Module 7: Part IV: Gibbs Sampling, Data Augmentation, Mixture Models

Rebecca C. Steorts

This goes with Lab 8, which has been prepared by Olivier Binette

Agenda

- Review of data augmentation
- A three component mixture model
- Model specification
- The Dirichlet-Multinomial
- Return to the three component mixture model problem
- This corresponds to Lab 8
- Homework this week is to go through the details of this lab on your own and bring detailed questions to your lab either with Olivier/Michael on Friday.

You can find Olivier's outline for Lab 8 here:
https://github.com/OlivierBinette/Labs-STA-360

Goal

The goal of this lecture, which corresponds with lab 8, is to introduce you to the **three component mixture model**.

This easily extends to any type of mixture model.

Background

In order to work with this module, we will need to know to work with the Multinomial and Dirichlet distributions.

Big picture

- We will build a model specification based upon a **three component mixture model**
- What does this model look like?
- The posterior will be intractable? Why?
- What should we do in order to help us solve the problem?
- This will be your homework for this coming week, and it will be solved and work in lab 8 with your TA's.
- You should have read through the lecture notes for class and attempted to work through the problems on your own before attending Lab 8.

Likelihood (three component mixture model)

For $i = 1, \ldots, n$

$$
\begin{aligned}
& p(Y_i | \mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) \\
& = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2) \\
& = w_1 \; N(\mu_1, \varepsilon^2) + w_2 \; N(\mu_2, \varepsilon^2) + w_3 \; N(\mu_3, \varepsilon^2)
\end{aligned}
$$

- w_1, w_2 and w_3 are the mixture weight of mixture components 1,2 and 3 respectively
- μ_1, μ_2 and μ_3 are the means of the mixture components
- ε^2 is the variance parameter of the error term around the mixture components.

Prior specification on likelihood terms

Let's specify the priors on

- w_1, w_2 and w_3 are the mixture weight of mixture components 1,2 and 3 respectively
- μ_1, μ_2 and μ_3 are the means of the mixture components
- ε^2 is the variance parameter of the error term around the mixture components.

Prior specification on likelihood terms

For $i = 1, \ldots, n$

$$p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2)$$

$$\mu_j|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0) \quad (1)$$

$$\varepsilon^2 \sim \text{InverseGamma}(2, 2) \quad (2)$$

$$(w_1, w_2, w_3) \sim \text{Dirichlet}(1, 1, 1) \quad (3)$$

What is the Dirichlet$(1, 1, 1)$? This is the multivariate distribution of the Beta distribution.

Complete the model specification

Let's specify the priors on

- μ_0
- σ_0^2

Finalizing model specification

For $i = 1, \ldots, n$

$$p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2)$$

$$\mu_j|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0) \tag{4}$$
$$\varepsilon^2 \sim \text{InverseGamma}(2, 2) \tag{5}$$
$$(w_1, w_2, w_3) \sim \text{Dirichlet}(1, 1, 1) \tag{6}$$

$$\mu_0 \sim N(0, 3) \tag{7}$$
$$\sigma_0^2 \sim \text{InverseGamma}(2, 2) \tag{8}$$

Three component mixture model (Lab 8)

In order to be able to work on this problem, we need to:

1. We need to realize that the full conditionals as written cannot be easily sampled from. (Lab 8).
2. Next, we want to re-write the model using latent allocation variables to make it easier to work with.
3. Finally, in order to work with this model, we need to know about two distributions — the Dirichlet and the Multinomial. It's also essential to note that the Dirichlet is the conjugate prior for the Multinomial.

Three component mixture model

- Recall the three component mixture of normal distribution with a common prior on the mixture component means, the error variance and the variance within mixture component means.
- The prior on the mixture weights w is a three component Dirichlet distribution.

$$
\begin{aligned}
p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) &= \sum_{j=1}^{3} w_i N(\mu_j, \varepsilon^2) \\
\mu_j|\mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\
\mu_0 &\sim N(0, 3) \\
\sigma_0^2 &\sim \text{InverseGamma}(2, 2) \\
(w_1, w_2, w_3) &\sim \text{Dirichlet}(1, 1, 1) \\
\varepsilon^2 &\sim \text{InverseGamma}(2, 2),
\end{aligned}
$$

for $i = 1, \ldots n$.

Task 1 and Task 2

Derive the joint posterior up to a normalizing constant. What do you observe?

Specifically, derive

$$p(w_1, w_2, w_3, \mu_1, \mu_2, \mu_3, \epsilon^2, \mu_o, \sigma_o^2 \mid y_{1:n})$$

up to a normalizing constant, where it may be helpful to let $\tau = \frac{1}{\epsilon^2}$, $\phi_o = \frac{1}{\sigma_o^2}$.

