IE 506: Machine Learning - Principles and Techniques **Jan-May 2023**

End-term Project Report : Probabilistic Latent Semantic Analysis

Team Name: OG *Team Member:* 22*N*0459

Abstract

In this project our main goal is to given a set of unknown documents we need infer about them without go through the entire documents. The algorithm proposed in this paper will solve this problem and it is also solves the problem of polysems and synonyms when understanding natural language by a computer. In this work we first talk about Probabilistic Latent Semantic Analysis (PLSA) which is inspired by the Latent Semantic Analysis. and next we move on Gibbs sampling technique to derive topics from a given document.

1 Introduction

This projects

We provide a survey of existing literature in Section 2. Our proposal for the project is described in Section 3. We give details on experiments in Section 5. A description of future work is given in Section 7. We conclude with a short summary and pointers to forthcoming work in Section 8.

2 Literature Survey

2.1 Latent Semantic Analysis

-This is the only method avialable before Probabilistic Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a statistical technique used in natural language processing to analyze relationships between a set of documents and the terms they contain. The basic idea behind LSA is to represent documents as vectors in a high-dimensional space, where each dimension represents a term or concept. LSA then applies a mathematical technique called Singular Value Decomposition (SVD) to reduce the dimensionality of the space and identify the underlying semantic structure of the documents.LSA can be used for a variety of tasks, including information retrieval, text classification, and document clustering. For example, LSA can be used to find documents that are similar to a given query or to group together documents that are topically related.One of the strengths of LSA is that it can capture the meaning of words based on their usage in context, rather than relying solely on their literal definitions. However, LSA has some limitations, such as its inability to handle word sense disambiguation and its reliance on a large corpus of training data to accurately capture the underlying semantics of a language.
Algorithm

- Given the documents we will find the document-word matrix. $N_{i,j} = n(d_i, w_j)$ where $n(d_i, w_j)$ indicates number of times the word w_j came in the document d_i
- Now we will apply Singular Value Decomposition on the document-word matrix.
- Now we will get three matrices

  - DxT - Document - Topic Matrix
  - TxT - Topic - Topic Matrix(diagonal matrix) - Each topics are singular values with ranking .
  - TXW - Topic - Word Matrix

- No for given topic k we will take that value in TxT matrix as 1 and rest as zero and matrix multiply of DxT,TxT - for documents in that topic, TxT,TxW for words in that topic.we will all words,documents whose data value ¿0 after multiplication.

