Result and Discussion- Programming Project 4

TASK1: (Gibbs Sampling)

On the Artificial Dataset:

The 3 most frequent word for each topic(k=2) is as following:

Topic0: [‘bank’,’water’,’river’]

Topic1: [‘loan’, ’bank’, dollars]

On the 20Newsgroups Dataset:

The Top 5 most frequent words for each topic in 20Newsgroups dataset is as follows:

Topic 0	*cars*	*big*	*design*	*radar*	*such*
Topic 1	*etc*	*phase*	*diesels*	*reason*	*fact*
Topic 2	*space*	*toronto*	*bright*	*lexus*	*surface*
Topic 3	*mission*	*hst*	*ground*	*alaska*	*bright*
Topic 4	*car*	*miles*	*engine*	*don*	*performance*
Topic 5	*edu*	*writes*	*years*	*don*	*bill*
Topic 6	*bill*	*great*	*email*	*moon*	*local*
Topic 7	*henry*	*astronomy*	*curious*	*wondering*	*extended*
Topic 8	*feel*	*orbit*	*operations*	*make*	*long*
Topic 9	*system*	*degree*	*spacecraft*	*insurance*	*oort*
Topic 10	*cost*	*second*	*life*	*case*	*black*
Topic 11	*edu*	*gif*	*ship*	*capability*	*incoming*
Topic 12	*car*	*couple*	*george*	*cars*	*stick*
Topic 13	*even*	*point*	*such*	*ship*	*high*
Topic 14	*edu*	*writes*	*years*	*use*	*based*
Topic 15	*shuttle*	*station*	*option*	*order*	*mission*
Topic 16	*insurance*	*geico*	*two*	*quite*	*make*
Topic 17	*engine*	*power*	*write*	*nice*	*original*
Topic 18	*sky*	*light*	*important*	*night*	*group*
Topic 19	*oil*	*insurance*	*lights*	*seats*	*cmu*

Observations:

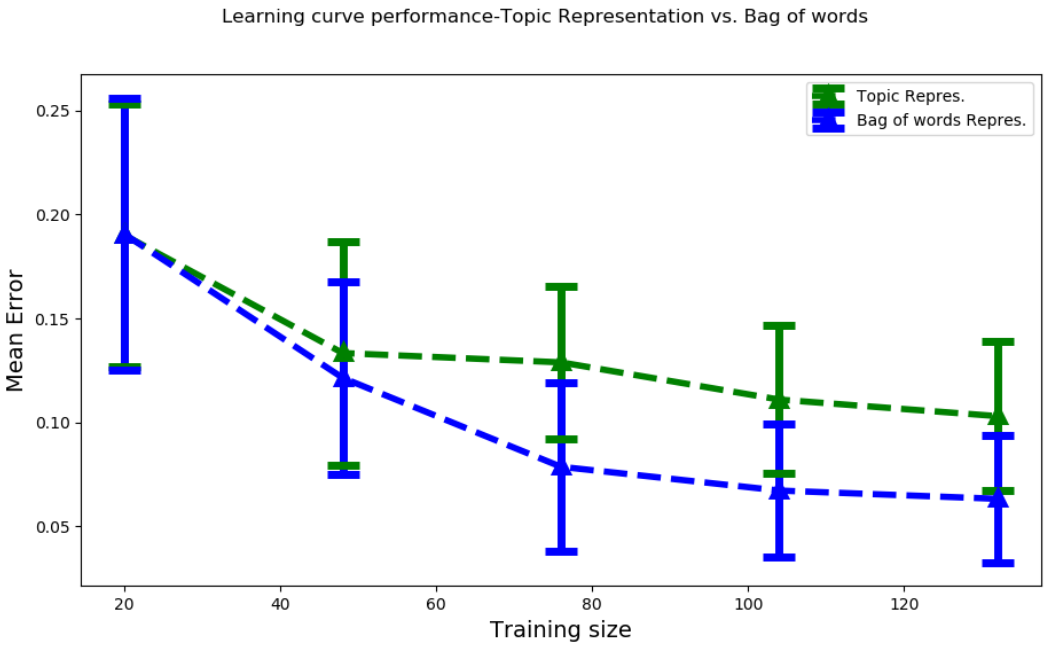
LDA model: The basic idea of LDA is that documents are represented as random mixtures over latent topics; and each topic, in turn, is characterized by a multinomial distribution over words.

Observations:

1. On the Artificial Dataset, the model has done pretty good job. As mentioned above, the topic0 is related to water bodies and topic1 is related to some financial vertical. As including the bank in the both groups make sense, as bank can be using in both ways in a certain context.
2. On the 20Newsgroup Dataset, the model hasn't done good job. As we can observe the words under each topic is random and it doesn't make sense calling the 5 words as a group.
   The model, couldn't have done well because of following reasons:
   Probabilistic models such as LDA exploit statistical inference to discover latent patterns of data. In short, they infer model parameters from observations. These parameters include the number of documents, the length of individual documents, the number of topics, and the Dirichlet (hyper)parameters. If any one of the above parameters isn't align, then model wouldn't be that efficient. One of the conditions mentions in the literature that if log D ≤ N is true (where D is number of documents and N is topics), then LDA could perform well. So LDA even though it is flexible but complex in nature.

TASK2: (Classification using Logistic regression)

The following is the graphs is learning curve performance of logistic regression on 2 representation: topic and "Bag-of-words")



Observations:

- It is evident from the above graph that on smaller training data, the Topic Representation outperforms the Bag-of-words Representation, but as the training size groups, the situation is just reversed, the Bag-of-words Representation performs better than Topic Representation. At max. training size, the mean error rate is about 7% and 12% for the Bag-of-words Representation and Topic Representation respectively.
- We can also observe that, as the training size increases the std of error is decreasing for both the representation.
- The better performance of 'Bag of words' Representation could be of following reasons: It captures enough of topic information and more complex
  representations are hard to model, since they considerably increase the dimensionality of the feature space.

- **The poor performance of the Topic representation could be many of reasons:**
  Probabilistic models such as LDA exploit statistical inference to discover latent patterns of data. In short, they infer model parameters from observations. These parameters include the number of documents, the length of individual documents, the number of topics, and the Dirichlet (hyper)parameters:
  (1) The number of documents plays perhaps the most important role; it is theoretically impossible to guarantee identification of topics from a small number of documents, no matter how long. Once there are sufficiently many documents, further increasing the number may not significantly improve the performance, unless the document length is also suitably increased.
  (2) The length of documents also plays a crucial role: poor performance of the LDA is expected when documents are too short, even if there is a very large number of them. Ideally, the documents need to be sufficiently long, but need not be too long.
  (3) When a very large number of topics than needed are used to fit the LDA, the statistical inference may become inescapably inefficient.
  (4) The LDA performs well when if the topics are concentrated at a small number of words. Another favorable scenario is concerned with the distribution of documents within the topic polytope: when individual documents are associated mostly with small subsets of topics, so that they are geometrically concentrated.