Topic modeling

What is this document *about*?

Latent Dirichlet allocation

- before we can model topics, we need to model documents
  → bag of words
    - order doesn't matter
- but, which words? Words drawn from topics.

- A document is associated with topics, each topic is associated with words.
- What do we mean by "associated with"? A multinomial distribution.

word distribution for topic "walking":

multinominal distribution

word	proportion
walk	0.18
stroll	0.09
amble	0.04
slog	0.04
step	0.12
...	0.53

$$p(w_1, w_2, ...) = \frac{n!}{w_1!w_2!...}\beta_1^{w_1}\beta_2^{w_2}...$$

w:counts of words

beta:word proportions within the topics

draw topics for a document and then for each topic we can draw words

topic distribution for document "lecture":

```
 topic           | proportion
-----------------+-----------------
 words           | 0.20
 topics          | 0.22
 documents       | 0.17
 walking         | 0.09
 dragons         | 0.01
 ...             | 0.31
```

if we draw the word ‘walk’ 18% of the time from walking topic
and the walking topic compose 9% of the documents
what proportion of the words in the document are ‘walk’ —> at least 18%*9%
(if ‘walk’ is included as part of an of these other topics

$$p(z_1, z_2, \ldots) = \frac{m!}{z_1!z_2!\ldots}\theta_1^{z_1}\theta_2^{z_2}\ldots$$

generate document:
what word should be said first in this document?
1) first we need to know what topic that word is going to be from—draw a topic—-p(z)
2) then, draw a word from p(w) based on p(z)
E.g. walking —> step

what we really want to do is not to generate these silly documents,
but given some real document to infer the topics that make it up
—> infer the entire distribution

- How do we model a *corpus* of documents?
- A distribution over topic proportions. 0) what we need to do before 1) and 2)
- A conjugate prior for the θ_i parameters of the multinomial...
- A Beta/Dirichlet distribution!

when we are drawing a topic, we are drawing it from a multi nominal distribution, which has parameters
Topic distribution: essentially drawing those parameters randomly

- Beta distribution

$$p(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}$$

- reformatted

$$p(\theta_1, \theta_2) = \frac{1}{B(a,b)} \theta_1^{a-1} \theta_2^{b-1} \mid \theta_1 + \theta_2 = 1$$

- Dirichlet distribution

we still need alpha and beta for the whole process

$$p(\theta) = \frac{1}{B(\alpha)} \prod_i \theta_i^{\alpha_i} \mid \sum_i \theta_i = 1$$

allow you to draw a distribution
a vector of probabilities that sum to one —> do 0)

if we have five topics, we can make a distribution of dimension 5 and we draw this five vectors out(probability sum to one)
which we take as out topic proportions (hypo parameters for a multinomial distribution)

- The words for the document are chosen from the appropriate topics.
- What do we mean "chosen"? Drawn from the appropriate distribution.
- LDA is a generative model.

Generative process

```
for each document
    draw topic proportions from Dirichlet
    [optional] draw a number of words from Poisson/Gaussian/uniform/etc.
    for each word    each word we want to write and we haven’t chosen yet
        draw a topic from document topic proportions
        draw a word from topic word proportions
```

the likelyhood of each topics(now we have 3 topics)

sampling from Dirichlet

```
import numpy as np

a = [2.1, 3.2, 4.3]
n = 10000
p = np.random.dirichlet(a, n)
```

each row is a document
n: number of documents
each row sums to 1

we can control the concentration of the probability may(for example when they are all equally likely)
a = np.array([1,1,1,1])*100 —> tends to distribute that probability evenly over the topics
a = np.array([1,1,1,1])*0.01 —-> tends to concentrate that probability in one topic

put in that alpha vector, we've decided how many topics we wanted to try to learn we assign those alphas to try again to get how many different topics we expect to be represented in a document and how their relative frequencies of (we expect some topics to be more popular than others, we can encode that information as well.)

posterior distribution is intractible

solution methods:

- Laplace approximation
- Markov chain Monte Carlo (MCMC) methods
- variational Bayes
- and more!

1:08:00左右是exercise的介绍

1) we made the bag of words assumption and it’s wrong
to the extent that the order of the words in the documents is reflective of the topic
we’ve lost that information along those lines

2) it’s not going to do as good a job capturing rare topics just because it’s sort of data quality issue
the way we sets that alpha, ideally it would be attuned to that
so the closer we can get that hypo parameter(that alpha) to the truth, the better

What are the shortcomings of the LDA model?

think about what modeling assumptions do we make and how would we expect those two impacts our topic modeling performance