Project Report

B 659 - Applied Machine Learning

LDA over the Discussions Data - Indiana University

Team Members:

Naren Suri - nsuri@iu.edu
Parag Juneja - pjuneja@iu.edu

Abstract:

This project focuses on the idea of using the “Topic Modeling” to classify the discussions done by the students in one particular class of Indiana University. The data given to us is an unsupervised data set with no-class label of topics given, So the idea was to do the topic modeling on the data and also observe if there are any new patterns in the results those are either expected or un-expected among the different discussions. For instance, the professor and AI’s of a class know which topics are usually related and what keywords come together. So, after applying the Topic Modeling if we see a new relation among keywords those extracted then that would be an interesting behaviour to further investigate . The idea was to understand the appropriateness of Topic Modeling - Latent Dirichlet Allocation Generative method to IU discussions data and infer the results. Also, we are implementing our own LDA code to learn the Graphical models and Bayesian Inference techniques Collapsed Gibbs Sampling to infer the results. All the work was done by reading the 3 research papers cited below.

Introduction:

Following its publication in 2003, Blei et al.’s Latent Dirichlet Allocation (LDA) has made topic modeling – one of the most popular and most successful paradigms for both supervised and unsupervised learning. The key update equations and other details on inference are discussed below, the intermediate steps used to arrive at these conclusions are given as direct steps by referring the papers.

In this technical report we will describe our work on discussions data and most importantly, how to implement a working system to perform learning with topic models. We focus on the theory of the stochastic approximate inference technique Gibbs Sampling and then we will discuss implementation details for building a topic model Gibbs sampler.

Data:

The data that we have for the LDA modeling is the Indiana University Students discussions data. This data is given in files saying that each document contains a mixed set of discussions being done over the threads. Also, there were no class labels given.

All we are given are the messages. We will attach the data and the couple of NLP techniques used to clean the data in the final version of our report.

Data Description:

The data that has been used in our problem set is the IU discussion data from one of the visualization course. There are 50 separate text files each containing multiple discussion messages that may be of different topics. In total there are 1609 total discussion messages in all the files combined.

I get that a lot too. Just keep refreshing and eventually it works. Or so its worked for me.
You can also download the slides. There should be a link underneath each embedded video on the pages.
Hi, I joined the course late (just this week). I am currently on 1-06 Sci2intro. Installed Sci2 ok on my 64bit 8GB laptop. Trying to run GUESS visualization on florentine data and got the following message: "The algorithm
Thanks, Orlando. I'm on a ThinkPad 64bit/8GB laptop. Couldn't figure out why it didn't work. Will try to uninstall it, download and re-install it again. I even tried rebooting, which didn't help.
Look at Frank Chum's issues on this discussion thread. Your issues might be memory related, so playing around with those settings could help. Let me know if that doesn't help.
Would it be possible to post a link for all of the Office Hours video sessions in one place? I think that the links for earlier sessions have disappeared.
is it essential to use Sci2 tool for the visualizations ? Is Sci2 useful for prototyping or can be used for publishing on web for example ?
I just checked with our lead developer and he wrote back, "Sci2 is based on the CiShell framework (cishell.org) that uses Java at its core, and which uses OSGi (osgi.org) architecture to allow the integration of a variety of different plugins.

The image above is a screenshot from one of the 50 discussion data files. As it can be seen, each file contains multiple discussion messages one after the other that may be of same or different topic.

