HW2 report

107041033 李其霖

I. Feature Engineering :

Based on the TA's advices, I utilize the Bagofwords to preprocess the train data and test data. I have already compared the Bagofwords method and Tokenizer method, and then I found that Bagofwords have better results when I latter apply those methods to my model.

II. Model Building :

In this section, I am going to discuss the model I build for this competition. Since Decision Tree, Neural Network, and LSTM Model are frequently discussed in the Kaggle competition in the past, I decide to construct those three models to compare the final results.

1. **Decision tree:**

   This model is much more simpler compared to the latter, so I simply set the random state as 1. Afterwards, I shuffle the data and start training. Eventually, I got the final results – approximately 0.24 score in the Kaggle competition.

2. **Neural Networks:**

   The training data is quite large with exactly 500 features. Therefore, I build not shallow neural network to solve it. Using 500, 128, and 64 as connecting units and setting the activation function as ReLU. In the end, the output layer activation function is softmax.

   To lower the training loss, this model need about 50 epochs to reach higher accuracy. In addition, I utilize the adam as optimizer to boost this model. Eventually, I reach about 0.38 score in the Kaggle competition.

3. **LSTM:**

   LSTM is renowned for solving the text question in Kaggle, and lots of people frequently discuss this model. Hence, I decide to adopt this model.

   I set the units as the same with the features with training data. In addition, since the dimensions of that is three, so I set the return_sequences as true. The units are too large so I need to dropout 20% of the data. In the end, I reach about 0.38 score in the Kaggle competition.

4. **Conclusion:**

Comparing the three model performances, I conclude that according to this three model performance, shallow neural networks achieve higher accuracy in this data.

III. Kaggle Rank :