TEXT RECOGNITION USING MACHINE LEARNING

By:THE MACHINISTS
Amey Anjarlekar
Shailee Suryawanshi
Atharva Pangarkar

July 2018

0.1 Acknowledgements:

- We would like to thank ERC , WNCC ,Aeromodelling club and all others who are and were involved in the foundation , and development of this programme called ITSP .

- We would also like to thank our mentor DEEP MODH who selflessly helped us and solved our queries even when he was having busy times.

0.2 Motivation:

The plan was basically to learn machine learning and go "deep" into deep learning . So , we took upon ITSP as an opportunity to learn machine learning by solving the "20 Newsgroups " problem of categorizing various texts into 20 classes .

0.3 The Journey

In the initial phase where the learning of theory required for this project was complete , we started with building a simple model to solve the problem undertaken .We planned to do it in two different ways . Using some basic functions from the scikit-learn library we were successful in solving this problem with an accuracy of around 85

This was done using backpropogation .Using Naive Bayes classifier we were able to enhance this to 90.6This was the first part of our project .

Then we took a part of the dataset (belonging to 5 classes of the 20) and built our own neural network and thus were fairly successful in classifying that data . The aim was to modify it such that its accuracy was close to the accuracy got with standard functions available in libraries . We were also able to modify the code such that it took considerable lesser time than the standard functions .

0.4 Some technical aspects : :

- For the first part we used the function MLPClassifier which is based on backpropogation algorithm . Also to convert text into numerical matrices we used the function tfidfvectorizer() .We used the bag of words model for extracting features from text files. Here each unique word corresponds to a new feature. This function finds the words relevant to the document and assigns them values accordingly . It comprises of two parts .If some words are present in certain documents but not in most they are given positive points

. This helps to distinguish the technical words from the rest . eg - words like “cricket” could mean sports. It also gives negative points to words like ‘the’ which could tamper with the machine learning by giving wrong/useless information . So , this part was mostly about getting acquainted with scikit learn and learning how to manage large data.It was also done by the Naive Bayes classifier approach using the inbuild sklearn functions which use the Naive Bayes classifier .The reason for using the inbuilt one is that the efficiency is very good and better than the backpropogation approach .High efficiency in NB classifier (0.906) was achieved by GridSearchCV.

- Also it provided us another approach for this problem and helped us to learn more about this approach .

- Then we moved on to write an extensive neural network code using backpropogation and modifying it by involving bias units . Here, the main challenge was to choose proper learning rates , debugging (because first time so more challenging :P).

0.5 Issues involved during the process :

There is a big question to ask why we went with only 5 categories during the second part . This was mainly because my computer was not able to process high amount of data while running backpropogation codes . It used to get froze/hang . Also there was this issue of a lot of time getting consumed during the iterations(which was solved later by efficient coding) . Another issue which we encountered was as we went ahead with less data (but comparatively more features) now , there was a problem of the data getting overfit which reduced our accuracy quite a bit .

0.6 Theory involved :

The first project involved functions from Naive Bayes classifier and backpropogation separately . As mentioned earlier we used backpropogation algorithm coupled with grading descent(which is necessary) . The theory was learnt from the courses of Andrew NG on coursera (upto week 6) .

0.7 Scope for further improvements :

We could have a better accuracy in the 2nd part of the project , might be by using RNN (Recursive Neural Networks) .

0.8 Component list :

Just a laptop . . . ..

0.9 Useful links :

- https://www.coursera.org/learn/machine-learning
- http://qwone.com/~jason/20Newsgroups/

0.10 Find Us At:

- https://github.com/ameyanjarlekar/20Newsgroupsprac/blob/master/README.md
- https://github.com/atharva244/ITSP