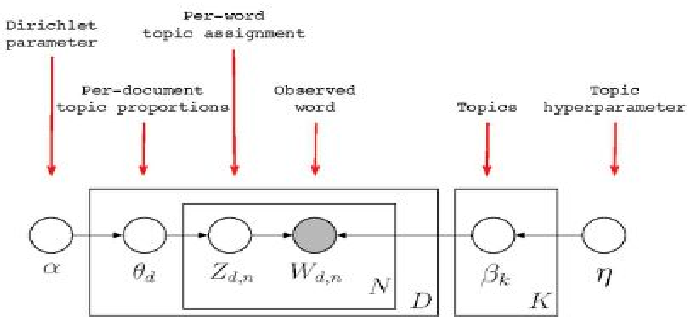
Data Cleaning:
In Order to implement the LDA, the text messages are to be cleaned to make sure that we don't have in-significant words. In Order to remove such words, NLTK package is used to for stopwords and synonyms assignment to words. Here is the sample snippet.

```
    for eachFileInFolder in glob.iglob(Location+"*.txt"):
        TotalFilesToProcess= TotalFilesToProcess + 1
        print "Now Processing the File : " + eachFileInFolder
        FileContents = Path(eachFileInFolder).read_text()
        tokens = tokenizer.tokenize(FileContents.encode('ascii',errors='ignore'))
        StopWordsRemovedText = [word for word in tokens if word not in english_stops]
        words=""
        for idx, word in enumerate(StopWordsRemovedText):
            words = words + " " +stemmer.stem(lemmatizer.lemmatize(word))
        DataHolder.append(words)
    print "Hello, basic stemming and stop word removal is done"
    return(DataHolder)
```

Latent Dirichlet Allocation
LDA is a generative probabilistic model for collections of grouped discrete data. Each group is described as a random mixture over a set of latent topics where each topic is a discrete distribution over the collection’s vocabulary.

K is the number of latent topics in the collection, φ (k) is a discrete probability distribution over a fixed vocabulary that represents the kth topic distribution, θd is a document-specific distribution over the available topics, zi is the topic index for word wi , and α and β are hyperparameters for the symmetric Dirichlet distributions that the discrete distributions are drawn from.

The unobserved (latent) variables z, θ, and φ are what is of interest to us. Each θd is a low-dimensional representation of a document in “topic”-space, each zi represents which topic generated the word instance wi , and each φ (k) represents a K × V matrix where φi,j = p(wi |zj).



Each piece of the structure is a random variable.

Algorithm used in Modeling this Problem:

Input: words w ∈ documents d
Output: topic assignments z and counts $n_{d,k}$, $n_{k,w}$,and n_k
begin
    randomly initialize z and increment counters
    foreach iteration do
        for i = 0 → N − 1 do
            word ← w[i]
            topic ← z[i]
            $n_{d,topic}$ -=1; $n_{word,topic}$ -=1; n_{topic} -=1
            for k = 0 → K − 1 do
                p(z = k|·) = $(n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$
            end
            topic ← sample from p(z|·)
            z[i] ← topic
            $n_{d,topic}$ +=1; $n_{word,topic}$ +=1; n_{topic} +=1
        end
    end
    return z, $n_{d,k}$, $n_{k,w}$, n_k
end

Gibbs Sampling

The MCMC algorithms aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations of stepping through the chain, sampling from the distribution should converge to be close to sampling from the desired posterior. Gibbs Sampling is based on sampling from conditional distributions of the variables of the posterior.

A simpler algorithm is used if we integrate out the multinomial parameters and simply sample zi. This is called a collapsed Gibbs sampler. The collapsed Gibbs sampler for LDA needs to compute the probability of a topic z being assigned to a word wi , given all other topic assignments to all other words. Discussed in detail at the end of the report

LDA Joint Probability :
To save space and not to lose the reader, we have not discussed the proof details clearly here, but we discussed all the technical details at the end of the report.

$$= \prod_{k=1}^{K} \mathrm{Dir}(\phi_k|\alpha) \times \prod_{m=1}^{M} \mathrm{Dir}(\theta_m|\beta) \times \prod_{m=1}^{M} \prod_{n=1}^{Nm} \mathrm{Disc}(z_{m,n}|\theta_m) \times \prod_{m=1}^{M} \prod_{n=1}^{Nm} \mathrm{Disc}(w_{m,n}|\phi z_{m.n})$$

And the statistically sufficient condition is derived later $p(z_{a,b}|z_{-(a,b)}, w, \beta, \alpha)$

