Gibbs Sampling for Bayesian Mixture

Carl Edward Rasmussen

November 27th, 2018

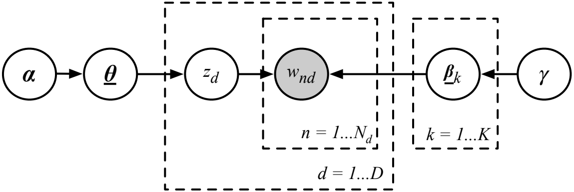
Key concepts

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

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

The mixture model has K components, so the parameters are $\underline{\boldsymbol{\beta}}_k, k = 1, \ldots K$. 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.



$$
\begin{aligned}
\text{length K.}\quad \underline{\theta} &\sim \mathrm{Dir}(\alpha) \\
\text{length M.}\quad \underline{\boldsymbol{\beta}}_k &\sim \mathrm{Dir}(\gamma) \\
z_d | \underline{\theta} &\sim \mathrm{Cat}(\underline{\theta}) \\
w_{nd} | z_d, \underline{\boldsymbol{\beta}} &\sim \mathrm{Cat}(\boldsymbol{\beta}_{z_d})
\end{aligned}
$$

Bayesian mixture model

The conditional likelihood is for each observation is

prob of the word sequence given that the doc has topic k.

$$p(\mathbf{w}_d|z_d = k, \boldsymbol{\beta}) \;=\; p(\mathbf{w}_d|\underline{\beta}_k) \;=\; p(\mathbf{w}_d|\underline{\beta}_{z_d}),$$

and the prior

$$p(\underline{\boldsymbol{\beta}}_k).$$ ~ Dir(gamma)

The categorical latent component assignment probability

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

with a Dirichlet prior

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

Therefore, the latent conditional posterior is

posterior prior likelihood

$$\underbrace{p(z_d = k|\mathbf{w}_d, \underline{\theta}, \underline{\boldsymbol{\beta}})}_{\text{posterior}} \;\propto\; \underbrace{p(z_d = k|\underline{\theta})}_{\text{prior}}\underbrace{p(\mathbf{w}_d|z_d = k, \boldsymbol{\beta})}_{\text{likelihood}} \;\propto\; \theta_k p(\mathbf{w}_d|\beta_{z_d}),$$

which is just a discrete distribution with K possible outcomes.

Gibbs Sampling

Iteratively, alternately, sample the three types of variables:

Component parameters

posterior prior likelihood of beta_k

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

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

The posterior latent conditional allocations

$$p(z_d = k|\mathbf{w}_d, \theta, \boldsymbol{\beta}) \propto \theta_k p(\mathbf{w}_d|\underline{\beta}_{z_d}),$$ from last page.

are categorical and mixing proportions

posterior prior likelihood of theta_

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

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

Hence c_1, ..., c_K sum up to N, the total no of documents.

Collapsed Gibbs Sampler

The parameters are treated in the same way as before.

If we marginalize over θ

prior

$$p(z_d = k|\mathbf{z}_{-d}, \alpha) \;=\; \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K} \alpha + c_{-d,j}},$$

Predicted prob of z_d given the categories of all the other documents, z_{-d}, is given by the posterior mean of z_{-d}, which equals a fraction defined by the Dirichlet parameters alpha and c as shown.

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

posterior likelihood prior

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

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.