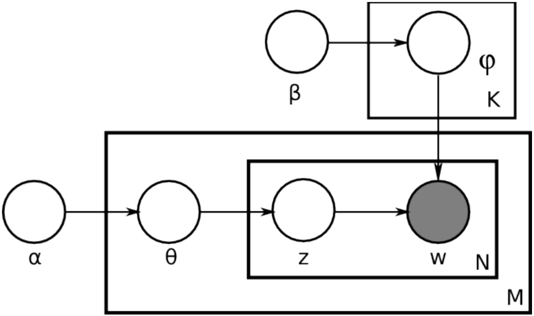
Lecture 19: Probabilistic topic models II: LDA (part 2)

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 19

LDA diagram and process



1. Choose term probabilities for each topic: $\Phi_i \sim \mathcal{D}(\beta)$
2. Choose topic probabilities for each document: $\Theta_d \sim \mathcal{D}(\alpha)$
3. Choose the topic of each token: $z_{dn} \sim \mathcal{M}(\theta_d)$
4. Choose the token: $w_{dn} \sim \mathcal{M}(\phi_{z_{dn}})$

The multinomial distribution

- Let X be a multinomial random variable
- A "realization" of X takes on k distinct values, $\{X_1, \ldots, X_i, \ldots, X_k\}$
- X has $k+1$ parameters:
  - $n > 0$, the number of "trials"
  - $\{p_1, \ldots, p_k\}$, the probability of each distinct value at each trial
    - $0 \leq p_i \leq 1$
    - $\sum_{i=1}^{k} p_i = 1$
    - i.e. $\mathbf{p}$ is a probability distribution over k values
- $X_j \in \{0, 1, \ldots, n\}$
- $\sum_{X_i} = n$
- Intuition
  - Roll a biased k-sided dice n times
  - Count number of times each face turns up
  - X_i is the number of times the i'th face turns up
- There is a formula but we don't need to worry about it!

The Dirichlet distribution

- Let ψ be a Dirichlet random variable
- A "realization" of ψ takes on k values, $\{\psi_k, \ldots, \psi_i, \ldots, \psi_k\}$
  - $0 \leq \psi_i \leq k$
  - $\sum_{i=1}^{k} \psi_i = 1$
  - i.e. a realization of ψ is itself a probability distribution
- ψ has k parameters, $\alpha = \{\alpha_1, \ldots, \alpha_k\}$, with $\alpha_i > 0$
- Let $A = \sum \alpha_i$
- The expected value of ψ_i is α_i / A (written $\mathbb{E}[\psi_i] = \alpha_i / A$)
- The greater A, the closer ψ_i is likely to be to α_i / A
- A realization of ψ can give us the parameters $\{p_1, \ldots, p_k\}$ for a multinomial variable X
- If $\alpha_i = \alpha_j \, \forall i, j$, we say that ψ is symmetric
- There is a formula but we don't need to worry about it!

The probability equations of LDA

$$
\begin{aligned}
P(\mathbf{w}_1, \ldots, \mathbf{w}_n | \alpha, \beta) &= \prod_{i=1}^{n} P(\mathbf{w}_i | \alpha, \beta) \\
P(\mathbf{w}_i | \alpha, \beta) &= \int P(\mathbf{w}_i | \alpha, \phi) P(\phi | \beta) \, \mathrm{d}\beta \\
P(\mathbf{w}_i | \alpha, \phi) &= \int P(\mathbf{w}_i, \theta_i | \alpha, \phi) \, \mathrm{d}\theta_i \\
P(\mathbf{w}_i, \theta_i | \alpha, \phi) &= P(\mathbf{w}_i | \theta_i, \phi) P(\theta_i | \alpha) \\
P(\mathbf{w}_i | \theta_i, \phi) &= \prod_{j=1}^{m} P(w_{i,j} | \theta_i, \phi)^{c_{i,j}} \\
P(w_{i,j} | \theta_i, \phi) &= \sum_{k=1}^{K} P(z_{i,j} = k | \theta_i) P(w_{i,j} | \phi_k)
\end{aligned}
$$

K number of topics

$\mathbf{w}_i$ bag of words (terms and $f_{d,t}$) for document i

The components: θ_i

- θ_i is the multinomial distribution over topics for document i
- There are K topics (where K is semi-arbitrarily chosen by us)
- Therefore θ_i has K parameters
- $\theta_{i,k}$ is the probability that an arbitrary word in document i will belong to topic k
- α is the prior to θ
  - That is, the probabilities $\{\theta_{i,1}, \ldots, \theta_{i,K}\}$ are a "realization" of a Dirichlet random variable with parameters $\{\alpha_1, \ldots, \alpha_K\}$
- The same Dirichlet RV is prior to all θ_i
- α is asymmetric, meaning we allow certain topics to be *a priori* more likely than others