LDA implementation Code Pieces:
The Initialization of Alpha and Eta parameters are shown in the code, where the Alpha, Eta is initially sampled from the Dirichlet distribution, Theta is sampled from the multinomial distribution.

```
def generateEachDocument(self):
    # We have previously shown in the function calling this function on how the distribution
    # is being ependent and calleed between each other for sampling
 """
    This is the very important piece of the Generation of document with some distributional
    assumptions:
        1) Topic proportions are to be sampled from the Dirichlet distribution.
        2) From the Multinomial Sample a topic index using the topic proportions from step 1).
        3) Sample a word from the Multinomial corresponding to the topic index from 2).
        4) Go to 2) if need another word.
    """
    theta = np.random.mtrand.dirichlet([self.alpha] * self.K_topics)
    vocabFreqCount = np.zeros(self.V_lengthOfVocab)
    for n in range(self.dafaultDummyDocLenth):
        # Sample from MultiNomial Distribution of the data
        topic_index_z_sampled =  np.random.multinomial(1,theta).argmax()
        word_w_sampled_withz =  np.random.multinomial(1,self.Topic_Word_Distrib[topic_index_z_sampled,:
        vocabFreqCount[word_w_sampled_withz] += 1

    return vocabFreqCount

```

Implemented a Gibbs Sampler over the model for created conditional distributions:

```python
def _Gibbs_estimate(self):
    """Start estimating gibbs sampling
    """
    print "Sampling %d iterations with burn-in of %d (B/S=%d)" \
          % (self.ITERATIONS, self.BURN_IN, self.THIN_INTERVAL)

    self.__init_state()

    for i in range(self.ITERATIONS):
        # one scan of all z_i
        for m in range(len(self.z)):
            for n in range(len(self.z[m])):
                # (z_i = z[m][n])
                # sample from p(z_i|z_-i, w)
                topic = self.__sampling(m, n)
                self.z[m][n] = topic

        # get statistics after burn-in
        if i > self.BURN_IN and i % self.SAMPLE_LAG == 0:
            self.__update_params()

    # nornalize theta and phi
    assert self.SAMPLE_LAG > 0
    if self.SAMPLE_LAG > 0:
        #import pdb; pdb.set_trace()
        self.__compute_theta()
        self.__compute_phi()

    print 'Estimated paramete value is:'
    pprint.pprint(self.theta)
    pprint.pprint(self.phi)
```

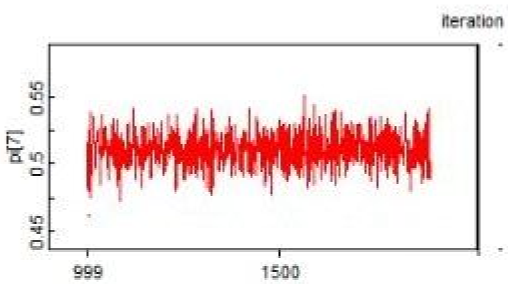
```
''' Here starts the implementation of the algorithm as shown in the
rrsearch paper and as we showed in our report '''
self.Topic_Assignment_EachWord_Z = []
for m in range(self.M_TotalDocuments):
    N = len(self.DataHolder[m].split())
    self.Topic_Assignment_EachWord_Z.append([0]*N)
    for n in range(N-1):
        topic = int (random.random() * self.K_topics)
        self.Topic_Assignment_EachWord_Z[m][n] = topic
        print "updating the counts"
        print n
        print m
        print self.DataHolder[m].split()[n]

        indexToUse = self.vocabulary.index(self.DataHolder[m].split()[n])
        self.n__Words_Topics[indexToUse][topic] += 1
        self.n__Documents_Topics[m][topic]+= 1
        self.total__words_InEach_Topic[topic]+=1
    self.total__words_InEach_Docu = N

## Gibbs and Bayes way to understand the nature of the iterations
if self.SAMPLE_LAG > 0:

    self.theta_sum_val = np.zeros((self.M_TotalDocuments, self.K_topics))
    self.theta = np.zeros((self.M, self.K_topics))
    # cumulative statistics of phi
    self.phi_sum = np.zeros((self.K_topics, self.V))
    self.phi = np.zeros((self.K_topics, self.V))
    # size of statistics
    self.numstats = 0
```

The Lag and Burn_In and convergence were observed for the various iterations. The below image shows that we had a good convergence at 1000 itself.



