Maching Learning assignment 3

Mei-Shin Wu

July 2016

1 Task 1 : Logistic regression

- Sklearn result
  - Unigram : 0.8435
  - Bigram : 0.8195
- Keras result
  - Unigram : 0.6285
  - Bigram : The code had error. So I can't report accuray here.

2 Task 2 : Multi-layer Perception

Every document was represented as the average of word vectors. The dimensions for unigram and bigram are 50 and 100 correspondingly.

- scikit-neuralnetwork result
  - Unigram : 0.664
  - Bigram : 0.6725

3 Task 3 : Convolutional neural network

3.1 Model description

I trimmed all the documents to equal length, which was 300 words. For the initial attempt, I simply use word sequence to represent each documents. For the second attempt, I insert the word vectors as the weight matrix.

Input layer

- model setting
  - Dropout = 0.2
  - `nb_filter`=200

  - `filter_length`=5
  - subsempling = 2
  - `nb_epoch`=20
- Accuracy : 0.76

4 Task 4: Review

Among all the models, the logistic regression achieved the highest accuracy by using sklearn library. However, the logistic regression model which was provided by Keras , performed a very low accuracy of 0.63.

The accuracy of second task has lowest accuracy among all the models. The input data transformed tokens into the average of all the word vectors. By doing so, the useful information was discard. Thus the second model performed a low accuracy.

The third model, convolutional neural network model reached 0.76 accuracy. It is the second best model among all the task. Nevertheless, convolutional neural network is time consuming (CNN 1 ran for 4 to 5 ourhs, CNN2 ran for over 8 hours). Without a huge amount of training data and information, the CNN model can't perform well.