CSE 515T Bayesian Method in Machine Learning

Project Final Report

Chunyuan Li
Jiarui Xing

Abstract

Topic model represents documents as mixtures of topics that spit out words with certain probabilities, and latent Dirichlet allocation (LDA) is probably the most widely used implementation of topics model. In our work, we explored the possibility of using LDA as a document vectorization method and tested if LDA works better than other vectorization methods on author classification task with different target vector dimension and training dataset size. We also tested how the parameter of prior Dirichlet distribution would affect the performance. The results show that LDA cannot give better result with both low target dimension and small training dataset. We also found that LDA with smaller document-topic distribution parameter α works better, which may due to smaller α will generate more sparse topics and better represent a document.

1. Introduction

LDA is a technique that automatically discovers topics that a set of documents contain. It assumes each document is written in the following way. (1) Decide on the number of words N the document will have; (2) Choose a document-topic distribution for the document; (3) Generate each word in the document by: first picking a topic according to the document-topic distribution that we sampled above, then using the topic to generate the word itself following a topic-word distribution. LDA also assumes that both document-topic distribution and topic-word distribution have prior Dirichlet distributions with different parameters. Figure 1.1 shows the plate notation of LDA model, where:

α is the parameter of the Dirichlet prior on the each document-topic distributions,

β is the parameter of the Dirichlet prior on the each topic-word distribution,

θ_m is the topic distribution for document m,

φ_k is the word distribution for topic k,

z_{mn} is the topic for the n-th word in document m, and

$w_{mn}$is the specific word.

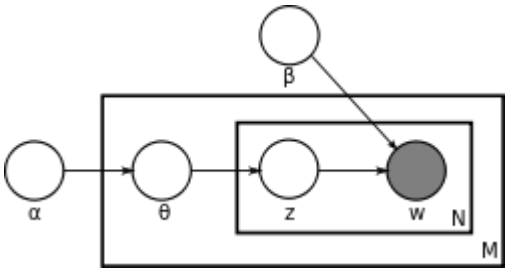

Figure 1.1 Plate notation of LDA model[1]

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to generate the collection. Each document can be seen as a weighted mixture of topics with their own topic distribution.

Based on the idea that LDA can represent a document as distribution of topics, we can take it as an unsupervised document vectorization method and test this idea on document classification tasks. Though we already have a great number of document vectorization method such as simpler ones like

term frequency and tf-idf, or more complex ones like doc2vec, we hope LDA can perform better in the following aspects:

- Low-dimension representation. Term frequency and tf-idf can be seen as a generalized one-hot encoding method, which gives every word a dimension and thus will often generate vectors with thousands of dimensions. On the contrary, each topic generated by LDA will correspond to several words and the number of topics is usually much smaller than number of words, and we hope LDA can generate equally effective vectors with much lower dimension compared with term frequency and tf-idf.
- Robust on small dataset. As one advantage of Bayesian method is the prior can help us avoid overfitting, we also hope LDA preforms better on small dataset.

The experiment part exams the two ideas above, comparing with other document vectorization methods. Furthermore, since the parameters of prior distribution (Dirichlet distribution) may also affect the performance of LDA, we also test the performance using different prior parameters.

2. Experimental Setup

In our experiments, as is shown in figure 2.1, after preprocessing the raw text data, we vectorize the corpus using seven different methods. Then training a multinomial naive Bayes classifier and comparing the performance of cross validation and test accuracies. The vectorization methods we used including:

- Term frequency and tf-idf. These two methods are used as baseline of our task.
- Term frequency and tf-idf with non-negative matrix factorization (NMF). These two methods first apply standard term frequency or tf-idf, then use NMF to reduce feature dimension to the same amount as LDA.
- Term frequency and tf-idf with most frequent terms. These are two more low dimension document vectorization methods. In these two methods we first do a standard term frequency or tf-idf, then simply keep the most frequent terms. In following part of this report we will call these last four methods and LDA as dimension-fixed methods.

