Latent Dirichlet Allocation For Text And Image Topic Modeling

Tsung-Yi Lin and Chen-Yu Lee
Department of Electrical and Computer Engineering
University of California, San Diego
{tsl008, chl260}@ucsd.edu
February 28, 2013

Abstract

Latent Dirichlet allocation (LDA) is a popular unsupervised technique for topic modeling. It learns a generative model which can discover latent topics given a collection of training documents. In the unsupervised learning framework, where the class label is unavailable, it is less intuitive to evaluate the goodness-of-fit and degree of overfitting of learned model. We discuss two measurements 1) harmonic mean and 2) perplexity to measure goodness-of-fit and degree of overfitting respectively. In this report, we apply LDA trained with Gibbs sampling to discover topics of text and image. The goal is to use Classic400 dataset for discovering topics of text and binary charecter image dataset for meaninful elements of images. We demonstrate the outcome of LDA by visualizing the topic probability given each document for text data and showing the image topic as meaningful parts of characters.

1 Introduction

Latent Dirichlet allocation (LDA) is an unsupervised learning algorithm which learns the generative model for words given a document with latent topics. It is first proposed by [1] and now widely used in data mining and computer vision community. In this report, we want to apply LDA to both text and image data to discover their latent topics. We implement Gibbs sampling algorithm to efficiently learn LDA model. We use Classic400 for text dataset [1] and binary character image dataset [2]. The performance evaluation of LDA is less clear than supervised learning algorithms which use the accuracy of label prediction as performance metric. We visualize the probability of topic given a document in 3D space for the text dataset and show topic images of binary character image dataset. We measure the goodness-of-fit and degree of overfitting by introducing harmonic mean [2] and perplexity [3]. We show the topic modeling can discover the meaningful topics for text and meaningful parts for images. We find harmonic mean and perplexity are two measurements allow to select hyperparameters to learn a model that have the balance between goodness-of-fit and overfiting with cross-validation.

2 Algorithm Design and Analysis

In this section, we introduce the principle of LDA. We discuss how to use Gibbs sampling for learning LDA. Harmonic mean and perplexity are introduced to measure the goodness-of-fit and degree of overfitting.

[1] http://cseweb.ucsd.edu/users/elkan/151/classic400.mat
[2] http://psiexp.ss.uci.edu/research/programs_data/exampleimages2.html

2.1 Latent Dirichlet Allocation

LDA is a generative model that has 3 hyperparameters topic number K, Dirichlet prior for topic distribution α, and Dirichlet prior for word distribution β. The latent topic model are topic distribution θ and word distribution ϕ. The generative model generate corresponding topic z for each word in each document. It works by selecting θ and α from $Dir(\alpha)$ and $Dir(\beta)$ and selecting topic z from θ and word w from ϕ_z. The next section we will talk about how to learn latent topic given 3 hyperparameters with Gibbs sampling.

2.2 Gibbs Sampling For LDA

The goal of Gibbs sampling is to infer the topic z for each word in each document. We compute probability of $p(z_i = j|\bar{z}', \bar{w})$ to assign word z_i to topic j which maximizes the probability. $\bar{z}'$ is fixed when estimating z_i and $\bar{w}$ is the corpus in the training data. The prime notation means the examples is from fixed training in current iteration of Gibbs sampling. The probability can be computed by:

$$p(z_i = j|\bar{z}', \bar{w}) \propto \frac{q'_{jw_i} + \beta_{w_i}}{\sum_t q'_{jt} + \beta_t} \frac{n'_{mj} + \alpha_j}{\sum_k n'_{mk} + \alpha_k} \tag{1}$$

where q'_{jw_i} is the counts of word w_i to topic j. n'_{mj} is topic j counts in document m. The meaning of Dirichlet prior is clear in this equation. They add pseudo topic and word counts. So if we know that words are ambiguous in many topics in prior, we should give higher value for α; otherwise, α should $<< 1$. Same for β, if we know each document tend to have many topics in prior, we should give higher value for β; otherwise, β should $<< 1$. The complexity of Gibbs sampling is $O(KN)$ to estimate the label for each word, where K is number of topic and N is all word counts in all document.

