Assignment 4 Part 2

Rohit Garla	Rg3365

Wednesday, November 28, 2018   4:10 PM

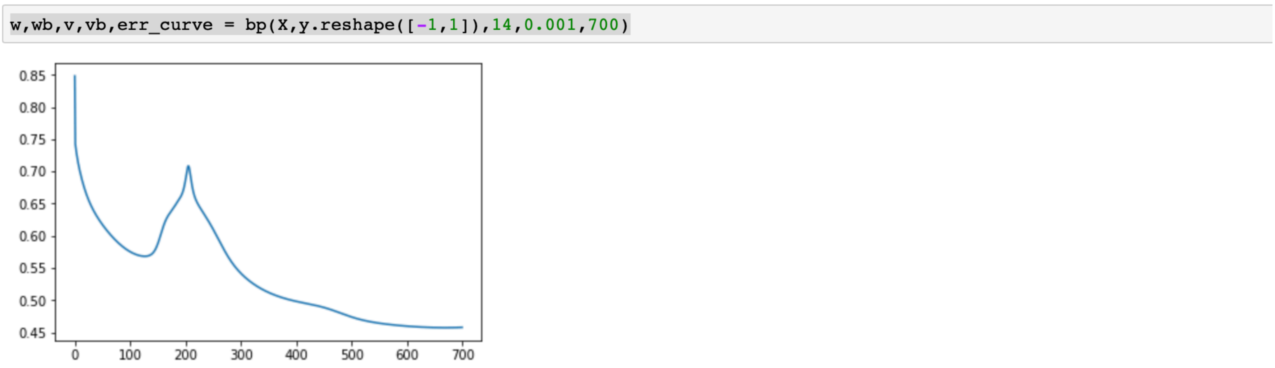
5)
A)
Most Frequent 5 tokens
with stopwords['.' , ',' , 'the' , 'a' , 'and']

B)
Top 5 Info Gain
['bad', 'best', "n't", 'too', 'moving']

C)
Confusion Matrix

129	98
62	11

Accuracy = 0.68 or 68 %
Eta = 0.001
Hidden nodes = 14



```
w,wb,v,vb,err_curve = bp(X,y.reshape([-1,1]),14,0.001,700)
```

D)
ZeroR accuracy = 54.6

E)

i) Why is it reasonable to think that increasing the number of attributes might increase accuracy ?

Answer:

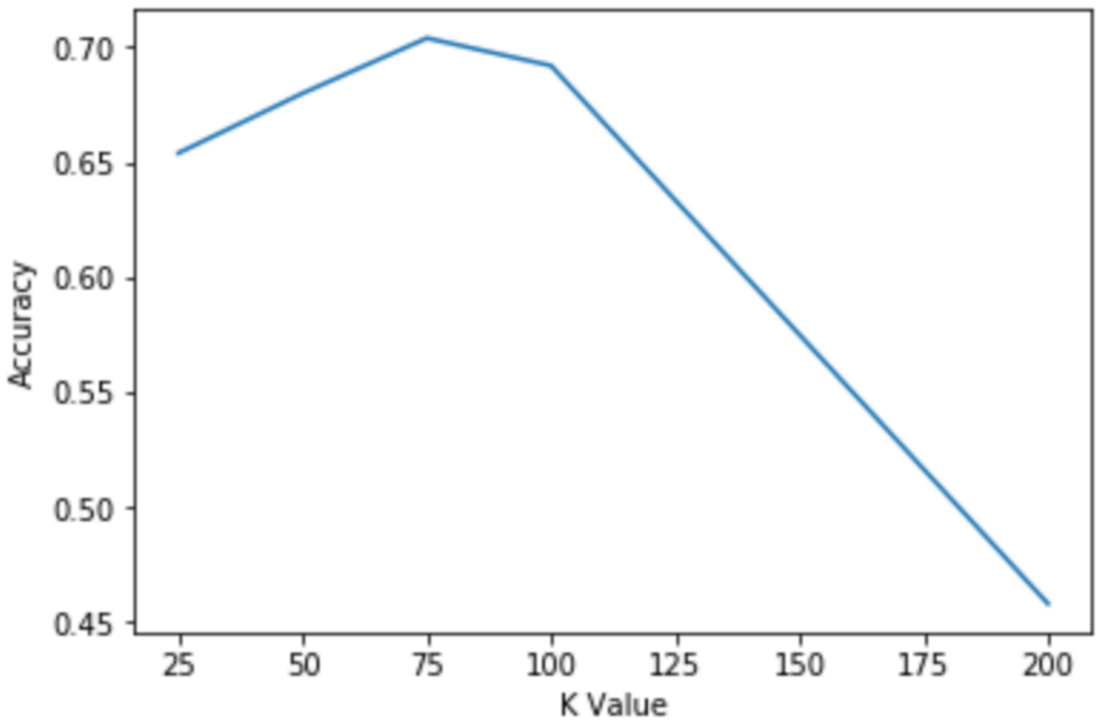
i. On increasing the no of attributes the model, we won't exclude data items in the test data which would have mattered. (Instruction it is given to ignore the tokens which are not present in *Vocabulary). For example there might be token which might have been ignored because it may have been absent in the top k attributes but it may have high info gain for test set. So on increasing the k value there is a possibility of this token to be included which might increase the accuracy*

ii) Why is it reasonable to think that decreasing the number of attributes might increase accuracy ?

Answer:

i. When the no of attributes is high there is a chance of overfitting. Over fitting occurs when the neural network model becomes too complex on the training data.

F) K list = [25,50,75,100,200]



- The answer some what surprised me for lower K value when the accuracy was higher compared to my expectations but it later it followed the same trend as expected. On further increasing the number of attributes the accuracy decreased proving overfitting
- No difficulty in running the experiments. Increasing the number of attributes might cost more computational power and execution time.
- But the set chosen by me was optimal for my machine