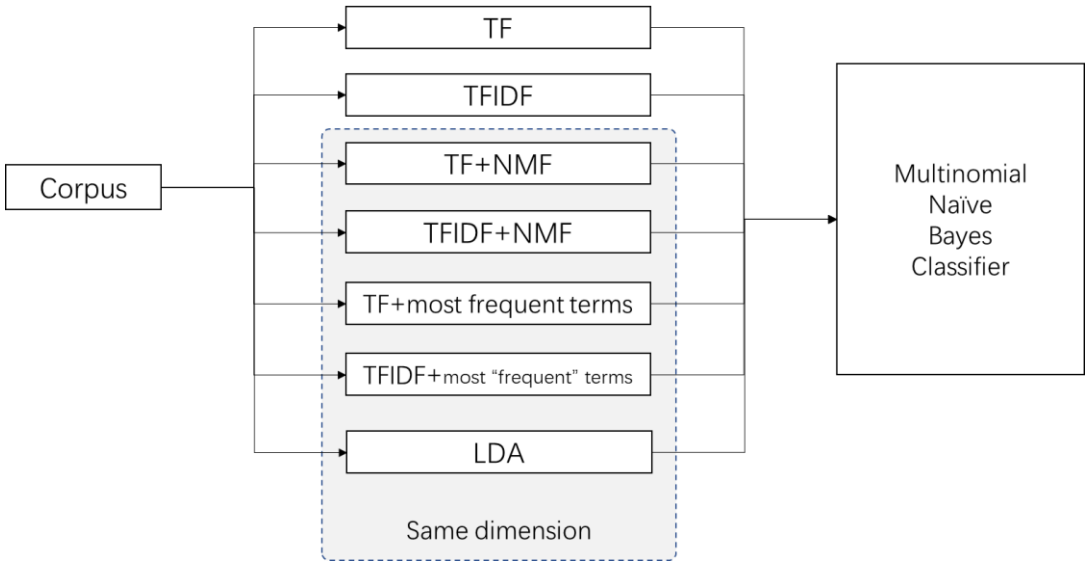

Figure 2.1 Experiment Pipeline

We experimented with two datasets: the state of union address and scripts from *The Simpsons*. In the former dataset we had addresses from 5 presidents, split whole address into several paragraphs,

filtered out too short paragraphs and tried to figure out the speaker of a given paragraph. In the latter one, we selected the script lines of the most active characters and filtered out too short scripts, and then tried to figure out the speaker given a script. However, with the second dataset we got very bad result (lower than 70% accuracy when distinguishing from 2 characters), which can't give useful information. This might be due to that there are many very short scripts that contains few topics and most of the script lines are daily conversations which have little difference in topics. Therefore, in the following sections we present the results base on the state of union address dataset.

