Programming Project – 4

Experiments with LDA

Harika Kanakam

Algorithm 1 –Gibbs Sampling:

* In this algorithm, we have implemented the collapsed Gibbs sampler for LDA.
* Topic representations per document is calculated based on the hidden topic variables associated with words in text.
* We considered 500 iterations to run the sampler with k number of topics. The Dirichlet parameter for topic distribution, alpha value is considered as (k/5) and Dirichlet parameter for word distribution, beta is considered as 0.01.
* In the sampler algorithm, K represents number of topics, w represents array of word indices, d represents array of document indices, z represents initial random topic indices, v represents number of words in vocabulary, and n represents the total number of words in all the documents.
* Also, two matrices C\_d and C\_t are initialized where C\_d is the topic counts per document and C\_t is the words count per topic.
* After the sampler is executed, the 5 most frequent words for each topic is stored in *topicwords.csv* as each row.
* Then, topic representations are calculated using the formula given for each topic in all the documents.

Algorithm 2 – Newton’s Method:

* Train dataset is considered as the 2/3rd part of the entire dataset and Model is trained 30 times for each training size.
* Weights are calculated for each training size by initializing alpha to 0.1 and w to a 0 vector.
* The first class is the labels that are 1s and the 2nd class is the labels that are 0s.
* Mean of both the classes are calculated and the S is calculated using S1 and S2.
* Weights W and Wo are calculated using mean and S.
* Once the weights are calculated, the error on test dataset is calculated for the remaining 1/3rd datasets 30 times of each train size.
* For each train size, the mean and standard deviations of 30 runs are calculated

Task 1 – Gibbs Sampling:

* Using the 1st algorithm, we calculate the 5 most frequency words in each topi is calculated.
* To validate the algorithm, we first run the algorithm with the artificial data and check the correctness of the algorithm.
* The 3 most frequent words of artificial data with 2 topics is as below:  
  Topic 0: bank,water,river

Topic 1: loan,dollars,bank

The most frequency words are validated which are meaningful.

* Now, we run the sampler for the 20newsgroups data. For this, the 5 most frequent words are stored in topicwords.csv and also mentioning below:  
  two,used,ford,back,option

edu,nasa,space,gif,incoming

bill,time,place,spacecraft,put

writes,article,engine,earth,question

edu,henry,writes,uci,find

don,writes,article,edu,insurance

fact,problems,another,bought,great

space,shuttle,shifter,work,sky

edu,space,apr,such,second

manual,mass,washington,sci,hst

article,engines,bit,extra,eliot

station,find,long,probe,redesign

point,comes,mustang,called,spencer

cars,make,clutch,small,feel

want,money,low,sky,nasa

speed,even,day,power,buy

car,driving,mail,interested,things

oil,good,people,diesels,launch

car,book,part,auto,world

high,system,solar,large,henry

* From the result, the words in each topic seems to be meaningful.
* We then calculate the topic representations for each topic in all the documents.

Task 2 – Classification:

* In this portion we will evaluate the dimensionality reduction accomplished by LDA in its ability to support document classification and compare it to the bag of words representation.
* We run the second algorithm to calculate the error rates of the test data with increasing training sizes.
* The results of error rates for both the LDA and bag of word representation is as below:

Table

Description automatically generated

Table

Description automatically generated

The plot of error rates for both is attached below:

Chart, line chart

Description automatically generated

Observations:

* The learning curve performance of the logistic regression algorithm on the two representations looks similar.
* Though there is the dimensionality reduction in LDA and less features compared to bag of words, we could see the similar learning curves for both the representations.
* Gibb’s Sampling algorithm takes more time as that data is huge and number iterations are more. Also, we would be calculating the topics for each and every word in the vocabulary 500 times.
* The time consumption for the classification algorithm is way lesser than the Gibb’s algorithm as we are only calculating the weights, labels and prediction label.