Technology Review

CS 410 Fall 2020 - Mack Teng (mackt2)

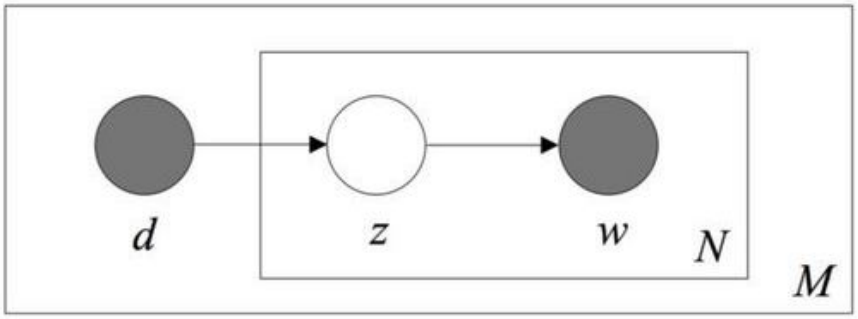
Introduction

Topic models allow us to extract the “hidden” or abstract topics that underlie a collection of documents, which can be used in various applications, including document classification, contextual matching, and recommendation systems based on content. This review will start by going over PLSA, moving on to the motivations for LDA and its inner workings.

Body

PLSA

We recall the plate diagram for PLSA from previous material:



The generative process is as follows:

1. given a document d, topic z is present in that document with probability P(z|d)
2. given a topic z, word w is drawn from z with probability P(w|z)

The probability of seeing a word with a document together is then given as:

$$P(D,W) = P(D)\sum_{Z} P(Z|D)P(W|Z)$$

A shortcoming of PLSA in that there is no good way to determine P(D) for an unseen document. The P(D) in PLSA is determined based on the observed corpus and is a fixed data point, so therefore cannot be used to determine P(D) for a new document not in the training set. Other downsides include the lack of a generative model for the topic mixtures, (as each P(Z|D) is a

separate parameter), which results in PLSA being prone to overfitting issues and increased computational requirements.

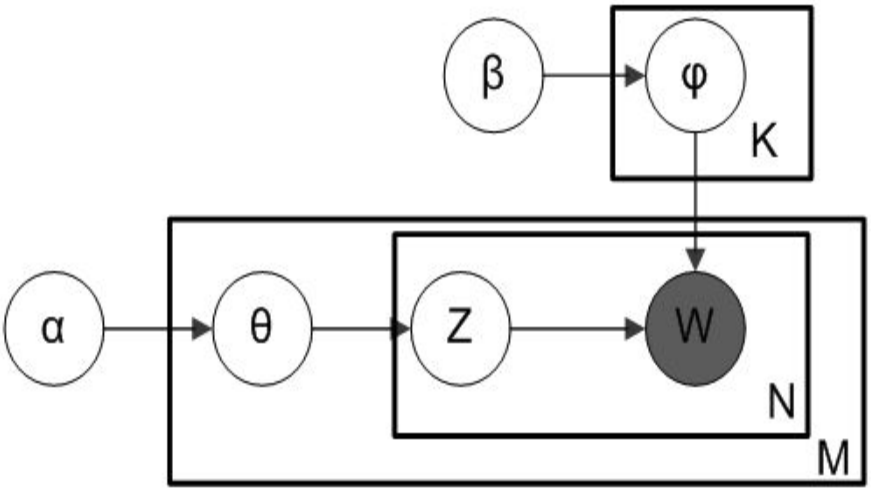
LDA

LDA aims to solve these problems by assuming a Dirichlet prior for the topic document distribution and word topic distribution. A k dimensional Dirichlet distribution can be thought of as a distribution on probability vectors with k elements which sum up to 1. This is how we encode the following intuitions:

1. A document is likely to be concentrated around a smaller number of topics.
    a. This is intuitive because most people write about a single topic in a single document. For example, it is very unlikely that you will find a document that is about science, sports, politics and fashion all at once.
2. A topic is likely to be concentrated around a smaller number of words.
    a. This is intuitive because inherently a topic is about a select relatively few number of words in the vocabulary (related groups of words).

A dirichlet prior helps with this because we can adjust the hyperparameters of the Dirichlet distribution to have higher probability densities for certain probability multinomials that reflect these intuitions. In other words, given 4 topics A, B, C and D, we can now model the question “How likely is a document going to exhibit 25% A, 25% B, 25%C and 25%?”.

The plate notation for LDA:



The generative process of LDA for a document is assumed as:

1. Choose N words for your document
2. Sample a topic distribution for your document

3. For each word n in N, choose a topic based on the topic distribution, and from that topic, pick a word based on a sampled word distribution for said topic.

α is the parameter for the dirichlet distribution of the topic document distributions which is how we control the topic concentration for a document. The higher the value of α, the more topics documents are composed of, and the lower the value of α, the fewer.

β is the parameter for the dirichlet distribution of the word topic distributions which is how we control the word concentration for a topic. The higher the value of β is, the more words topics are composed, and the lower value of β, the fewer.

Θ and φ are respectively the document topic distribution and topic word distribution sampled from the dirichlet distributions. Z represents the topic assignment for a given (document, word) pair.

Estimation

Fitting an LDA model onto a corpus is essentially trying to answer the question, which configurations of these parameters (Θ , φ, and Z) result in the highest probability of the corpus being generated?

There are different ways to estimate this, but a common one is by collapsed Gibbs sampling, which samples each document, word-topic assignment individually assuming other assignments are correct:

$$p(z_{d,n} = k|\vec{z}_{-d,n}, \vec{w}, \alpha, \lambda) = \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Credit: Jordan Boyd-Graber

This will result in higher probabilities for a topic z for a document word assignment, if the document is topic z heavy, or if the topic z favors said word.

Conclusion

LDA improves over PLSA by imposing a dirichlet prior on the topic-document and word-topic distributions. Over the years, there have been many extensions and applications of LDA for domain specific-tasks. Anywhere a document, word analogy can be drawn, LDA can be used to extract abstract topics over the “word”. For example, in bio-informatics, gene K-mers can be thought of “words” in a gene sequence (“document”).

References

1. Topic Modeling with LSA, PLSA, LDA & lda2Vec
2. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling