Alin Tomescu, http://people.csail.mit.edu/~alinush

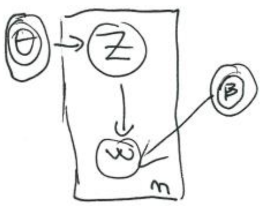
Lecture 18

Latent Dirichlet Allocation

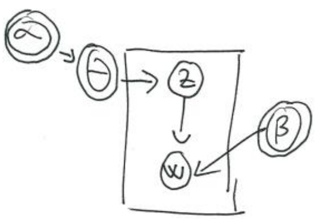
We have data for a single document, it's viewed as a sequence of n words: $w_1, w_2, \ldots, w_n$.

Exchangeable: I can exchange the words and obtain the same probability of getting the document.

Each word is assumed to have been generated from a topic $z_1, z_2, \ldots, z_n \in \{1, \ldots, k\}$



$$\prod_{i=1}^{n} \theta_{z_i} \beta_{w_i|z_i}$$



$$P(\theta; \alpha) \prod_{i=1}^{n} \theta_{z_i} \beta_{w_i|z_i}$$

(these are not document probabilities)

What is $P(d; \alpha, \beta)$?

$$P(d; \alpha, \beta) = \sum_{z_1, \ldots, z_n} \int_{K-simplex} P(\theta; \alpha) \prod_{i=1}^{n} \theta_{z_i} \beta_{w_i|z_i} \, d\theta$$

But this doesn't look like the module we defined last time. It seems like a mixture model with an exponential number of components. How do we get back the previously defined model?

$$P(d; \alpha, \beta) = \sum_{z_1, \ldots, z_n} \int_{K-simplex} P(\theta; \alpha) \prod_{i=1}^{n} \theta_{z_i} \beta_{w_i|z_i} \, d\theta = \int_{K-simplex} P(\theta; \alpha) \prod_{i=1}^{n} \sum_{z_i=1}^{k} \left(\theta_{z_i} \beta_{w_i|z_i}\right) d\theta$$

The summation with the $z_1, \ldots, z_n$ means we are summing over all possible assignments where $z_i \in \{1, \ldots, k\}$. There are k^n options for this, since each z_i can take k values.

There are a lot of options for the prior $P(\theta;\alpha)$, we will choose the most mathematically convenient: the Dirichlet distribution.

$$P(\theta;\alpha) = \frac{1}{z(\alpha)} \prod_{z=1}^{k} \theta_z^{\alpha_z - 1}$$

$$k = \dim \alpha$$

$$z(\alpha) = \frac{\prod_{z=1}^{k} \Gamma(\alpha_z)}{\Gamma\left(\sum_{z=1}^{k} \alpha_z\right)}, \Gamma(k+1) = k\Gamma(k) = k!$$

How do the points concentrate within the simplex? The α_z's are **hyperparameters** and they must somehow specify the cloud of points. What is the mean of this cloud of points?

$$E\{\theta_z|\alpha\} = \frac{\alpha_z}{\sum_{z'} \alpha_{z'}}$$

How concentrated the points are around the mean? It depends on the sum $\sum_{z'} \alpha_{z'}$:

- sum is large, then they are tight
- sum is small, then they are spread out

If someone gave us the topics $Z = \{z_1, \dots, z_n\}$. What is the likelihood of this data?

$$P(Z|\theta) = L(\theta; Z) = \prod_{i=1}^{n} \theta_{z_i} = \prod_{z=1}^{k} \theta_z^{n(z)}, n(z) = \text{\# of times that z occurred in } z_1, \dots, z_n$$

$$P(Z|\theta)P(\theta;\alpha) = P(\theta|Z;\alpha)P(Z;\alpha) \propto P(\theta|Z;\alpha)$$

If LHS is Dirichlet, then so is RHS

$$P(\theta|Z,\alpha) \propto P(\theta;\alpha)P(Z|\theta) = \frac{1}{z(\alpha)} \prod_{z=1}^{k} \theta_z^{\alpha_z - 1} \prod_{z=1}^{k} \theta_z^{n(z)} = \frac{1}{z(\alpha)} \prod_{z=1}^{k} \theta_z^{\alpha_z + n(z) - 1}$$

The posterior as a Dirichlet is:

$$\text{Dirichlet}(\alpha_1 + n(1), \dots, \alpha_k + n(k))$$

The prior was:

$$\text{Dirichlet}(\alpha_1, \dots, \alpha_k)$$

Dirichlet's are convenient conjugate priors for multinomial (http://stats.stackexchange.com/questions/44494/why-is-the-dirichlet-distribution-the-prior-for-the-multinomial-distribution).

