment - 2**

1. Introduction

For this assignment, the students will train and optimize neural network models for sentiment analysis based on the IMDB dataset. Its goal is to train a model to distinguish between positive and negative movie reviews and to do that using different configurations in the neural network including different numbers of hidden layers using different activation functions, loss functions, and regularization techniques all different methods of configuring them across. Evaluation was done using accuracy and loss on the training and test set datasets.

2. Dataset Overview

IMDB dataset contains fifty thousand extremely positive and negative movie reviews. The reviews are divided into a training set which is of samples of 25,000 and the test set which also contains 25000 samples; The positive and negative samples are in equal proportion in both sets. For processing:

Each review is a n-vector of integers where n is the number of words in given review and every integer is correspondingly to the place of word in given sequence.

The 10000 most often occurring words are included and all the rest are excluded.

To handle the integer values, multi-hot encoding was applied to the data changing the integer sequences to binary matrices, where the presence of a word is marked as 1 and vice versa for 0.

3. Model Configurations

For the current assignment, the various neural networks needed to be trained, and their performance compared. Here are the key modifications made:

One Hidden Layer (32 Units):

Only one hidden layer is employed, and it contains 32 hidden units as a maximum number.

Activation Function: tanh is used instead of the most popular today ReLU. This is why tanh is selected since it can manage both positive and negative criteria. The principal disadvantage of the ReLU function is its inability to handle negative values of the input data, which may be a problem with certain data sets.

Regularization: To prevent overfitting an L2 regularization coefficient of 0.005 is placed in order to discourage large weights.

Dropout: Dropout (with a dropout of 0.5) is employed where during the training phase, 50 percent of the neurons are eliminated at random to prevent over fitting.

Loss Function: binary\_crossentropy is used to compute the loss for the binary classification problem/ Brownian motion and log-normal models are used for predicting stock prices.

Optimizer: The present model uses Adam optimizer for gradient descent.

Purpose: This model was used as a template against which to determine the performance of a neural network with a single hidden layer.

Two Hidden Layers (32, 32 Units):

Two new hidden layers were added, each has 32 units.

Interestingly, the activation function tanh and the L2 term are used in both layers.

There is an addition of Dropout after each of the hidden layers to reduce the problem of overfitting.

The model still employs binary\_crossentropy as the loss function, and the Adam optimizer.

Purpose: This model was built hoping that the inclusion of a second hidden layer would enhance the model’s extent to interpret more compound patterns in the data.

Three Hidden Layers (32, 32, 32 Units) Using mse Loss:

The next newly introduced model is the adding of the third hidden layer which consist of 32 units in each layer.

The activation function is still tanh, and L2 regularization is used.

Dropout (0.5) is applied to fend off the effect after each hidden layer in order to minimize overfitting.

Other than binary\_crossentropy, mse is used for loss function usually used in regression problems but used here in comparison to classification loss functions.

Purpose: This configuration investigated whether increasing network depth and changing the loss function would improve performance.

Smaller Hidden Units (16, 16, 16 Units) Using tanh Activation and mse Loss:

This model also prunes the number of hidden units to the nearest 16 systems in each of the 3 layers.

There is no change in parameters such as activation function which is tanh, L2 regularization and dropout settings.

The mse loss function is used as it was done in the previous model.

Purpose: This configuration explores results obtained when the model’s capacity is pruned by decreasing the number of hidden units with the aim of testing whether it would generalize.

4. Results and Observations

All the configurations were made on the IMDB test set. The only quantitative measure of performance used was accuracy and the loss values were also reported. Here's a breakdown of the results:

One Hidden Layer (32 Units):

Accuracy on Test Set: ~85%

Loss: Both the training and the validation loss were low and there was no issue of overfitting.

Explanation: In this way, the chosen architecture turned out to be sufficient for the model to learn most of the patterns in the data because the model had only one hidden layer. Thus tanh was enabled in handling the positive and negatives inputs with equal convenience, regularization to prevent overfitting and a dropout which was also of importance in preventing overfitting.

Two Hidden Layers (32, 32 Units):

Accuracy on Test Set: ~86%

Loss: This was a slightly better accuracy than the single hidden layer model, but after few iterations, validation loss is increasing and hence overfitting occurs.

Explanation: This caused the network to become more complex and adding a second hidden layer has increased the capacity of the feed forward neural network to capture these underlying patterns in the data. However, this also Increased the probability of developing overfitting despite the constant application of the regularization and dropout methods.

Three Hidden Layers (32, 32, 32 Units) Using mse Loss:

Accuracy on Test Set: ~83%

Loss: Mse loss did bring accuracy down from the binary\_crossentropy models but made training and validation lose curves less juddering, less sign of over fitting.

Explanation: Even though mse is commonly applied to regression problems only, it was used in binary classification to evaluate its impact. The model had slightly inferior accuracy but appeared to train better and maybe because mse yields a smoother optimization curve.

Smaller Hidden Units (16, 16, 16 Units) Using tanh Activation and mse Loss:

Accuracy on Test Set: ~82%

Loss: Therefore, the accuracy of the model was a bit lower compared to other models, however, the reduction of the capacity was adequate in preventing overfitting.

Explanation: The number of hidden units was therefore decreased to reduce the model’s capacity to fully learn the training data, hence helping the model avoid overfitting. While the accuracy rate was lower, this model offered better generalizability of the given task in comparison to another model described in the paper.

5. Conclusion

The experiments show that a certain depth of a neural network is beneficial as it increases the performance, while adding extra layers results in difficulties and risk of overlearning. Tanh as activation function and L2 regularization together with dropout increased model generalization especially when a deeper network was being used. Binary cross entropy was generally more effective than mse when applied to this binary classification problem.

Thus, the best configuration of the model was two hidden layers with 32 neurons each, tanh activation function, L2 for regularization, binary cross entropy loss function was used. This model tested at about 86 percent in accuracy when using the test set.