Gibbs Sampling for Bayesian Mixture

Ayush Tewari

December 1st, 2025

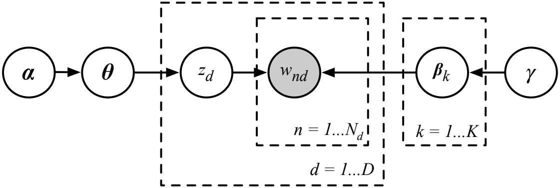
Adapted from Carl Edward Rasmussen

Key concepts

- General Bayesian mixture model
- We derive the Gibbs sampler
- Marginalize out mixing proportions: collapsed Gibbs sampler

Bayesian document mixture model

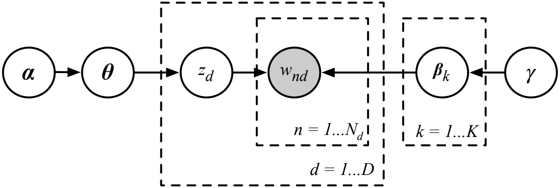
- Our mixture model has observations $\mathbf{w}_d$ the words in document $d = 1, \ldots, D$. The parameters are $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}$, and latent variables $\mathbf{z}$.



$$
\begin{aligned}
\boldsymbol{\theta} &\sim \text{Dir}(\alpha) \\
\boldsymbol{\beta}_k &\sim \text{Dir}(\gamma) \\
z_d | \boldsymbol{\theta} &\sim \text{Cat}(\boldsymbol{\theta}) \\
w_{nd} | z_d, \boldsymbol{\beta} &\sim \text{Cat}(\boldsymbol{\beta}_{\text{z}_\text{d}})
\end{aligned}
$$

Bayesian document mixture model

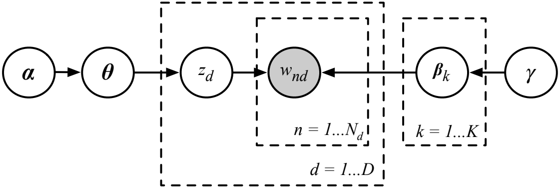
- Our mixture model has observations $\mathbf{w}_d$ the words in document $d = 1, \ldots, D$. The parameters are $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}$, and latent variables $\mathbf{z}$.
- The mixture model has K components, so the parameters are $\boldsymbol{\beta}_k, k = 1, \ldots K$. Each $\boldsymbol{\beta}_k$ is the parameter of a categorical over possible words, with prior $p(\boldsymbol{\beta})$. The discrete latent variables $z_d, d = 1, \ldots D$ take on values $1, \ldots K$.



$$\begin{aligned} \boldsymbol{\theta} &\sim \mathrm{Dir}(\alpha) \\ \boldsymbol{\beta}_k &\sim \mathrm{Dir}(\gamma) \\ z_d|\boldsymbol{\theta} &\sim \mathrm{Cat}(\boldsymbol{\theta}) \\ w_{nd}|z_d, \boldsymbol{\beta} &\sim \mathrm{Cat}(\boldsymbol{\beta}_{\mathrm{z_d}}) \end{aligned}$$

Bayesian document mixture model

- Our mixture model has observations $\mathbf{w}_d$ the words in document $d = 1, \ldots, D$. The parameters are $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}$, and latent variables $\mathbf{z}$.
- The mixture model has K components, so the parameters are $\boldsymbol{\beta}_k, k = 1, \ldots K$. Each $\boldsymbol{\beta}_k$ is the parameter of a categorical over possible words, with prior $p(\boldsymbol{\beta})$. The discrete latent variables $z_d, d = 1, \ldots D$ take on values $1, \ldots K$.
- Note, that in this model the observations are (the word counts of) entire documents.



$$
\begin{aligned}
\boldsymbol{\theta} &\sim \mathrm{Dir}(\alpha) \\
\boldsymbol{\beta}_k &\sim \mathrm{Dir}(\gamma) \\
z_d | \boldsymbol{\theta} &\sim \mathrm{Cat}(\boldsymbol{\theta}) \\
w_{nd} | z_d, \boldsymbol{\beta} &\sim \mathrm{Cat}(\boldsymbol{\beta}_{\mathrm{z_d}})
\end{aligned}
$$