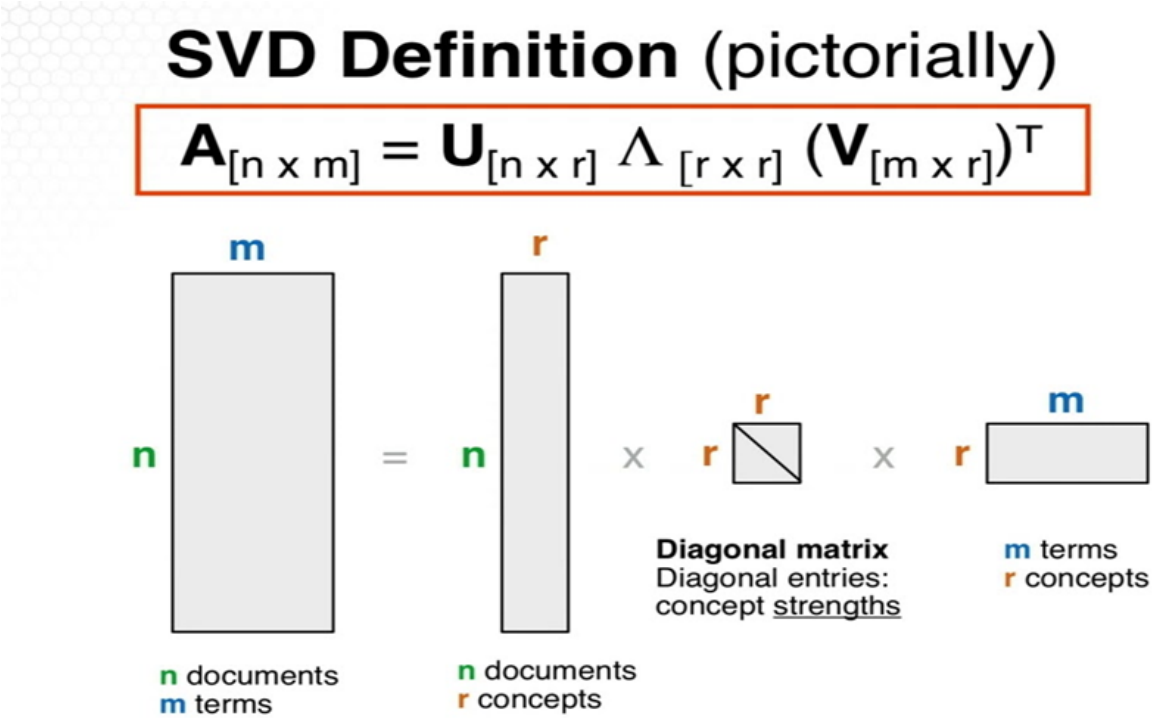

Figure 1: Latent Semantic Analysis

3 Methods and Approaches

3.1 PLSA

Probabilistic Semantic Analysis (PLSA) is a technique for latent semantic analysis that models the semantic relationships between words in a corpus using a probabilistic approach. The main assumption of PLSA is that given a latent class variable z the the words and documents are conditionally independent to each other.here z is the latent class variables which represents the topics.

$$P(d,w) = P(d)(P(d|w) \qquad P(w|d) = \sum_{z \in Z} P(w|z)P(d|z)$$

Now we need to calculate the likelihood function which is

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\ \log P(d_i, w_j)$$
$$= \sum_{i=1}^{N} n(d_i)\left[\log P(d_i) + \sum_{j=1}^{M}\frac{n(d_i, w_j)}{n(d_i)}\ \log\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i)\right],$$

Figure 2: Likelihood function of PLSA

In order to maximize the likelihood function and find the posterior distributions $P(w|z) and P(z|d)$ we use expectation and maximization algorithm .

Expectation Step

--

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l)P(z_l \mid d_i).}$$

Maximization Step Derivation Steps Finding the expectation of likelihood function

$$\mathbf{E}\left[\mathcal{L}^c\right] = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\sum_{k=1}^{K} P(z_k \mid d_i, w_j)\ \log\left[P(w_j \mid z_k)P(z_k \mid d_i)\right].$$

After adding the normalizing constraints the Lagrangian form of expected likelihood function is

$$\mathcal{H} = \mathbf{E}\left[\mathcal{L}^c\right] + \sum_{k=1}^{K}\tau_k\left(1 - \sum_{j=1}^{M} P(w_j \mid z_k)\right) + \sum_{i=1}^{N}\rho_i\left(1 - \sum_{k=1}^{K} P(z_k \mid d_i)\right).$$

After Solving the the above equations **Maximization Step**

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m) P(z_k \mid d_i, w_m)},$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k \mid d_i, w_j)}{n(d_i)}.$$

We need iterative use of these steps Estep and MStep one after other using values from one step used in the another step until they converge
Algorithm

- First randomly assign values to the posterior distribution $P(w|z)$ and $P(z|d)$
- Use Estep to calculate $P(z_k|d_i, w_j)$ for all k topics, i documents,j words
- Use Mstep to update parameter $P(w|z)$ and $P(z|d)$
- Repeat the above two steps one after other repeatedly
- We can stop the algorithm when all the parameter values converge or keeping a threshold value and if change in likelihood is less than that threshold value after each Estep and MStep iteration .

3.2 Gibbs Sampling

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method that can be used to estimate the posterior distribution of the hidden variables in a probabilistic model. In this answer, I will explain how to implement PLSA using Gibbs sampling.

Step 1: Define the PLSA model PLSA assumes that there is a set of latent topics in the corpus, and that each document is a mixture of these topics. Each word in the document is generated from one of the topics with some probability. The generative model for PLSA can be defined as follows:

For each topic k:

Draw a distribution over words ϕ_k from a Dirichlet prior with parameter α For each document d: Draw a distribution over topics θ_d from a Dirichlet prior with parameter β

For each word w in document d:
* Draw a topic z from the topic distribution θ_d

* Draw a word from the word distribution ϕ_z

- Step 2: Initialize the variables Initialize the topic assignments z for each word in the corpus randomly.
- Step 3: Define the likelihood function, The likelihood function for the PLSA model is:
  $P(w, z, \theta, \phi) = P(\theta|\beta) * P(\phi|\alpha) * P(z|\theta) * P(w|z, \phi)$

- Step 4: Define the conditional probabilities Using Bayes' rule, we can derive the conditional probabilities for the Gibbs sampler:

  $P(z_i = k|w, z_{-i}, \theta, \phi)(n_{d,i,k} + \beta_k)/(n_{d,i} + sum(\beta)) * (n_{w,i,k} + \alpha_i)/(n_k + sum(\alpha))$

  where

  - z_i is the current topic assignment for word i
  - w is the entire corpus
  - z_{-i} is the set of topic assignments for all words except i
  - $n_{d,i,k}$ is the number of words in document d assigned to topic k, excluding word i
  - $n_{d,i}$ is the total number of words in document d, excluding word i
  - $n_{w,i,k}$ is the number of times word i is assigned to topic k, excluding its current assignment
  - n_k is the total number of times topic k is assigned in the corpus
  - β_k is the parameter for the Dirichlet prior over the topic distribution for document d
  - α_i is the parameter for the Dirichlet prior over the word distribution for topic k

- Step 5: Run the Gibbs sampler At each iteration of the Gibbs sampler, we update the topic assignment for each word in the corpus using the conditional probabilities derived in step 4. After a burn-in period, the samples can be used to estimate the posterior distribution over the topic and word distributions.

- Step 6: Calculate the posterior distribution To estimate the posterior distribution over the topic and word distributions, we can use the samples generated by the Gibbs sampler to calculate the empirical mean of the topic and word distributions.

  This process can be repeated for multiple iterations until convergence. Once the PLSA model has converged, the resulting topic and word distributions can be used for various natural language processing tasks such as document classification, topic modeling, and information retrieval.

3.3 Calculating Topics

From the above two algorithms we are able to calculate the topic-word distribution and document - topic distribution and we require l words in each topic.Now in order to calculate the topics given a topic k and the we will see the probabilities of the words and sort them in descending and take the top l words as the words in that topic.

3.4 Work done before mid-term project review

As if we want explain work done before mid term project review then it can be explained as follows:

- Reading the research paper and understood the complete methodology described in the paper.
- Implement the PLSA algorithm and experimented on the 3 datasets mention the data set section.
- implemented the python code of data cleaning procedure
- Wrote the code from scratch

3.5 Work done after mid-term project review

As of till now we have completed the idea discussed in the paper by author and also our own ideas what we thought. It all can be viewed as follows:

- Read and Understood the Gibbs sampling technique.
- Implemented the Gibbs Sampling technique from scratch in the python
- Performed Gibbs Sampling on all 3 data sets
- Performed Gibbs sampling and PLSA o ELon Musk tweets data set after data cleaning which is also implemented from scratch.

4 Data set Details

- **Data Set 1**The first data set is 16 documents about one piece from Wikipedia.
- **Data Set 2**The second data set is 100 documents from the Associated Press.
- **Data Set 3(Japense)**The third data set is 50 documents from sina.
- **Elon Musk Tweets Data**Elon Musk is an founder and/or cofounder and/or CEO of SpaceX, Tesla, SolarCity, OpenAI, Neuralink, Musk is famously active on Twitter. This dataset contains all tweets made by @elonmusk, his official Twitter handle, between November 16, 2012 and September 29, 2017.From this data set In this paper we consider the tweet column as the documents

5 Experiments

- Performed data cleaning on the Data Sets 1,2,3
- Applied the PLSA and Gibbs Sampling Algorithm on the Data Set1,Data Set 2, Data set 3
- Performed data cleaning by removing and unnecessary details and used bag of words transformation .
- Applied PLSA and Gibbs Sampling Algorithm on the Elon Musk Tweets Data Set

6 Results

The results what are obtained by our work are described as follows:

1. **For Data Set1**
   Using Expecataion Maximization Algorithm : Time taken - 7iterations(Max 20) – Time taken – 3 sec (Document, word) dimension – (16,742)
   Topics:
   - luffy ,hat,dressrosa, flame,straw ,fruit, save, leading ,weapons, ancient
   - Piece, grand, haki, manga, red, blue, bur, mountain, series, color

- Island, pose, alabasta, baroque, pirates, fishman, log, magnetic, crew, grand
- Sea, devil, fruit, user, fruits, water, series, called, powers, north
- Luffy, crew, pirates, roger, navy, franky, government, island, pirate, zou

Using Gibbs Sampling : Time taken – 7 iterations (Max 20) times taken – 3 sec
Topics:

- Luffy,pirates ,crew, straw, alabasta, baroque, island, piece, hat, named
- Luffy,haki, crew, island, pirates, ace, battle, navy, color, alliance
- Devil, fruit, user, sea, fruits, luffy, zou, water, powers, pirates
- Grand, sea, island, su, red, blue, bur, mountain, pose, called
- Manga, series, piece, dressrosa, animals, north, luffy, flame, America. video