Learning LDA models

We are trying to maximize the log-likelihood with respect to alpha and beta:

$$\max_{\alpha,\beta} \sum_{t=1}^{T} \log P(d^t; \alpha, \beta)$$

In principle we can apply EM algorithm:

$$\log P(\theta; \alpha) + \sum_{i=1}^{n} \log \theta_{z_i} + \sum_{i=1}^{n} \log \beta_{w_i|z_i}$$

Latent variables: θ, z_i.

$$E\left\{\log P(\theta; \alpha) + \sum_{i=1}^{n} \log \theta_{z_i} + \sum_{i=1}^{n} \log \beta_{w_i|z_i} \middle| d, \alpha^{[m]}, \beta^{[m]}\right\}$$

If we are only learning β, we need to take an average of $\sum_{i=1}^{n} \log \beta_{w_i|z_i}$ assuming we know the prior $P(\theta; \alpha)$.

Gibbs sampling

In the E-step we need to figure out $P\left(z_1, \dots, z_n \mid \alpha^{[m]}, \beta^{[m]}\right)$. But there are too many combinations so we just have to sample:

$$(\hat{z}_1, \dots, \hat{z}_n) \sim P\left(z_1, \dots, z_n \mid \alpha^{[m]}, \beta^{[m]}\right)$$

How can we draw that sample? Gibbs' sampling.

$$P(z_1, \dots, z_n)$$

Since there are k^n possible configurations it's very hard to draw?

Start with $z_1^0, \dots, z_n^0$ and randomly update that configuration one coordinate at a time

$$\text{For } i = 1, \dots, n \text{ (in random order) do:}$$

Ask what would be a better value for z_i?

$$z_i = P\left(z_i \middle| z^{-i}\right) = P(z_i \mid z_1, \dots z_{i-1}, z_{i+1}, \dots, z_n),$$

where $z_1, \dots z_{i-1}, z_{i+1}, \dots, z_n$ are fixed

And then reiterate quite a few times to forget you started from $z_1^0, \dots, z_n^0$. Hard to say where that is exactly, but it seems that you need to definitely do it more than n times.

How can you sample from $P(z_i \mid z_1, \dots z_{i-1}, z_{i+1}, \dots, z_n)$?

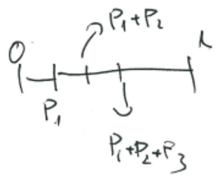
$P_1 \dots P_k$

P_1

$P_1 + P_2$

$0.3 \rightarrow$ $P_1 + P_2 + P_3$

...



Pick a random number between $[0, 1]$

What is $P(z_i \mid z_1, \dots z_{i-1}, z_{i+1}, \dots, z_n)$?

$$z_1, z_2 \dots z_{i-1}, z_i, z_{i+1}, \dots, z_n$$

$$w_1, w_2 \dots w_{i-1}, w_i, w_{i+1}, \dots, w_n$$

$$P(z_i \mid \hat{z}_1, \dots \hat{z}_{i-1}, \hat{z}_{i+1}, \dots, \hat{z}_n, d, \alpha, \beta)$$

If θ were fixed, words are independent of each other?

$$P(z_i \mid \hat{z}_1, \dots \hat{z}_{i-1}, \hat{z}_{i+1}, \dots, \hat{z}_n, d, \alpha, \beta) \propto \beta_{w_i \mid z_i} \theta_{z_i}$$

If we had a single word only:

$$P(z_i \mid \hat{z}_1, \dots \hat{z}_{i-1}, \hat{z}_{i+1}, \dots, \hat{z}_n, d, \alpha, \beta) \propto \beta_{w_i \mid z_i} \frac{\alpha_{z_i}}{\sum_{z'} \alpha_{z'}}$$

$$P(z_i \mid \hat{z}_1, \dots \hat{z}_{i-1}, \hat{z}_{i+1}, \dots, \hat{z}_n, d, \alpha, \beta) \propto \beta_{w_i \mid z_i} \frac{\alpha_{z_i} + n^{-i}(z_i)}{\sum_{z'} \alpha_{z'} + n - 1}$$

$$\sum_z n^{-i}(z) = n - 1$$