2.3 Harmonic Mean Method

Goodness-of-fit is an evaluation of how well the trained model can fit to the training data. [3] suggests that harmonic mean could be a way to do the evaluation. Harmonic mean for LDA model in [3] is defined as:

$$\frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{K} \sum_{k=1}^{K} \theta_{m,k} \right)^{-1} \tag{2}$$

where $\theta_{m,k}$ is the topic distribution for document m and topic k. Harmonic mean could measure the degree of differentiation of a distribution θ over all topics. If the topic distribution has high probability for only few topics, then it would have lower harmonic mean value. Namely, the model would have better ability to separate documents into different topics.

2.4 Perplexity

A common criterion to evaluate overfitting of a clustering model is to use perplexity suggested in [2]. Perplexity is defined as the reciprocal geometric mean of the likelihood of testing data given the trained model $\mathbf{M} = \{\phi, \theta\}$. Therefore, the lower perplexity value indicates that the model could fit the testing data better. Perplexity is defined as:

$$p(\mathbf{W}|\mathbf{M}) = \prod_{m=1}^{M} p(w_m|\mathbf{M})^{-\frac{1}{N}} = \exp - \frac{\sum_{m=1}^{M} \log p(w_m|\mathbf{M})}{\sum_{m=1}^{M} N_m} \tag{3}$$

$\log p(w_m|\mathbf{M})$ can be directly expressed as a function of the multinomial parameters:

$$\begin{aligned}\log p(w_m|\mathbf{M}) &= \prod_{n=1}^{N_m}\sum_{k=1}^{K} p(w_n = t|z_n = k)p(z_n = k|d = m) \\ &= \prod_{n=1}^{N_m}\left(\sum_{k=1}^{K}\phi_{k,t}\theta_{m,k}\right)^{n_m^t}\end{aligned} \tag{4}$$

where **W** is all words in testing data, **M** is number of documents, N is number of all words, N_m is number of words in document m, w_m is words in document m, $\phi_{k,t}$ is the word distribution for topic k and word t, $\theta_{m,k}$ is the topic distribution for document m and topic k.

3 Experimental Results

In this section, we introduce two datasets, Classic400 and binary image character, for experiments. We explain our approach to find good parameters to learn the LDA model by inspecting harmonic mean and perplexity. The result of Classic400 is demonstrated by visualizing probability of topics given labels of words. The result of binary image character is shown by visualizing image topics of a character.

3.1 Experiment Setup

We use two datasets in this report. First we use Classic400 which contains 400 documents and average 78.79 words in each document and consists of 6205 vocabulary. The documents are collected from documents with three different scientific topics. Since these three topics are fairly uncorrelated, we assume that 3 topic models are sufficient to capture concepts in the training dataset. We use 90% of all training data for learning LDA model. 150 epochs are used for training to converge. We apply harmonic mean to measure the goodness-of-fit from the training example. The 10% held-out set is used to measure the perplexity that indicates the degree of overfitting.

We also interested in different usage of LDA for image data. We assume characters consist of parts which may be shared with different characters (like vertical line of 1 and L) or very specific to one character. LDA is an appropriate tool to discover the meaningful part structures of characters. So the main goal of this experiment is to discover meaningful topic in the form of images that compose a character. We use binary image character dataset to demonstrate the discovery result. The dataset consists of 36 characters that are 0-9 and A-Z and 39 examples for each character. Each example is a binary image of size 20x16. The visualization of dataset is shwon in Figure 2. We take each training example as a document, so there are 1404 documents. Each pixel in an image is a word which can have word count 0 or 1. For each document, it has 320 binary words. Due to the simplicity of this dataset, we use 20 topics for topic discovery.

3.2 Topic Discovery of Classic400 Dataset

Figure 1 visualizes $\theta_{k,m}$ the probability of topics given a document. Each document is a point in the figure and the all the points lie on plane where $\sum_{k=1}^{3}\theta_{k,m} = 1$. The three sets of hyperparameter we test are ($\alpha = 0.1$, $\beta = 2$), ($\alpha = 0.1$, $\beta = 0.2$), and ($\alpha = 0.01$, $\beta = 2$).

