LDA Tutorial

Chris Tufts

Contents

Preface

The purpose of this book is to provide a step by step guide of LDA utilizing Gibbs Sampling. It is heavily inspired by the Gregor Heinrich's *Parameter Estimation for Text Analysis*(Heinrich 2008).

Background

I've often found that resources covering LDA are either hard to follow, do to a heavy reliance on calculus derivations, or the resources are extremely high level so you get the idea of what LDA accomplishes, but you never fully grasp how it works. The book focuses on LDA for inference via Gibbs Sampling. To aid in understanding both LDA and Gibbs sampling all probability distributions used in LDA will be reviewed along with a variety of different approaches for parameter estimation. Following the introduction of these components, LDA will be presented as a generative model. This will lay the groundwork for understanding how LDA can be used for inference of topics in a corpus.

Layout of Book

- What is LDA? - High level overview
- Parameter Estimation Methods (general overview using Bernoulli distribution) - ML, MAP, Bayesian Inference, Gibbs Sampling
- Multinomial Distribution - Explains the relationship of Bernoulli to Multinomial, then goes into Gibbs Sampling
- Word/Document Structures - Bag of Words, Word Document Matrix
- LDA - A Generative Model
- LDA - Inference

Chapter 1

What is LDA?

Latent Dirichlet Allocation (LDA) is a generative probablistic model for collections of discrete data developed by Blei, Ng, and Jordan. (Blei, Ng, and Jordan 2003) One of the most common uses of LDA is for modeling collections of text. The general idea is that each document is generated from a mixture of topics and each of those topics is a mixture of words. This is known as a hierarchical model since it is built on distributions of topics built on top of distributions of words.

In regards to the model name, you can think of it as follows:

- Latent: Topic structures in a document are 'latent' meaning they are hidden structures in the text.
- Dirichlet: The Dirichlet distribution is used as the prior for both the topic distributions and the word distributions. If you have no idea what a Dirichlet distribution or a prior is do not despair. We will be going through all of this in the upcoming chapters.
- Allocation: Allocation of words to a given topic.

So to review: we have latent structures in a corpus (topics), taking into account Dirichlet priors for the word and topic distributions, to allocate words to a given topic and topics to a given document.

Throughout this book I will work through all of the building blocks which make LDA possible, but to help get an understanding of what LDA is and why it is useful, I will offer a quick example first.

1.1 Animal Generator

The majority of this book is about words, topics, and documents, but lets start with something a bit different: animals and where they live. One of the ways you can classify animals is by where they spend the majority of their time - land, air, sea. Obviously there are some animals that only dwell in one place, for example a cow only lives on land. However, there are other animals, such as some birds, that split their time between land, sea, and air.

You are probably asking yourself where I'm going with this. We can think of land, air, and sea as topics that contain a distribution of animals. In this case we can equate animals with words. For example, on land I am much more likely to see a cow than a whale, but in the sea it would be the reverse. If I quantify these probabilities into a distribution over all the animals (words) for each type of habitat (land,sea, air - topics) I can use them to generate sets of animals(words) to populate a given location (document) which may contain a mix of land, sea, and air (topics).

So let's move on to generating a specific location. We know that different locations will vary in terms of which habitats are present. For example, a beach contains land, sea, and air, but some areas inland may only contain air and land like a desert. We can define the mixture of these types of habitats in each location. For example, an example beach is 1/3 land, 1/3 sea, and 1/3 air. We can think of the beach as a single

Table 1.1: Animal Distributions in Each Habitat

vocab	land	sea	air
	0.00	0.12	0
	0.00	0.12	0
	0.00	0.12	0
	0.00	0.12	0
	0.00	0.12	0
	0.05	0.06	0
	0.05	0.06	0
	0.05	0.06	0
	0.05	0.06	0
	0.10	0.00	0
	0.10	0.00	0
	0.05	0.06	1
	0.05	0.06	0
	0.10	0.00	0
	0.10	0.00	0
	0.10	0.00	0
	0.10	0.00	0
	0.10	0.00	0

document. To review: a given location (document) contains a mixture of land, air, and sea (topics) and each of those contain different mixtures of animals (words).

Let's work through some examples using our animals and habitats. The examples provided in this chapter are oversimplified so that we can get a general idea of what is going on with LDA. The rest of the book will handle all the nuts and bolts of the model, but for now let's try and get a handle on how this works.