Experimental Results over different hyper-parameters(alpha,eta):

The gibbs sampler is run over 1000 iterations.

Different values of alpha and eta are used to analyze the results.

(i) alpha = 5 and eta = 5:

The results coming for these values of parameters are meaningful with the topics assigned. Many topic are grouped together as seen in the below picture.

```
0.07463 aesthetics relationships opinion.better criteria password file email
0.07463 low fuel employ unc tests clarification trace brightness pleased properly instantly journal match haven't point couldn't
0.07463 rules brings
0.07463 individual manageable window hi,pfa gamma
0.07463 north ups translates dail typography instantly assumed thin amercican alt co-occurence asterisk variable transition looked understand green needed l

0.07463 models simulation animations conferences deployment exter nyt paths north datawatch http://wearedata.watchdogs.com writing modeling microstrategy's
ctive onwards fellows adjust stay
0.07463 data sci nsf visualization map funding i'm file network work word assignment analysis tool found proportional research homework files gephi
0.07463 high-resolution navigate recommend deadlines jobs full-time downloading narrow they're macs gephi guess
0.07463 ginda blondel recent eric recycling enjoyed poster lewis api delimiter structure bipartite thing people hand-ons pay stream
0.07463 system matt.....there exam chat cloud-based express patharkarsrinivasa afternoon timings pst suites hold proper gap certificate certification tests
ngle score
0.07463 exceeded rupture playlists consolation series type services element form american informative book default unique
0.07463 unavailable tomorrow asks frustration silent cite materials closely studied amercican date occurrence sharing caused
0.07463 sexual compromised call time's seemingly queries concerned descriptions quantitative carefully sitting david areas users opened change program
0.07463 low dropout hasanmichael calculation basically butcher's vizualization fantastic leave stru element bilirubin district config
0.07463 hey kristin food couple award university graph security
0.07463 amazing canada lynchings student editing important feel
0.07463 shift inches orfeb spoil amused alsodependent stanford lookin days longer
0.07463 pagechoosing cns webserver pls dots belongs federal neater assignment.in linked shows graduate info
```

(ii) **alpha = 0.1 and eta = 0.1:**

As shown in figure below for these values of alpha and eta, not many topics are being assigned by the model. But the results are very good, we can clearly see the topics all related to the visualizations tools are strongly classified.

```
0.00149 nsf funding map data sci proportional symbol research state visualization zip chose geospatial scholarly term i'm tableau assignment code found
0.00149
0.00149
0.00149
0.00149
0.00149 exam questions time question data updated don't answers test i'm publish change correct program make file sci large class page
0.00149 http://www.nytimes.com/interactive thisvisualization keplers smart system york show dave....you've moyamoya sharing
0.00149 skype project michael predictive pst tomorrow morning free works good response tools group client sci reached chat microstartegy satishpatharkar lets

0.00149
0.00149
0.00149 personal
0.00149
0.00149
0.00149
0.00149 burst analysis word visualization cloud temporal assignment data bursts wikipedia mesothelioma weight article work idea attached great graph inkscape
```

(iii) alpha = 5 and eta = 0.1:

These values of parameters are very mixed, but not that great compared to the results in ii. Some topics are assigned good related to exam and geospatial vis.

```
 0.07463 belongs easier download?download_frd web
 0.07463 visualizations geospatial visualization data map favorite maps humanitarian facebook pro time simple totally google earth day love interactive based l
gs
 0.07463 i'd reading utterly
 0.07463
 0.07463 temporal time user visualizations visualization blog satish chess taxi class strikes book list attention fantastic called understanding opinion read c

 0.07463 network nsf data bipartite graph files image work created cornell working assignment file homework investigators searched time quiz sci directed
 0.07463 started design join group sara setup background click libr fairly contribute confidence shaken reuters thomson individually dive deeper html/css/javas
front
 0.07463 board lines
 0.07463 content correct
 0.07463 historical project dear client questions understand mitch technical bit side professional analyst dilemma modern query georectified scholarly clarifyi
e compare
 0.07463 tree directory visualization map folder sci radial data visualizations created work view files tool found prefuse treeviz treemap edwin including
 0.07463 treatment notion actively compromised interactively lancashire educating money team drl funding wonderful elicits shor enjoying directory prompt color
```

(iv) For the first case alpha = 0.1 and eta = 5:

These set of parameters give the worst set of results. Most of the topics are unassigned and the keywords are not being assigned.

```
0.00149
0.00149
0.00149
0.00149 proportional
0.00149
0.00149
0.00149
0.00149
0.00149
0.00149 created
0.00149
0.00149 level characters
0.00149
0.00149 effective read
0.00149
0.00149
0.00149
0.00149
0.00149 confusing certificate vague chris manageable window search
0.00149
0.00149
0.00149
```

Future Analysis:
We want to continue the analysis further with the background words selection, which filters out all irrelevant to this domain, this may help improve the model better. Also, want to use the various parameters of the Gibbs and Hyper parameters to test this more.

References :
Implementation of Latent Dirichlet Allocation using the below Research Papers

Latent Dirichlet Allocation David M. Blei , Andrew Y. Ng
https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf

Bob Carpenter. Integrating out multinomial parameters in latent dirichlet
#allocation and naive bayes for collapsed gibbs sampling. Technical report,
#Lingpipe, Inc., 2010.

#https://lingpipe.files.wordpress.com/2010/07/lda3.pdf

Proof that was followed to estimate the p(z/.) LDA Sampling Model:

$M \in N_+$ is the number of documents. $N_m \in N_+$ is the number of words in the m-th document. J is the number of distinct words. K the number of topics. $w_{m,n} \in 1{:}J$ is the n-th word of the m-th document and $z_{m,n} \in 1{:}K$ is the topic to which it is assigned. $\theta_m \in [0, 1]^K$ is the topic distribution for document m. $\phi_k \in [0, 1]^J$ is the word distribution for topic k. $\beta \in R^k_+$ is the vector of prior counts (plus 1) for topics in documents and $\alpha \in R_+^J$ is the vector of prior counts (plus 1) for words in a topic.

In sampling notation, we draw the word distribution for topic k by

$\phi_k \sim \mathrm{Dir}(\alpha)$ for $1 \le k \le K$ (1)

For each document m, we draw its topic distribution,

$\theta_m \sim \mathrm{Dir}(\beta)$ for $1 \le m \le M$ (2)

For each word n in document m, we first draw the topic $z_{m,n}$ from the distribution over topics for the document m,

$z_{m,n} \sim \mathrm{Disc}(\theta_m)$ for $1 \le m \le M$ and $1 \le n \le N_m$ (3)

then draw the word $w_{m,n}$ itself from the word distribution for the word's topic, $z_{m.n}$,

$w_{m,n} \sim \mathrm{Disc}(\phi_{zm,n})$ for $1 \le m \le M$ and $1 \le n \le N_m$ (4)

LDA Joint Probability

Given the model, the joint probability for all of the parameters in the LDA model is

$p(w, z, \theta, \phi|\beta,\alpha)$ (5)