The harmonic mean is 8.58×10^{-5} for Figure 1d. Low harmonic mean indicates high goodness-of-fit for the training data. However, the perplexity of Figure 1d is quite high, which means the learned model is hard to generalize the unseen held-out data. On the other hand, Figure 1c has higher harmonic mean 6.56×10^{-3} and the documents do not look to be well separated. So this model has low goodness-of-fit compare to Figure 1d. Figure 1b has both low harmonic mean and perplexity, which means this model has better goodness-of-fit without overfitting to the training data. The experiment shows the trade-off between harmonic mean and perplexity. When one want to fit the model very well from the given training data, the learned model is very likely to be overfitting to the training data and fail to generalize to unseen testing data. Table 1 shows words with highest probability for each topic with the best model we can get from Figure 1b.

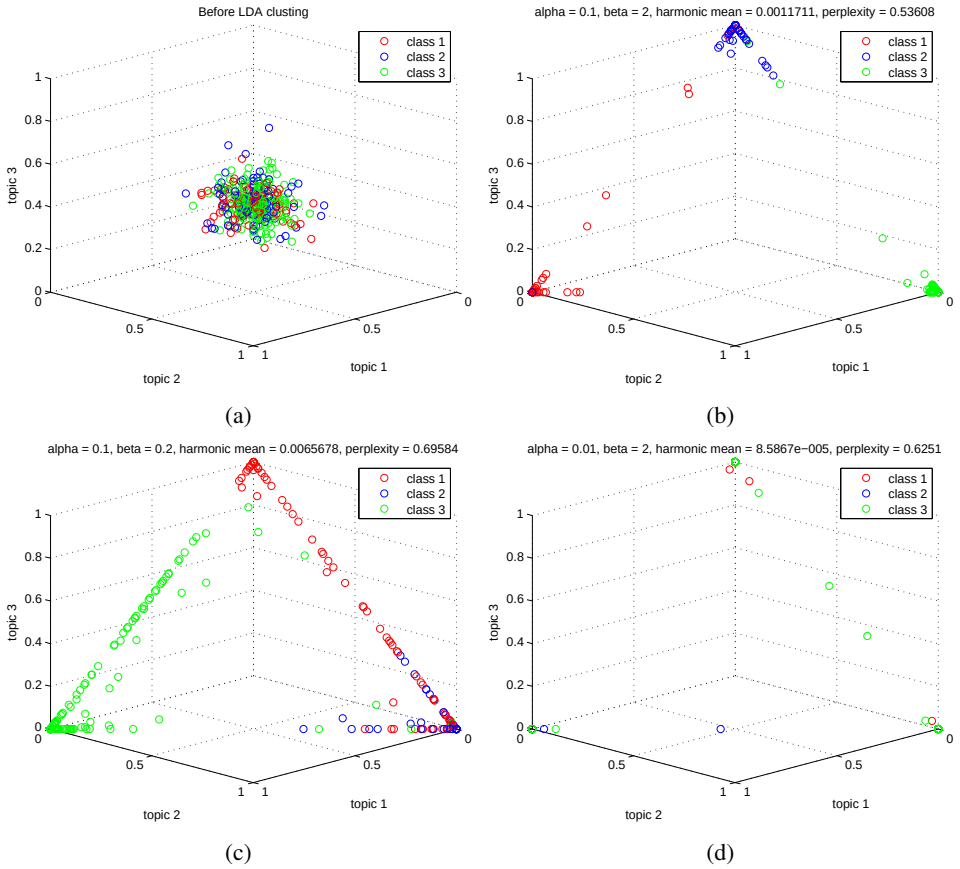


Figure 1: Probability of document given topic with different set of hyperparameters α and β. (a) Initialization of document distribution (b) LDA result with $\alpha = 0.1, \beta = 2$ (c) LDA result with $\alpha = 0.1, \beta = 0.2$ (d) LDA result with $\alpha = 0.01, \beta = 2$. Based on harmonic mean and perplexity, (b) provides high goodness-of-fit and low overfitting.