3. Experimental Result

3.1 Dimension

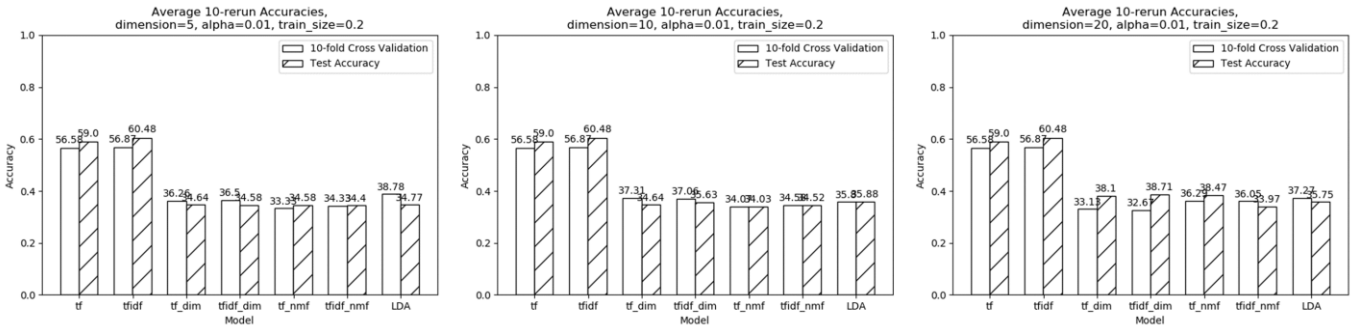

Figure 3.1 Accuracies with different target dimensions

As it is shown in figure 3.1, When the number of dimension increases, LDA's test accuracies keeps nearly unchanged. Moreover, LDA doesn't give better result compared to other dimension-fixed methods and preform much worse than standard term frequency and tf-idf. This suggests that if we want a low dimension document vectorization, it will be better to simply take the most frequent terms in term frequency or tf-idf, since they give fairly accuracies result and are much faster than other methods.

3.2 Dataset size

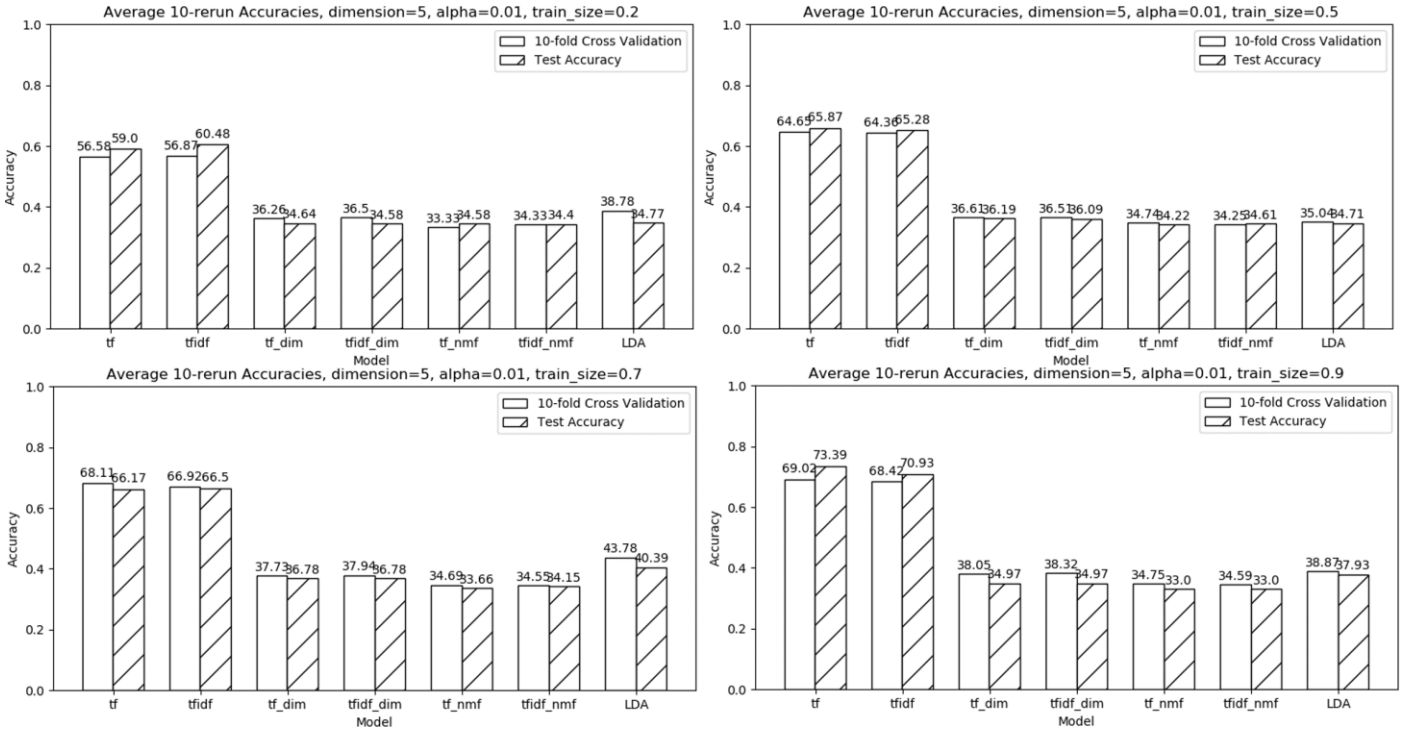

Figure 3.2 Accuracies with different training dataset size

Figure 3.2 shows accuracies with training datasets of different size (for example, train_size=0.2 means we take 20% samples from whole dataset as training set). When training dataset is small (train_size=0.2, 0.5), test accuracies of LDA do not show much difference against other dimension-fixed methods. Furthermore, with larger datasets LDA preforms better than other dimension-fixed methods, which suggests LDA works better on larger datasets. Prior distribution cannot bring significant advantage to LDA when dataset is small.