2. **Data Set 2**
Using Expecation Maximization Algorithm : Time taken - 20 iterations(Max 20) – 4 min Document – word dimension – (100,6250)
Topics:

- bank global warming percent president immigration police american el summit
- bush dukakis people president administration government campaign told noriega rating
- percent oil company prices rate soviet report government month billion
- soviet gorbachev central fbi tuesday degrees people city polish expected
- people fire police barry roberts thursday officials waste greyhound magellan

Using Gibbs Sampling : Time taken – 20 iterations(Max 20) – 4 min Topics:

- percent police people fbi california government city barry agents york
- percent soviet government prices oil people rose saudi congress jewish
- bank company percent duracell magellan global summit spacecraft warming thursday
- soviet central union people noriega gorbachev officials official school greyhound
- bush dukakis people fire president campaign roberts rating children city

3. **Data Set 3 - Japenese Dataset**
Using Expectation Maximization Algorithm : Time taken - 20 iterations(Max 20) – 2 min, Document – word dimension – (50,5757)
Topics: (Translated using google translator)

- He Tian, child, Liu Yao, mobile phone, Heyuan, reporter, police, tell old man, parent
- police, man, reporter, Jay Chou, accident, vehicle, police, high speed
- Villagers, dividends, police, scenic spots, counterfeit money, tourists, Longchi, rescue, search and rescue, logistics
- New New Kids Driver Mosaic Police Weibo School Janus Van Discovery
- Traffic police propose marriage dividends Wang reporter policeman ticket villager woman in red

Using GibbsSampling :Time taken – 20 iterations(Max 20) - time taken 3 min
Topics:

- The police, the police, the reporter, the man Jay Chou, the counterfeit currency, the hospital, the police, the logistics Yao, Heyuan, Police, Man, Reporter, Child, Policeman, Red Clothes, Take away, Report

- Dividends, villagers, scenic spots, tourists, Longchi, Dadun, rescue, Tengchong, banks, search and rescue
- Police Man Driver Mosaic Marriage Proposal Van Weibo Reporter Woman Ms. Li
- He Tian, child, new, new, mobile phone, school, happened, Jian Feng, drag racing, thief, Guangzhou

Elon Musk Tweets Data Set[Ref.7]
Using Expectation Maximization Algorithm: Document - Word dimension : (3109,3163)
Topics:

- the, of, and, in,good,
- for, the, to, and, on,
- is, of,but,yes,the,
- to,in,at,model,for
- the,to,be,on,of
- to, the,is,we,that

Using Gibbs Sampling Algorithm
Topics:

- a4 , aaronpaul8 , fellow , fight , fig
- a4 , fear , fight , fig , field
- a4 , aaronpaul8 , fellow , fight , fig
- a4 , aaronpaul8 , fellow , fight , fig
- a4 , aaronpaul8 , fellow , fight , fig

7 Future Work

As of now we are able to derive the topics from the documents but we can use those topics to form text summarizing.This topics can be used to retrieve document summaries which are useful for the user without reading through the entire document.We can improve the algorithm to calculate topic- word distribution and document topic distribution which is faster then gibbs and EM algorithm used in this work.

8 Conclusion

We cans see from the use of PLSA and Gibbs sampling we were able to derive meaningful topics when the documents has stories.In the case of elon musk tweets dataset we where not able to infer much information from the topic derived . This was happened beacause each tweet will be short in size and does mean a story most of times.

References

1. T. Hofmann, “Probabilistic latent semantic analysis,” in Proc. 15th Conf. Uncertainty Artif.Intell., Stockholm, Sweden, Jul. 1999,pp. 289–296
2. S. Deerwester, S. T. Dumais, G. W. Furnas, Landauer.T.K., and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 1990.
3. T. Hofmann, J. Puzicha, and M. I. Jordan. Unsupervised learning from dyadic data. In Advances in Neural Information Processing Systems, volume11. MIT Press, 1999.
4. Gelfand, A. (2000). Gibbs Sampling. Journal of the American Statistical Association,95(452): 1300–04
5. Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent Dirichlet Allocation. J. Mach. Learn. Res. 2003, 3, 993– 1022
6. Initial datasets and coding where taken from this github repo
   link
7. Elon Musk Data Set taken from kaggle.
   https://www.kaggle.com/datasets/kulgen/elon-musks-tweets