The components: ϕ_k

- ϕ_k is the multinomial distribution over terms for topic k
- There are $m = |V|$ terms in the vocabulary
- Therefore ϕ_k has m parameters
- $\phi_{k,i}$ is the probability that anarbitrary word produced by topic k will be V_i
- β is the Dirichlet prior to ϕ
    - That is, the probabilities $\{\phi_{i,1}, \ldots, \phi_{i,m}\}$ are a "realization" of a Dirichlet random variable with parameters $\{\beta_1, \ldots, \beta_m\}$
- β is symmetric, meaning that all words are *a priori* as likely for each topic
    - This does *not* mean that
        - the posterior distribution ϕ_k over terms for topics will be flat
        - each ϕ_k will give the same distribution over terms

Symmetric and asymmetric priors

Why is β symmetric, but α asymmetric? Following Wallach et al., 2009

- Asymmetric α leads to more stable results
- In particular, models are more stable to choice of number of topics K
  - Think back to LSI, where some topics are "more important" than others
  - And model for topic $k \leq K$ is independent of choice of K
- However, asymmetric β does not improve stability
- Asymmetric requires more parameters to fit than symmetric
- Therefore, only employ asymmetric priors if they provide some advantage

(In the full/pure Bayesian model, we apply yet another prior to α, known as a gamma prior; in practice, this is approximated using empirical methods.)

Deciphering the formulae: $P(w_{i,j}|\theta_i, \phi)$

$$P(w_{i,j}|\theta_i, \phi) = \sum_{k=1}^{K} P(z_{i,j} = k|\theta_i)P(w_{i,j}|\phi_k) \qquad (1)$$

- $P(w_{i,j})$ is the prob that an arbitrary term in doc i is the term j
  - We don't care about the word position in the doc
  - . . . a standard assumption of the unigram term model
- ϕ_k is the probability distribution over terms for topic k
- Therefore, $P(w_{i,j}|\phi_k)$ is just $P(j|\phi_k)$
- $z_{i,j}$ is the topic that generates term j of document i
- $P(z_{i,j} = k|\theta_i)$ is the probability that this topic is k
- θ_i is the prob dist over topics for document i
- Therefore, $P(z_{i,j} = k|\theta_i)$ is just $P(k|\theta_i)$

Deciphering the formulae: $P(w_{i,j}|\theta_i, \phi)$ (cont.)

$$P(w_{i,j} = t|\theta_i, \phi) = \sum_{k=1}^{K} P(z_{i,j} = k|\theta_i) P(w_{i,j} = t|\phi_k) \tag{2}$$

- $P(z_{i,j} = k|\theta_i)P(w_{i,j}|\phi_k)$ is the probability that the term is j and the topic is k
- The term must come from exactly one topic
- Therefore, we sum these probabilities over all K topics
- And this gives us $P(w_{i,j}|\theta_i, \phi)$
  - That is, the probability that an arbitrary term in doc i is j

Deciphering the formulae: $P(\mathbf{w}_i|\theta_i, \phi)$

$$P(\mathbf{w}_i|\theta_i, \phi) = \prod_{j=1}^{m} P(w_{i,j}|\theta_i, \phi)^{c_{i,j}} \tag{3}$$

- $c_{i,j}$ is the number of times that term j occurs in document i
- We assume (unigram model) that these occurrences are independent
- Therefore the probability that term j occurs $c_{i,j}$ times is the probability of each occurrence, raised to the $c_{i,j}$'th power
- We also assume (unigram model) that the occurrence of different terms is independent
- Therefore, the probability of the bag-of-words representation $\mathbf{w}_i$ of document i is just the product of all the individual probabilities

Deciphering the formulae: $P(\mathbf{w}_i, \theta_i | \alpha, \phi)$

$$P(\mathbf{w}_i, \theta_i | \alpha, \phi) = P(\mathbf{w}_i | \theta_i, \phi) P(\theta_i | \alpha) \tag{4}$$

- $P(\mathbf{w}_i, \theta_i | \alpha, \phi)$ is the (joint) probability of observing:
  - the bag-of-words representation $\mathbf{w}_i$ of document i
  - document i's distribution over topics θ_i
- We have previously figured out $P(\mathbf{w}_i | \theta_i, \phi)$
- The multinomial θ_i is an "observation" of the Dirichlet RV α
- So $P(\theta_i | \alpha)$ is the prob of the multinomial θ_i given the prior α
  - There is a formula for this (that we won't worry about!)
- We assume conditional indepence of $\mathbf{w}_i$ and θ_i
- So their joint probability is just the product of their individual (marginal) probabilities

Deciphering the formulae: $P(\mathbf{w}_i|\alpha,\phi)$

$$P(\mathbf{w}_i|\alpha,\phi) = \int P(\mathbf{w}_i,\theta_i|\alpha,\phi)\,\mathrm{d}\theta_i \tag{5}$$

- $P(\mathbf{w}_i|\alpha,\phi)$ is
  - the probability of the bag-of-words document i
  - given our prior for document distributions over topics α
  - and the (list of k) topic distributions over words ϕ
- θ_i is the document distribution over topics for doc i
- We already have $P(\mathbf{w}_i,\theta_i|\alpha,\phi)$ (from previous step)
- We can "remove" θ_i by integrating over all θ_i's
  - If you're unfamiliar with calculus, think of the integral as analogous to a sum over a continuous variable

