LDA

Saki Kuzushima

2019-05-03

1. The Dirichlet is a conjugate prior of the multinomial

The Dirichelt distribution and the multinomial distribution are the key distributions for the latent Dirichlet allocation. First we review the key features of the two distributions, and confirm that the Dirichlet distribution is a conjugate prior to the multinomial distribution.

The multinomial distribution has the following pmf. If

$$X \sim Multinomial(\theta, n)$$

where $X = (X_1, ..., X_n)$, $\theta = (\theta_1, ...\theta_k)$. Then,

$$p(x|\theta, n) = \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \tag{1}$$

where $x = (x_1, ..., x_n)$, and $\sum_{i=1}^{K} x_i = n$.

The Dirichlet distribution has the following pdf. If

$$\theta \sim Dirichlet(\alpha)$$

Then,

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{2}$$

where $\sum_{i=1}^{K} \theta_i = 1$ and $\theta_i \geq 0, \forall i$. (i.e. θ belongs to the $K-1$ simplex.)

We can intuitively think how these distributions are related. We can think of the multinomial distribution as a result of n consecutive rolls of a die, where $X = (X_1, ..., X_6)$ is the number of each face showing up during the n rolls. Then, the Dirichlet distribution is the distribution over the probability of each face showing up in each roll. (If a die is fair, then $\theta = (1/6, ...1/6)$).

We can mathmatically show that the Dirichlet distribution is a conjugate prior of the multinomial distribution. The posterior of θ is

$$p(\theta|X, \alpha) \propto p(X|\theta, \alpha)p(\theta|\alpha)$$

Then, the RHS is

$$
\begin{aligned}
& p(X|\theta,\alpha)p(\theta|\alpha) \\
&= \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\
&\propto \prod_{i=1}^{K} \theta_i^{x_i + \alpha_i - 1}
\end{aligned}
$$

This is the kernel of the Dirichlet pdf. So we find that the posterior is also the Dirichlet distribution. By adjusting the constant term, we find that

$$
p(\theta|X) = \frac{1}{B(\alpha + x)} \prod_{i=1}^{K} \theta_i^{x_i + \alpha_i - 1} \tag{3}
$$

Therefore, the posterior distribution of θ is

$$
\theta^{post} \sim Dirichlet(\alpha + x)
$$

Latent Dirichlet Allocation model

The original model

We firstly define the key terms used in the model. - Word: an element of a vocabulary. - Vocabulary: a set of V unique words appear in the entire corpus. - Document: a sequence of N words, denoted by $d = (w_1, w_2, , ..w_N)$ - Corpus: a set of M documents, denoted by $D = (d_1, d_2 ... d_M)$ - Topic: a distribution of words. We assume that the number of topics is K.

LDA is the following generating probabilistic model.

- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...M]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$

For instance, we have three topics on a news article: sports, politics, and technology ($K = 3$). θ_j, which is the distribution of these topics over the document d_j was drawn from the Dirichlet distribution. We assume that the sampled $\theta_j = [\frac{1}{3}, \frac{1}{4}, \frac{5}{12}]$, where the order of topics are [sports, politics, technology]. Substantively, this means that the document d_j is a mixture of 1/3 "sports", 1/4 "poliitcs", and 5/12 "technology".

Next, we draw the topic of each word $z_{j,n}$ from $Multinomial(\theta)$. We assume this happens to be "politics." Then, we sample each word $w_{j,n}$ from $Multinomial(\phi_{zj,n})$.

Assuming that the documents and the words are independent, $d_j \perp\!\!\!\perp d_{j'}$ and $w_{j,n} \perp\!\!\!\perp w_{j,n'}$, we can write the joint distribution of W, Z, Θ given α, Φ in the following way.

$$
p(W, Z, \Theta|\alpha, \Phi) = \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{j=1}^{K} p(z_{j,n}|\theta_j) p(w_{j,n}|\phi_{z_{j,n}}) \tag{4}
$$

Smoothed LDA

Alternatively, we can assign a prior to ϕ. This model is called smoothed LDA.

- For each topic, $\phi_i \sim Dirichlet(\beta), i \in [1, ..., K]$
- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...M]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$

In this model, we can write the joint distribution of W, Z, Θ given α, β.

$$p(W, Z, \Theta|\alpha, \beta) = \prod_{i=1}^{K} p(\phi_j|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{j=1}^{K} p(z_{j,n}|\theta_j) p(w_{j,n}|\phi_{z_{j,n}}) \tag{5}$$

Variational Inference (original method by Blei et al.)

The original inference method uses variational inference. Variational inference essentially approximate the posterior with a distribution that varies across some parameters, and tune the parameters to minimize the difference between the posterior and the approximate distribution.