Topic 1	Topic 2	Topic 3
system	boundary	patients
scientific	layer	ventricular
retrieval	wing	fatty
research	mach	left
science	supersonic	nickel
language	ratio	cases

Table 1: Words with highest probability for each topic with hyperparameters $\alpha = 0.1$, $\beta = 2$.

3.3 Topic Discovery of Binary Image Character Dataset

We use hyperparameter $\alpha = 0.5$ and $\beta = 0.001$ this experiment. Figure 3a-3d visualize discovered topics from random initialization to 30 epochs after. We can see as time progress, the topic groups initial random pixels (words) into some semantically meaningful part structures (topics). Figure 3e provides 10 digits and their top 5 corresponding topics. We can find that topics are very close to parts that compose the digit. The discovered topics provide a nice part feature for human to analyze the composition of character. Another usecase in computer vision task, we can consider probability of learned topics in an image as the feature which may be a good representation for character classification.

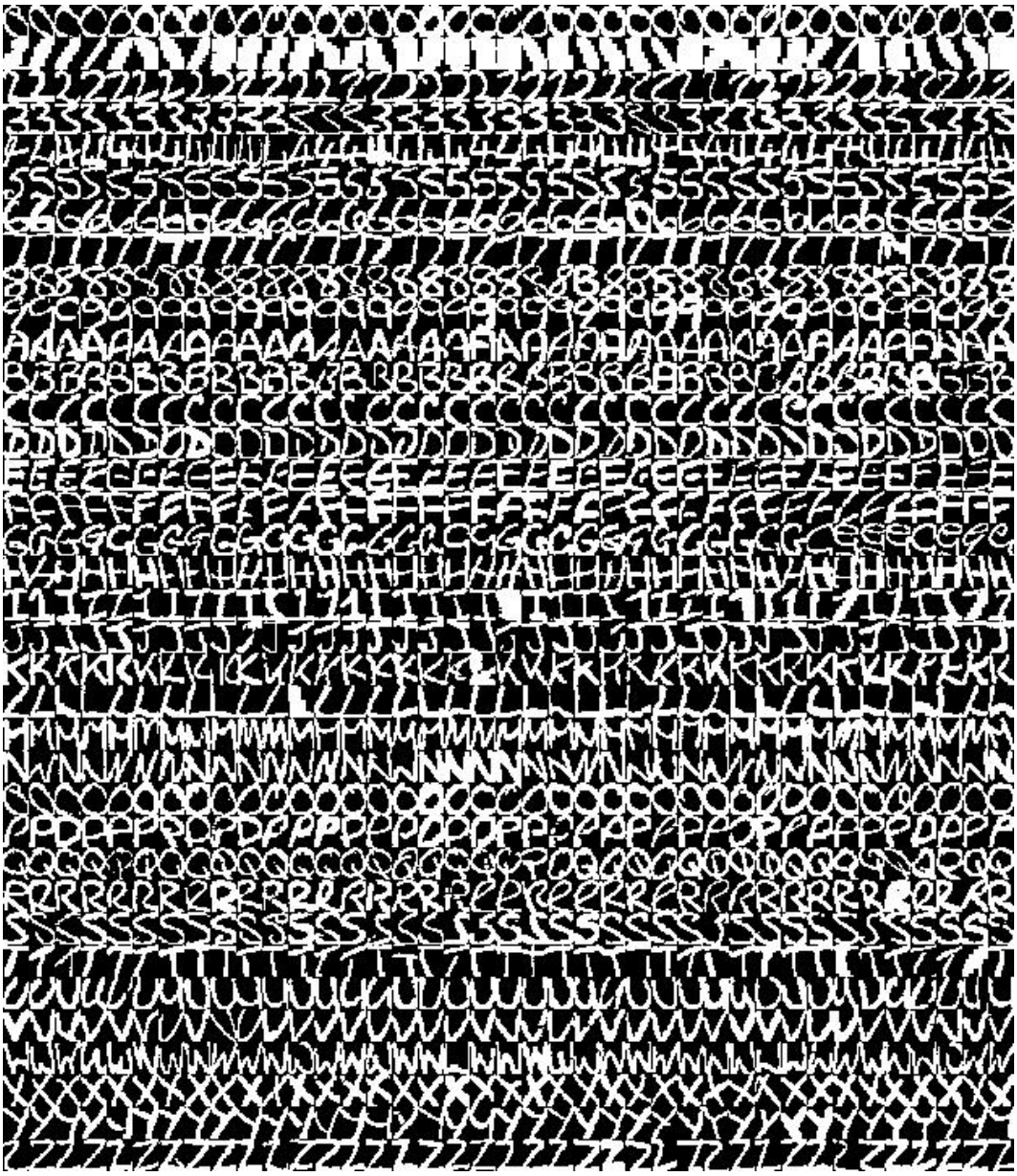
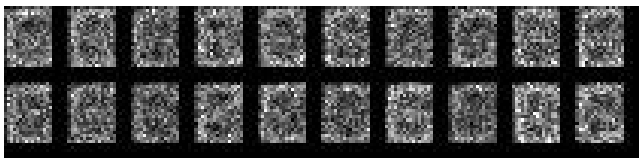