Bayesian mixture model

The conditional likelihood is for each observation is

$$p(\mathbf{w}_d | z_d = k, \boldsymbol{\beta}) \; = \; p(\mathbf{w}_d | \beta_k) \; = \; p(\mathbf{w}_d | \beta_{z_d}),$$

and the prior

$$p(\boldsymbol{\beta}_k | \gamma) \; = \; \mathrm{Dir}(\gamma)$$

Bayesian mixture model

The conditional likelihood is for each observation is

$$p(\mathbf{w}_d|z_d = k, \boldsymbol{\beta}) \; = \; p(\mathbf{w}_d|\beta_k) \; = \; p(\mathbf{w}_d|\beta_{z_d}),$$

and the prior

$$p(\boldsymbol{\beta}_k|\gamma) \; = \; \mathrm{Dir}(\gamma)$$

The categorical latent component assignment probability

$$p(z_d = k|\boldsymbol{\theta}) \; = \; \theta_k,$$

with a Dirichlet prior

$$p(\theta|\alpha) \; = \; \mathrm{Dir}(\alpha).$$

Bayesian mixture model

The conditional likelihood is for each observation is

$$p(\mathbf{w}_d|z_d = k, \boldsymbol{\beta}) \ = \ p(\mathbf{w}_d|\beta_k) \ = \ p(\mathbf{w}_d|\beta_{z_d}),$$

and the prior

$$p(\boldsymbol{\beta}_k|\gamma) \ = \ \mathrm{Dir}(\gamma)$$

The categorical latent component assignment probability

$$p(z_d = k|\boldsymbol{\theta}) \ = \ \theta_k,$$

with a Dirichlet prior

$$p(\theta|\alpha) \ = \ \mathrm{Dir}(\alpha).$$

Therefore, the latent conditional posterior is

$$p(z_d = k|\mathbf{w}_d, \boldsymbol{\theta}, \boldsymbol{\beta}) \ \propto p(z_d = k|\boldsymbol{\theta})p(\mathbf{w}_d|z_d = k, \boldsymbol{\beta}) \ \propto \ \theta_k p(\mathbf{w}_d|\beta_{z_d}),$$

which is just a discrete distribution with K possible outcomes.

Gibbs Sampling

The Goal: Sample from the Joint Posterior of all variables:

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \mathbf{w})$$

To achieve this, we iteratively sample from the conditionals:

Gibbs Sampling

The Goal: Sample from the Joint Posterior of all variables:

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \mathbf{w})$$

To achieve this, we iteratively sample from the conditionals:

1. Component parameters (Dirichlet)

$$p(\beta_k | \mathbf{w}, \mathbf{z}) \ \propto \ p(\beta_k) \prod_{d: z_d = k} p(\mathbf{w}_d | \beta_k),$$

which is now a categorical model, the mixture aspect having been eliminated.

Gibbs Sampling

The Goal: Sample from the Joint Posterior of all variables:

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{w})$$

To achieve this, we iteratively sample from the conditionals:

1. Component parameters (Dirichlet)

$$p(\beta_k|\mathbf{w}, \mathbf{z}) \ \propto \ p(\beta_k) \prod_{d:z_d=k} p(\mathbf{w}_d|\beta_k),$$

which is now a categorical model, the mixture aspect having been eliminated.

2. Latent assignments (Categorical)

$$p(z_d = k|\mathbf{w}_d, \boldsymbol{\theta}, \boldsymbol{\beta}) \ \propto \theta_k p(\mathbf{w}_d|\beta_{z_d}),$$

Gibbs Sampling

The Goal: Sample from the Joint Posterior of all variables:

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{w})$$

To achieve this, we iteratively sample from the conditionals:

1. Component parameters (Dirichlet)

$$p(\beta_k|\mathbf{w}, \mathbf{z}) \ \propto \ p(\beta_k) \prod_{d: z_d = k} p(\mathbf{w}_d|\beta_k),$$

which is now a categorical model, the mixture aspect having been eliminated.

2. Latent assignments (Categorical)

$$p(z_d = k|\mathbf{w}_d, \boldsymbol{\theta}, \boldsymbol{\beta}) \ \propto \theta_k p(\mathbf{w}_d|\beta_{z_d}),$$

3. Mixing proportions (Dirichlet)

$$p(\theta|\mathbf{z}, \alpha) \ \propto \ p(\theta|\alpha) p(\mathbf{z}|\theta) \ \propto \ \mathrm{Dir}(\mathbf{c} + \alpha).$$

where $c_k = \sum_{d: z_d = k} 1$ are the counts for mixture k.