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right)\left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right)\left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right)\left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

$$p(\mu_0) = N(\mu_0; 0, 3),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right)\left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

$$p(\mu_0) = N(\mu_0; 0, 3),$$

$$p(\phi_0) = \mathrm{Gamma}(\phi_0; 2, 2),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right)\left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

$$p(\mu_0) = N(\mu_0; 0, 3),$$

$$p(\phi_0) = \text{Gamma}(\phi_0; 2, 2),$$

$$p(\tau) = \text{Gamma}(\tau; 2, 2).$$

Task 2

Using Task 1, what are the following conditional distributions below

- $p(w_1, w_2, w_3 | \mu_1, \mu_2, \mu_3, \varepsilon^2, Y_1, ..., Y_N) \propto$
- $p(\mu_1 | \mu_2, \mu_3, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\mu_2 | \mu_1, \mu_3, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\mu_3 | \mu_1, \mu_2, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\varepsilon^2 | \mu_1, \mu_2, \mu_3, Y_1, ..., Y_N) \propto$
- $p(\mu_0 | \mu_1, \mu_2, \mu_3, \sigma_0^2) \propto$
- $p(\sigma_0^2 | \mu_0, \mu_1, \mu_2, \mu_3) \propto$

Task 2

Hint: most of the full conditionals are available from Task 1.

Observe that the likelihood is difficutl to work with, and thus we will utilize latent variables in order to help us proceed as we did in the last module.

Using latent variables

Neither the joint posterior nor any of the full conditionals involving the likelihood are of a form that is easy to sample from.

Using latent variables

We will introduce an additional set of random variables $\{Z_i\}_{i=1}^{N}$ that assign each observation to one of the mixture components with the proability of assignment being the respective mixture weight.

If we condition on Z_i we can then write the likelihood of Y_i as

$$p(Y_i|Z_i, \mu_1, \mu_2, \mu_3, \varepsilon^2) = \sum_{j=1}^{3} N(\mu_j, \varepsilon^2)\delta_j(Z_i) = \sum_{j=1}^{3} N(\mu_{Z_i}, \varepsilon^2)$$

$$P(Z_i = j) = w_j.$$

Latent variables

- Conditional on Z_i we no longer have a sum of Normal pdfs in our likelihood, resulting in a significant simplification.
- Conditional on the $\{Z_i\}$ updates will be straightforward, only depending on the mixture component that any given Y_i is currently assigned to.
- The drawback is that we also have to update $\{Z_i\}_{i=1}^{N}$ as well, introducing extra steps into our sampler.

The updated model

The model is now

$$
\begin{aligned}
Y_i \mid Z_i, \mu_1, \mu_2, \mu_3, \epsilon^2 &\sim \sum_{i=1}^{3} N(\mu_{Z_i}, \epsilon^2) \\
\mu_j \mid \mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\
Z_i \mid w_1, w_2, w_3 &\sim \mathrm{Cat}(3, \boldsymbol{w}) \\
\boldsymbol{w} = (w_1, w_2, w_3) &\sim \mathrm{Dirichlet}(1, 1, 1) \\
\mu_0 &\sim N(0, 3) \\
\sigma_0^2 &\sim IG(2, 2) \\
\epsilon^2 &\sim IG(2, 2)
\end{aligned}
$$

$i = 1, \ldots, n$ $j = 1, \ldots, 3$

Task 3

Where necessary, (re)derive the full conditionals under the data augmentation scheme.

(See the lab solutions).

Task 4

In task 3 you derived all the full conditionals, and due to data augmentation scheme they are all in a form that is easy to sample. Use these full conditionals to implement Gibbs sampling using the data from "Lab8Mixture.csv".

Task 5

- Show traceplots for all estimated parameters
- Show means and 95% credible intervals for the marginal posterior distributions of all the parameters

Now suppose you re-run the sampler using 3 different starting values, are your results in a,b the same? Justify your reasoning with visualizations.

Sample code

Partial code for this problem can be found at
https://github.com/resteorts/modern-bayes/tree/master/labs/08-gibbs-augmentation

Recap of Module 8 (Part I – Part IV)

1. We introduced the two-stage Gibbs sampler.
2. You should be able to derive conditional distributions. for two-stage Gibbs samplers. (See Part I, Module 8 for examples).
3. Be familar with diagnostic plots.
4. We then looked at a three-stage sampler and generalized to the multi-stage Gibbs sampler.
5. We looked at an application to censoring (a type of missing data here).
6. Why would we use latent variables in a Gibbs sampler? (We looked at these for Gaussian mixture models). Notice that the hierarhical modeling setup was more complicated here, which is why we used this trick.
7. In short, we saw many ways to use Gibbs sampling in many applications and various tricks that one needs to use in order to derive the full conditionals in closed form. This is always driven by the data and will vary by the model specified.