We'll start by generating a beach location with 1/3 land animals, 1/3 sea animals, and 1/3 air animals. Below you can see our collection of animals and their probability in each topic. Note that some animals have zero probabilities in a given topic, i.e. a cow is never in the ocean, where some have higher probabilities than others (a crab is in the sea sometimes, but a fish is always in the sea). You may notice that there is only 1 animal in the air category. There are several birds, but only 1 of them is cabable of flight in our vocabulary.

(NOTE: These are the probability of a word given the topic and therefore the probabilities of each habitat(column) sum to 1.)

To generate a beach (document) based off the description we would use those probabilities in a straightforward manner:

```
words_per_topic <- 3
equal_doc <- c(vocab[sample.int(length(vocab),words_per_topic, prob=phi_ds$land, replace = T)],
               vocab[sample.int(length(vocab),words_per_topic, prob=phi_ds$sea, replace = T)],
               vocab[sample.int(length(vocab),words_per_topic, prob=phi_ds$air, replace = T)])
cat(equal_doc)
```

```
##
```

NOTE: In the above example the topic mixtures are even, so each habitat (topic) contributes 3 animals to the beach.

Ok, now let's make an ocean setting. In the case of the ocean we only have sea and air present, so our topic distribution in the document would be %50 sea, %50 air, and %0 land.

Table 1.2: Animals at the First Two Locations

Document	Animals
1	
2	

Table 1.3: Distribution of Habitats in the First Two Locations

Document	Land	Sea	Air
1	0.3579681	0.4060102	0.2360217
2	0.2460988	0.0981821	0.6557192

```
words_per_topic <- 3
ocean_doc <- c(vocab[sample.int(length(vocab),words_per_topic, prob=phi_ds$sea, replace = T)],
               vocab[sample.int(length(vocab),words_per_topic, prob=phi_ds$air, replace = T)])
cat(ocean_doc)
```

```
##
```

NOTE: In the example above only the air and land contribute to the ocean location. Therefore they both contribute an equal number of animals to the location.

1.2 Inference

We have seen that we can generate collections of animals that are representative of the given location. What if we have thousands of locations and we want to know the mixture of land, air, and sea that are present? And what if we had no idea where each animal spends its time? LDA allows us to infer both of these peices of information. Similar to the locations (documents) generated above, I will create 1000 random documents with varying length and various habitat mixtures.

The habitat (topic) distributions for the first couple of documents:

With the help of LDA we can go through all of our documents and estimate the topic/word distributions and the topic/document distributions.

This is our estimated values and our resulting values:

The document topic mixture estimates are shown below for the first 5 documents:

```
##      1    2    3
## 1 0.31 0.37 0.32
## 2 0.18 0.09 0.73
## 3 0.64 0.25 0.11
## 4 0.22 0.46 0.32
## 5 0.41 0.25 0.34
## 6 0.50 0.30 0.20
```

Here are our real mixtures for comparison:

Table 1.4: Estimated word distribution for each topic

	Topic 1	Topic 2	Topic 3
	0.01	0.12	0.01
	0.00	0.12	0.01
	0.01	0.12	0.01
	0.00	0.12	0.01
	0.00	0.14	0.01
	0.05	0.07	0.00
	0.04	0.08	0.00
	0.03	0.08	0.01
	0.04	0.06	0.01
	0.08	0.00	0.00
	0.08	0.00	0.00
	0.12	0.04	0.90
	0.05	0.05	0.01
	0.10	0.00	0.00
	0.09	0.00	0.00
	0.09	0.00	0.00
	0.10	0.00	0.00
	0.10	0.00	0.00

Table 1.5: The word distribution for each topic used to build the documents

	land	sea	air
	0.00	0.12	0
	0.00	0.12	0
	0.00	0.12	0
	0.00	0.12	0
	0.00	0.12	0
	0.05	0.06	0
	0.05	0.06	0
	0.05	0.06	0
	0.05	0.06	0
	0.10	0.00	0
	0.10	0.00	0
	0.05	0.06	1
	0.05	0.06	0
	0.10	0.00	0
	0.10	0.00	0
	0.10	0.00	0
	0.10	0.00	0
	0.10	0.00	0

