Fall 2019 Dhruva Bhavsar

B555: Machine Learning

Programming Project 4

Task 1:

The most frequent words obtained for each topic in both the datasets are as below:

Dataset 'Artificial':

Topic	Word1	Word2	Word3
0	bank	water	river
1	loan	dollars	bank

Dataset '20newsgroups':

Topic	Word1	Word2	Word3	Word4	Word5
0	etc	earth	things	day	similar
1	car	ford	nice	probe	dealer
2	car	clutch	shifter	miles	manual
3	sky	insurance	uiuc	geico	light
4	make	don	even	two	use
5	nasa	science	space	gov	internet
6	george	info	idea	howell	great
7	engine	power	toyota	small	seat
8	don	writes	people	edu	want
9	system	point	such	good	each
10	mission	hst	shuttle	solar	pat
11	oil	engine	service	change	bmw
12	space	bill	moon	program	long
13	edu	article	writes	eliot	washington
14	edu	gif	uci	ics	incoming
15	edu	writes	article	good	apr
16	cars	heard	diesels	air	matter
17	launch	station	option	cost	shuttle
18	book	part	another	body	blue
19	henry	edu	toronto	spencer	article

The topics obtained do make sense to a great extent. The words in each topic are mostly correlated. For example, in the artificial dataset we can see that the words in each topic like bank, river and water as well as bank, loan and dollars have relation. Similarly, in dataset '20newsgroups', for topic 11 we can see that having the words oil, engine, service, change and bmw in one sentence will completely make sense.

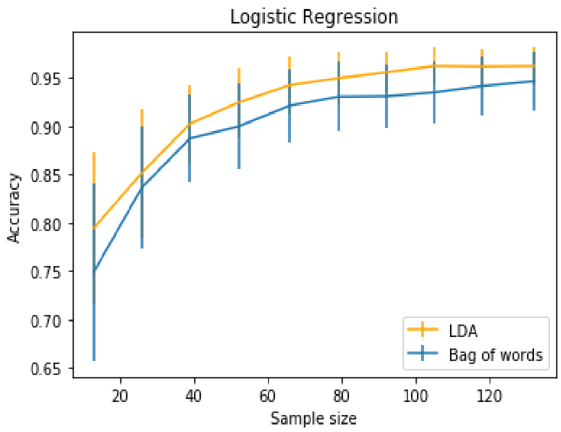
So, Gibbs sampling for LDA is very efficient except for a few mistakes.

The approximate time taken by the code for both the datasets is as below:

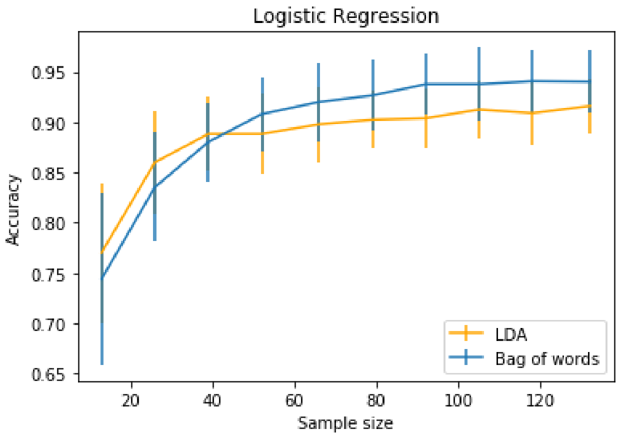
- Artificial: 15.359375 sec
- 20newsgroups: 823.15625 sec

Task 2:

The learning curve for both the topic representation(LDA) and bag of words representation is as shown below:



Case (1)



Case (2)

- While plotting the topic representation vs bag of words representation, sometimes I got more accuracy with topic representation than with bag of words(Case 1) while sometimes I got more accuracy with bag of words representation(Case 2). Although, mostly the accuracy of bag of words representation was more than that of topic representation.
- While the accuracy of bag of words remains the same, the accuracy of topic representation varies everytime. This may be because Gibb's sampling is somewhat random and can produce slightly different results everytime.
- The accuracy for both these representations increases with increase in the sample size.
- The average time and number of iterations taken by both is as below:

Type of representation	Average Time	Average no. of iterations
Topic	0.003125	4.8
Bag of words	1.6578125	11.0

- Bag of words takes more time per iteration compared to topic representation for Logistic regression.