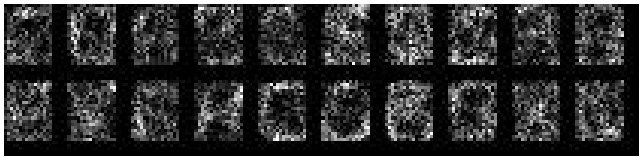
Figure 2: Visualization of binary character image database.

4 Discussion

In this report, we apply LDA to text and image data and visualize the result by showing probability of topic for text data and topic images for image data. The success of mining latent topics suggests the algorithm can be generalized and applied to both text and image data. We introduce harmonic mean and perplexity to select hyperparameters to learn LDA model. We discover the trade-off between harmonic mean and perplexity. The model usually can not have low harmonic mean and low perplexity at the same time. But with cross-validation, one is able to find a balance point where harmonic mean and perplexity both have good scores that indicates our model not only fits the training data well but also generalizes to validating data.

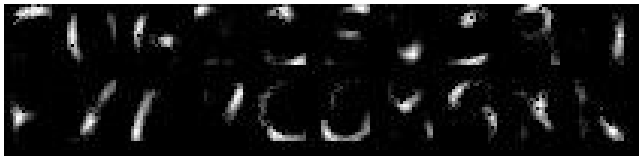


(a) Initial 20 topics.

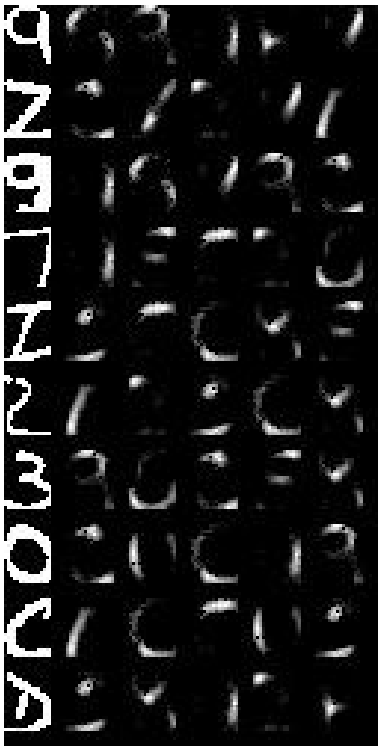


(b) Topics after 5 epochs.

(c) Topics after 10 epochs.



(d) Topics after 30 epochs.



(e) 10 digits and top 5 topics.

Figure 3: Left column shows 20 Topics learned from binary character image database with different training epochs. Right column shows 10 selected digit and their top 5 corresponding topics.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] G. Heinrich. Parameter estimation for text analysis. In *Technical report, vsonix GmbH and university of Leipzig, Germany.*, 2005.

[3] H. M. Wallach, I. Murray, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada.*, 2009.