CS506 hw5 report

Lijun Xiao ljxiao@bu.edu (kaggle username: ljxiao)

1. The algorithm I have implemented:

a. First, I use word2vec to clean up the text, deleting meaningless symbols and keep the text only.

b. Then I transfer these text into feature vector arrays. I changed the input dimension which is the length of the vector array and see what this influence the accuracy.

c. I split the training dataset into training data and validation data, I first trained the model based on the training data and then check the model accuracy on the validation dataset.

d. Then I transfer the corresponding Y_train (that is the score corresponding to each comment) into categorical vectors (one hot encoding). Since there are 1.0-5.0 scores, the number of classes is 5.

e. After these steps, I tried four different models:

   i. First is the Sequential model, I added four Dense layers, between them I added some dropout layers. I used sgd as optimizer.

   ii. Then I used some API to make prediction and compare to the Sequential model:

      1. Random Forest Classifier
      2. XGB Classifier

2. Special tricks to improve result

a. I find that the input dimension has great impact on the final accuracy of the model. After setting this to be larger, the accuracy get improved a lot

b. Also the batch_size has great influence too. Full_batch training maybe slow but with great accuracy.

c. First is change the input dimension which is the length of the feature vector. Because this is the input training date to the model, I found it influence the accuracy result pretty obviously.

d. Second is the the parameters of the CNN model, by adding dropout layers the accuracy get improved a lot.

e. I have also tried to change the split rate between training set and validation set. By put more data into the training set, the accuracy of the trained model gets improved a lot.

3. Strategy for selecting a particular algorithm

a. By using cross validation. I randomly split the original training dataset into training set and validation set. The purpose of this is to test the accuracy of the model both on training set and validation set. If during the training process, the accuracy on the training set keeps improving but the accuracy on the validation set stops improving then it means the model might be overfitting.

b. By comparing the results of the above three methods, I found that the Sequential model seems to perform better than RandomForestClassiier and XGB Classifier

4. How cross validation is performed

a. Cross validation is performed by randomly splitting the training dataset into training set and validation set.