$= p(\phi|\alpha)\ p(\theta|\beta)\ p(z|\theta)\ p(w|\phi, z)$ (6)

$$= \prod_{k=1}^{K} p(\phi_k|\alpha) \times \prod_{m=1}^{M} p(\theta_m|\beta) \times \prod_{m=1}^{M}\prod_{n=1}^{Nm} p(z_{m,n}|\theta_m) \times \prod_{m=1}^{M}\prod_{n=1}^{Nm} p(w_{m,n}|\phi z_{m.n}) \quad (7)$$

$$= \prod_{k=1}^{K} Dir(\phi_k|\alpha) \times \prod_{m=1}^{M} Dir(\theta_m|\beta) \times \prod_{m=1}^{M}\prod_{n=1}^{Nm} Disc(z_{m,n}|\theta_m) \times \prod_{m=1}^{M}\prod_{n=1}^{Nm} Disc(w_{m,n}|\phi z_{m.n}) \quad (8)$$

Integrating out Multinomials in LDA

The collapsed sampler needs to compute the probability of topic $z_{a,b}$ being assigned to $y_{a,b}$, the b-th word of the a-th document, given $z_{-(a,b)}$, all the other topic assignments to all the other words.

$$p(z_{a,b}|z_{-(a,b)}, w, \beta, \alpha) \quad (11)$$

By the definition of conditional probability,

$$= \frac{p(za,b|z-(a,b)\ , w|\ \beta, \alpha)}{p(z-(a,b)\ , w|\ \beta, \alpha)} \quad (12)$$

Remove the denominator, which does not depend on $z_{a,b}$,

$$\propto p(z_{a,b},z_{-(a,b)}, w|\ \beta, \alpha) \quad (13)$$

Note that $z_{a,b},z_{-(a,b)}$ is just z,

$$= p(y, z|\ \beta, \alpha) \quad (14)$$

Using the sum rule (or rule of total probability), integrate out the topic distributions for each document, θ, and the word distributions for each topic, ϕ,

$$= \iint p(w, z, \theta, \phi|\beta,\alpha)\, d\theta\, d\phi \quad (15)$$

Expand the integrand given the model defined in (6),

$$= \iint p(\phi|\alpha)\, p(\theta|\beta)\, p(z|\theta)\, p(w|\phi, z)\, d\theta\, d\phi \quad (16)$$

$$=\int p(z|\theta)\, p(\theta|\beta)\, d\theta \times \int p(y|\phi, z)\, p(\phi|\alpha)\, d\phi \quad (17)$$

And then expand out the terms again according to the independence assumptions in (7),

$$= \int \prod_{m=1}^{M} p(z_m|\theta_m)\, p(\theta_m|\beta)\, d\theta \times \int \prod_{k=1}^{K} p(\phi_k|\alpha) \prod_{m=1}^{M}\prod_{n=1}^{Nm} p(w_{m,n}|\phi z_{m.n})\, d\phi \quad (18)$$

For the same reason as we could separate two products, we may separate multiple products when other terms are constant, so we may distribute the multivariate integrals through the products over the dimensions,

$$= \prod_{m=1}^{M} \text{p}(\text{z}_\text{m}|\theta_\text{m})\ \text{p}(\theta_\text{m}|\beta)\ \text{d}\theta_\text{m} \ \times \ \prod_{k=1}^{K} \int \text{p}(\phi_\text{k}|\alpha) \ \prod_{m=1}^{M} \prod_{n=1}^{N_m} \text{p}(\text{w}_{\text{m,n}}|\phi \text{z}_{\text{m.n}}\ )\ \text{d}\phi_\text{k} \tag{19}$$

Expand out the Dirichlet priors and the discrete distributions according to their usual definitions,

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum_{k=1}^{K} \beta_k)}{\prod_{k=1}^{K} \Gamma(\beta_k)} \prod_{k=1}^{K} \theta_{m,k}^{\beta_k - 1} \prod_{n=1}^{N_m} \theta_{m,z_{m,n}} \text{d}\,\theta_m \ \times \ \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\sum_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} \phi_{k,j}^{\alpha_j - 1} \prod_{m=1}^{M} \phi_{z_{m,n} y_{m,n}} d\phi_k \tag{20}$$