Table 1.6: The Real Topic Distributions for the First 5 Documents

Land	Sea	Air
0.36	0.41	0.24
0.25	0.10	0.66
0.58	0.27	0.16
0.16	0.58	0.26
0.48	0.19	0.32
0.40	0.40	0.20

Chapter 2

Parameter Estimation

LDA is a generative probabilistic model, so to understand exactly how this works we need to understand the underlying probability distibutions. In this chapter we will focus on the Bernoulli distribution and the Beta distribution. Both of these distributions are very closely related to (and also special cases of) the multinomial and Dirichlet distributions utilized by LDA, but they are a bit easier to comprehend. Once we have made our way through Bernoulli and beta, the following chapter will go into detail about multinomial and Dirichlet distributions and how all these peices are connected.

Throughout the chapter I'm going to build off of a simple example - a single coin flip. Let's begin.

2.1 Distributions

2.1.1 Bernoulli

When you flip a coin you get either heads or tails. This single coin flip is known as a Bernoulli trial. You can think of any single trial with two possible outcomes as a Bernoulli trial. The Bernoulli distribution is the probability distribution of Bernoulli trials, basically a model of the single coin flip.

2.1.1.1 Bernoulli: A Special Case of the Binomial Distribution

You will often see Bernoulli distribution mentioned as a special case of the Binomial distribution. The binomial model consists of n bernoulli trials, where each trial is independent and the probability of success does not change between trials.(Kerns 2010). The bernoulli distribution is the case of a single trial or n=1.

To clarify, if I want to calculate the probability of getting heads on a single coin flip I will use a bernoulli distribution. However, if I want to know the probability of getting 2 heads (or more) in a row, this is where the binomial distribution comes in. For our purposes we are only concerned about the outcome of a single coin flip and will therefore stick to the Bernoulli distribution.

2.1.1.2 Bernoulli - Distribution Notation

The probability mass function of the bernoulli distribution is shown below.

Probability Mass Function

$$f_x(x) = P(X = x) = \theta^x(1-\theta)^{1-x}, \qquad x = \{0, 1\} \tag{1}$$

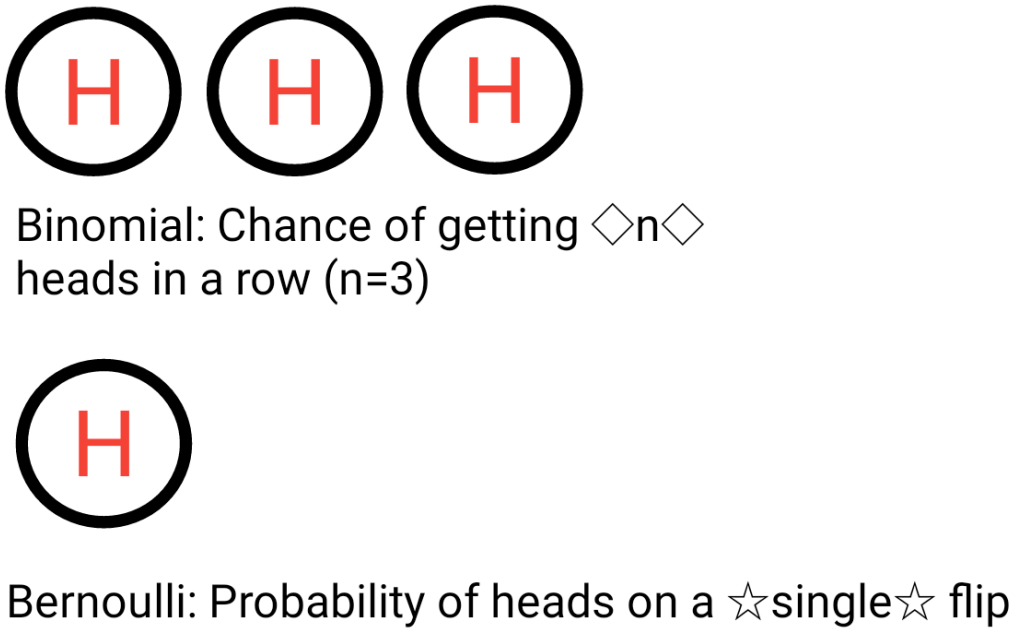

Figure 2.1: Bernoulli and Binomial Distributions