Programming Project 4 for CSCI B555

Saurabh Mathur

December 5, 2018

Task 1

Code

The solution code for Task 1 resides in task1.py and can be run as

```
$ python3 task1.py
```

Questions

Do the topics make sense for the dataset ?
Yes, the topics do seem to make sense. For example 1,2,3 and 5 clearly seem to talk about automobiles and 4 and 7 clearly talk about space.

topicwords.csv

```
diesels,blah,turbo,matter,cars
don,people,cars,power,low
car,ford,shifter,seat,probe
station,shuttle,launch,option,two
engine,feel,big,toyota,small
edu,writes,system,article,apr
mission,hst,shuttle,solar,pat
insurance,geico,each,want,area
oil,service,change,come,lights
clutch,cars,sho,drive,shift
henry,toronto,spencer,edu,writes
bill,moon,earth,etc,back
car,make,even,time,extra
edu,writes,article,apr,find
edu,writes,mustang,article,mail
edu,gif,uci,ics,incoming
sky,george,people,light,rights
space,nasa,long,program,world
don,two,used,point,another
science,internet,part,spacecraft,mars
```

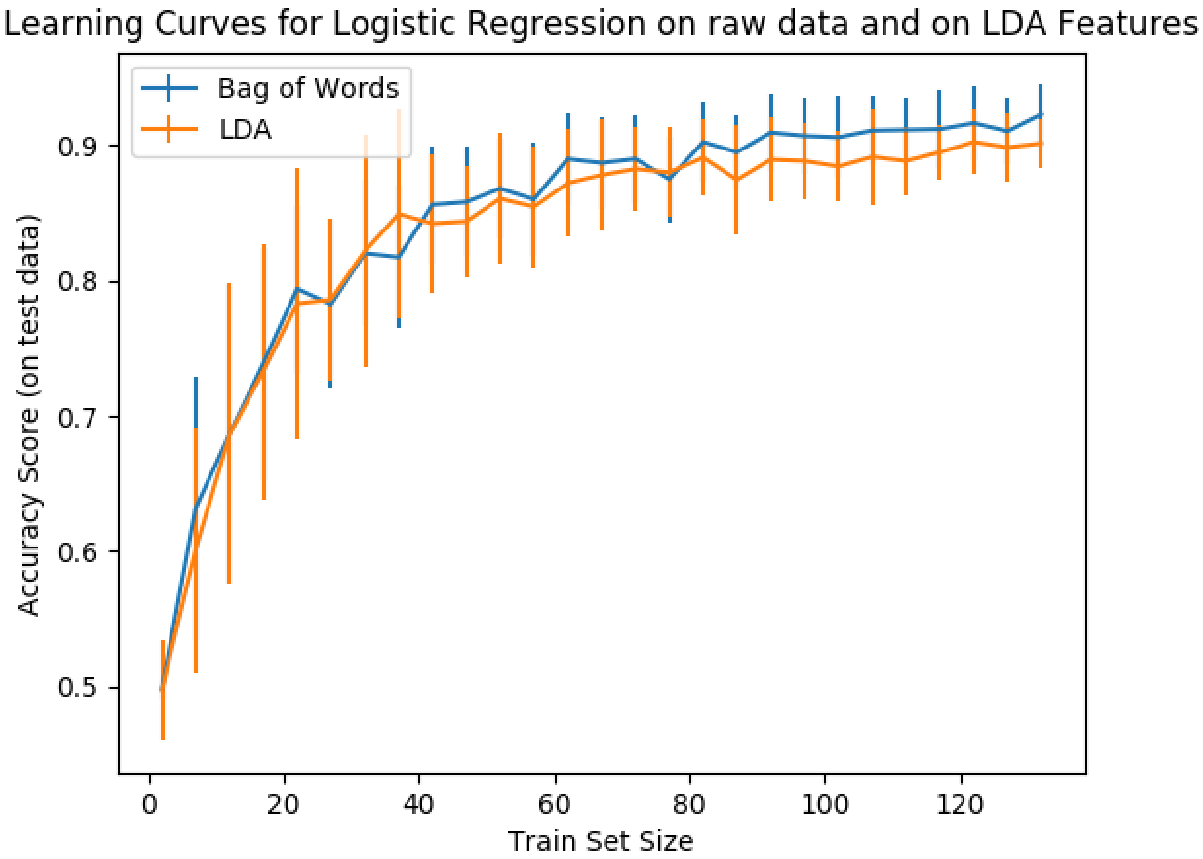

Figure 1: Test set accuracy as a function of increasing train set size

Task 2

Code

The solution code for Task 2 resides in task1.py and can be run as

```
$ python3 task2.py
```

Discussion of results

In the bag of words model each word is a feature and so it has close to 400 features. On the other hand, for the LDA based logistic regression model there are only 20 features, one feature for each topic. So, the size of the LDA based model is about 20 times smaller than the bag of words model.

Figure 1 depicts the learning curve for bag of words (raw data) based and LDA based logistic regression. For less training data, the LDA model seems to perform at par and even better than the bag of words model. However, as more data is available the bag of words model starts performing better. The difference in performance is small and does not seem to grow with further increase in size of training data.