Collapsed Gibbs Sampler

The parameters are treated in the same way as before.

Collapsed Gibbs Sampler

The parameters are treated in the same way as before.

If we marginalize over θ

$$\begin{aligned} p(z_d = k|\mathbf{z}_{-d}, \alpha) &= \int p(z_d = k|\theta) p(\theta|\mathbf{z}_{-d}, \alpha) d\theta \\ &= \int \theta_k p(\theta|\mathbf{z}_{-d}, \alpha) d\theta = \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K} \alpha + c_{-d,j}}, \end{aligned}$$

where index $-d$ means all except d, and c_k are counts;
we derived this result when discussing pseudo counts.

Collapsed Gibbs Sampler

The parameters are treated in the same way as before.

If we marginalize over θ

$$\begin{aligned} p(z_d = k|\mathbf{z}_{-d}, \alpha) &= \int p(z_d = k|\theta) p(\theta|\mathbf{z}_{-d}, \alpha) d\theta \\ &= \int \theta_k p(\theta|\mathbf{z}_{-d}, \alpha) d\theta = \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K} \alpha + c_{-d,j}}, \end{aligned}$$

where index $-d$ means all except d, and c_k are counts;
we derived this result when discussing pseudo counts.

The collapsed Gibbs sampler for the latent assignements

$$p(z_d = k|\mathbf{w}_d, z_{-d}, \boldsymbol{\beta}, \alpha) \propto p(\mathbf{w}_d|\beta_k) \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K} \alpha + c_{-d,j}},$$

where now all the z_d variables have become dependent (previously they were conditionally independent given θ).

Collapsed Gibbs Sampler

The parameters are treated in the same way as before.

If we marginalize over θ

$$\begin{aligned} p(z_d = k|\mathbf{z}_{-d}, \alpha) &= \int p(z_d = k|\theta) p(\theta|\mathbf{z}_{-d}, \alpha) d\theta \\ &= \int \theta_k p(\theta|\mathbf{z}_{-d}, \alpha) d\theta = \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K} \alpha + c_{-d,j}}, \end{aligned}$$

where index $-d$ means all except d, and c_k are counts;
we derived this result when discussing pseudo counts.

The collapsed Gibbs sampler for the latent assignements

$$p(z_d = k|\mathbf{w}_d, z_{-d}, \boldsymbol{\beta}, \alpha) \propto p(\mathbf{w}_d|\beta_k) \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K} \alpha + c_{-d,j}},$$

where now all the z_d variables have become dependent (previously they were conditionally independent given θ).

Notice, that the Gibbs sampler exhibits the rich get richer property.

Per word Perplexity

In text modeling, performance is often given in terms of per word perplexity. The perplexity for a document is given by

$$\exp(-\ell/n),$$

where ℓ is the log joint probability over the words in the document, and n is the number of words. Note, that the average is done in the log space.

Per word Perplexity

In text modeling, performance is often given in terms of per word perplexity. The perplexity for a document is given by

$$\exp(-\ell/n),$$

where ℓ is the log joint probability over the words in the document, and n is the number of words. Note, that the average is done in the log space.

A perplexity of g corresponds to the uncertainty associated with a die with g sides, which generates each new word.

Per word Perplexity

In text modeling, performance is often given in terms of per word perplexity. The perplexity for a document is given by

$$\exp(-\ell/n),$$

where ℓ is the log joint probability over the words in the document, and n is the number of words. Note, that the average is done in the log space.

A perplexity of g corresponds to the uncertainty associated with a die with g sides, which generates each new word.

Example:

$$p(w_1, w_2, w_3, w_4) = \frac{1}{6}\frac{1}{6}\frac{1}{6}\frac{1}{6} \tag{1}$$

$$\frac{1}{n}\log p(w_1, \ldots, w_4) = \frac{1}{4}\log(\frac{1}{6})^4 = -\log 6 \tag{2}$$

$$\text{perplexity} = \exp(-\frac{1}{n}\log p(w_1, \ldots, w_4)) = 6 \tag{3}$$