Deciphering formulae: $P(\mathbf{w}_i|\alpha, \beta)$

$$P(\mathbf{w}_i|\alpha, \beta) = \int P(\mathbf{w}_i|\alpha, \phi)P(\phi|\beta)\,\mathrm{d}\beta \tag{6}$$

- β is our (symmetric) prior for topic distributions over terms
- We've already calculated $P(w_i|\alpha, \phi)$
- We can write:

$$P(\mathbf{w}_i|\alpha, \beta) = \int P(\mathbf{w}_i|\alpha, \beta, \phi)P(\phi|\alpha, \beta)\,\mathrm{d}\beta \tag{7}$$

  by the law of total probability
  - Analogous to:

$$P(A) = \sum_b P(A|B = b)P(B = b) \tag{8}$$

- w_i is independent of β, given θ, and ϕ is independent of α
- Therefore, Equation 7 simplifies to Equation 6

Deciphering the formulae: $P(\mathbf{w}_1, \ldots, \mathbf{w}_n | \alpha, \beta)$

$$P(\mathbf{w}_1, \ldots, \mathbf{w}_n | \alpha, \beta) = \prod_{i=1}^{n} P(\mathbf{w}_i | \alpha, \beta) \tag{9}$$

- We assume that documents are probabilistically independent
- Therefore the probability of generating a set of documents $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$
- ... is the product of the probability of generating each individual document

The probability equations of LDA

$$
\begin{aligned}
P(\mathbf{w}_1, \ldots, \mathbf{w}_n | \alpha, \beta) &= \prod_{i=1}^{n} P(\mathbf{w}_i | \alpha, \beta) \\
P(\mathbf{w}_i | \alpha, \beta) &= \int P(\mathbf{w}_i | \alpha, \phi) P(\phi | \beta) \, \mathrm{d}\beta \\
P(\mathbf{w}_i | \alpha, \phi) &= \int P(\mathbf{w}_i, \theta_i | \alpha, \phi) \, \mathrm{d}\theta_i \\
P(\mathbf{w}_i, \theta_i | \alpha, \phi) &= P(\mathbf{w}_i | \theta_i, \phi) P(\theta_i | \alpha) \\
P(\mathbf{w}_i | \theta_i, \phi) &= \prod_{j=1}^{m} P(w_{i,j} | \theta_i, \phi)^{c_{i,j}} \\
P(w_{i,j} | \theta_i, \phi) &= \sum_{k=1}^{K} P(z_{i,j} = k | \theta_i) P(w_{i,j} | \phi_k)
\end{aligned}
$$

Solving LDA

Directly solving LDA would involve finding parameters that maximize the empirical likelihood $\mathcal{L}$ of the observed documents $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$:

$$\mathcal{L} = \prod_{j=1}^{m} \prod_{i=1}^{n} P(w_{i,j}|z_{i,j}, \phi) P(z_{i,j}|\theta_i) P(\theta_d|\alpha) P(\phi|\beta) \qquad (10)$$

- Note: parameters found are not α, β
- Rather, they are parameters to ϕ, θ

These parameters cannot be directly solved as $z_{i,j}$ not observed.

Instead, an approximation method must be used.

Gibbs sampling

A common approach is to use *Gibbs sampling*

- A Monte-Carlo Markov Chain method from statistical physics
  - "Monte Carlo" means based on random simulation
  - "Markov Chain" describes a random process in which each state depends only on the previous state
- Basic idea is in a complex model with many dependent variables:
  - Sequentially sample each variable, dependent upon state of all other variables
  - Observe averages over very large number of samples as probability estimates

Collapsed Gibbs sampling

Collapsed Gibbs method developed by Griffiths and Steyvers, 2006:

- Marginalize out θ, ϕ
- Instead, estimate $P(\mathbf{z}|\mathbf{w})$ (that is, $P(z_{i,j}|w_{i,j})$ for all i, j)
- They derive the approximation:

$$P(z_{i,j}|\bar{z}_{i,j}) \propto (N_{i,z} + \alpha_z)(N_{z,j} + \beta) \tag{11}$$

where

$\bar{z}_{i,j}$	all other topic assignments to words
$N_{i,z}$	number of times topic z has been assigned to words in document i
$N_{z,j}$	number of times word j has been assigned to topic z

- Iterate many, many times
- Count how many times each word assigned to each topic
- Normalize these counts to estimate θ_i, ϕ_k

Further reading

- Blei, Ng, and Jordan, "Latent Dirichlet Allocation", JMLR, 2003
- Crain, Zhou, Yang, and Zha, "Dimensionality Reduction and Topic Modeling", Chapter 5 of Aggarwal and Zhai (ed.), *Mining Text Data*, 2012 (brief summary of Gibbs sampling).
- Sun, Deng, and Han, "Probabilistic Models for Text Mining", Chapter 8 of Aggarwal and Zhai (ed.), *Mining Text Data*, 2012 (gives $P(\mathbf{w}_1, \ldots, \mathbf{w}_n | \alpha, \beta)$).
- Griffiths and Steyvers, "Finding Scientific Topics", *PNAS*, 2004 (collapsed Gibbs sampling for solving LDA)