PP4 Report

Logan Bontrager

November 10, 2020

Task 1)

The top five words for each topic are shown below where each line corresponds to a topic and the words are presented in decreasing order.

```
earth,question,large,temperature,things
don,driving,shift,cars,mph
space,science,internet,world,information
station,shuttle,option,launch,redesign
mission,hst,pat,solar,shuttle
car,clutch,shifter,drive,gear
edu,gif,insurance,uci,ics
car,ford,probe,seat,nice
cars,heard,diesels,people,matter
edu,writes,article,apr,find
space,nasa,gov,such,long
back,used,small,sho,use
oil,time,service,come,change
sky,bill,moon,light,rights
car,book,mustang,rear,price
system,oort,edu,david,ray
cost,second,years,cars,comes
engine,turbo,power,steering,temp
henry,edu,toronto,spencer,writes
make,don,point,want,good
```

Here, we can see that most topics have well defined themes. It appears that the two underlying topics present in the data are cars and space. One thing that sticks out to me about the model is that some words carry more weight

in terms of human theme definition than others. For example, topic one contains the word things which doesn't have much weight in terms of personal interprepability although I suppose it could be useful in other ways.

Task 2)

The graphs below show the test error vs. training size for the bayesian logistic regression model trained on both the topic representation data and the bag of words. we can see that both datasets performance similarly. The main disticintion is that the topic representation data appears to preform better on smaller training sizes than the bag of words. This could be advantageous under data scarcity. The error bars from each graph look rather consistent with each other given the y scale.