Text Classifier Model with Artificial Neural Network

CMPT 310
Simon Fraser University
Fall 2018
Group name: ILikeKNN
Group members: Site Li(301244297),
Tianyang Zhou(301244114),
Liheng Ou(301270058)

Introduction

In this project, we are propose to train a text classifier model which can distinguish whether the news is real or fake with artificial neural network. Python is a excellent tool to preprocess text data, train model with cross-validation, and visualizing data due to strong libraries. Another reason that we pick Python rather than Weka is we can test lots of parameters by loops automatically.

Method

Dataset

The dataset are downloaded from coursys. It includes 3000 rows fake and real articles.

Text Preprocessing

Firstly We read the text as string to store in list from text files, and tokenize the strings to words. Then we remove all the stopwords and change all the remain words to lowercase. After that we create a set which can count all the words, and we pick first 500 word frequency as my corpus. We use the corpus to label each text whether they have the words of corpus as the features of data, the y label should be 0 or 1 which indicates fake or real articles.

Model Training

We train the neural network model with 10-fold cross validation, and repeat 10 times, the accuracy of trained model is the average accuracy of 10 times training.

We have tried two options which are the influence of neural network size and the influence of different learning rates.

In option 1, we used two hidden layer for model, and we assigned fixed two nodes in layer 2, and using a loop from 2 to 100 for the size of layer 1. The learning rate is 0.25.

In option 2, mainly we would like to research the effect of learning rate for

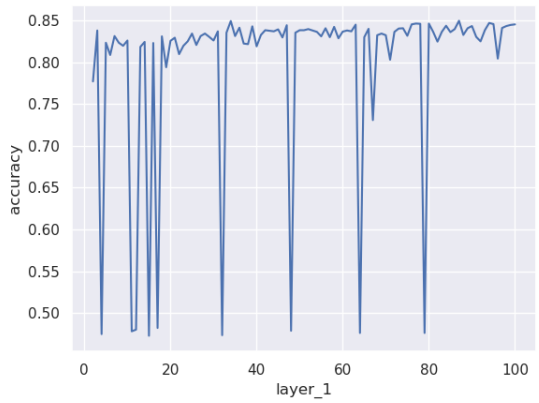
accuracy. We used a fixed hidden layer which the first layer has 3 nodes and the second layer is 20 nodes, the accuracy of this kind of structure is higher. The range of learning rate which we tested is from 0.01 to 1.

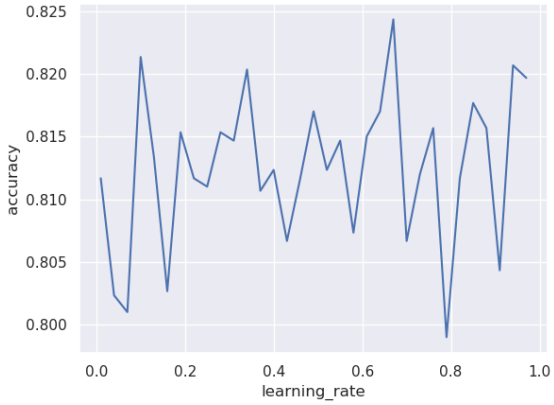
Result

According to the results of two options, we found that the size of layer 1 is 80, we have the highest accuracy 0.847, and when learning rate is 0.67, we have the highest accuracy 0.8243.

Insight

According to the two figures below, in the first plot, if we ignore data of the low accuracy, we can find the size of layer 1 can influence the accuracy slightly, almost data are distributed in range (0.82, 0.85). In the second plot, we cannot find any effect of learning rate for model, the data are distributed randomly between 0.8 and 0.825.





Compared to decision tree, in this task, the performance of two models are too close, neural network can have higher accuracy if we continue to increase the number of nodes or try to others advanced neural network. However neural network is unexplainable, it is not like decision tree which can show the word feature by tree node, we cannot get any implication from node and weights. So we conclude this is the only one clear advantages of decision trees over neural networks.