3.3 Topic Distribution Parameter - Alpha

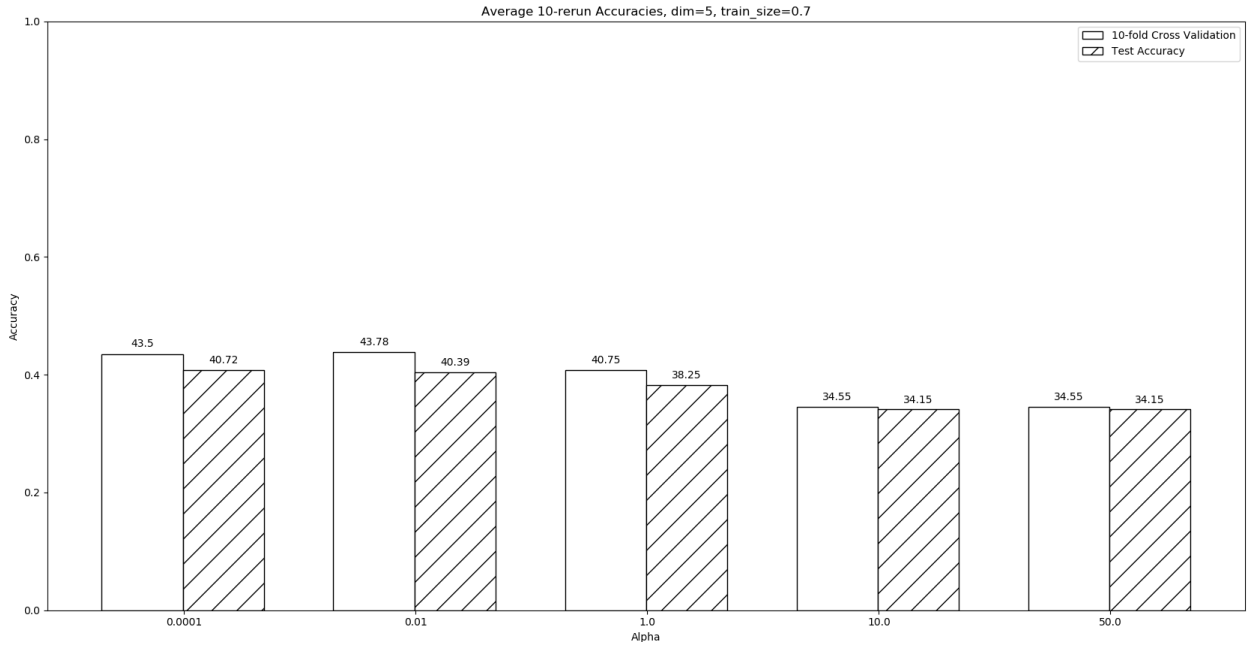

Figure 3.3 Accuracies of LDA with different α values

Figure 3.3 shows accuracies of LDA with different α values. As α increases, both cross validation and test accuracies decreases, which agree with our initial guess.

4. Summary

In our experiments we explored the possibility of LDA as a document vectorization method. The results show that LDA is not a good method for low-dimension document vectorization or vectorization with small dataset. And results also show that LDA do better with smaller α value. That's to say, a sparse distribution of topic can better represent a document.

The standard LDA we used is an unsupervised method. However, since we have labeled data, using the supervision information is likely to improve the performance. For future work, we plan to use supervised LDA trying to find how it performs.

5. References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[2] Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. arXiv preprint arXiv:1711.04305.