A Bayesian 2-Mixture Model

Jason D. M. Rennie
jrennie@gmail.com

March 11, 2005

1 Bayesian Formulation

We are interested in a two-component mixture model where each component itself is a distribution. The generative model is as follows. For each word in a document of length, n, a mixture component is drawn, $c \sim \text{Bernoulli}(\lambda)$. Then, a unigram parameter is drawn, $\mu \sim \text{Beta}(\alpha_c, \beta_c)$, where α_1, β_1 are the parameters for component 1 and α_2, β_2 are the parameters for component 2. Finally, a number of word occurrence is drawn, $h \sim \text{Binomial}(\mu)$.

Let D be the set of documents. Let i index the documents. Let h_i be the number of occurrences (heads); let n_i be the length (number of flips).

We want to find parameters (λ, α_1, β_1, α_2, β_2) to maximize the likelihood of the observed data. We make a few definitions for convenience. Let

$$G(a,b) \triangleq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}, \tag{1}$$

$$B(x|a,b) \triangleq G(a,b)x^{a-1}(1-x)^{b-1}, \text{ and} \tag{2}$$

$$U(h|n,x) \triangleq x^h x^{n-h}. \tag{3}$$

Note that $B(x;a,b)$ is a distribution. Also note that

$$\int U(h|n,x)B(x|a,b)dx = \frac{G(a,b)}{G(h+a,n-h+b)}. \tag{4}$$

Let θ represent the parameters of the model. The likelihood of the data is

$$p(D|\theta) = \prod_i \int p(h_i|n_i,\mu)p(\mu|\theta)d\mu. \tag{5}$$

The first term is the simple unigram probability of seeing h_i heads in n_i flips using a coin with μ probability of heads. The second term involves the mixture:

$$p(D|\theta) = \prod_i \int U(h_i|n_i,\mu)\Big[\lambda B(\mu|\alpha_1,\beta_1) + (1-\lambda)B(\mu|\alpha_2,\beta_2)\Big]d\mu, \tag{6}$$

$$= \prod_i \left(\lambda \frac{G(\alpha_1,\beta_1)}{G(h_i+\alpha_1,n_i-h_i+\beta_1)} + (1-\lambda)\frac{G(\alpha_2,\beta_2)}{G(h_i+\alpha_2,n_i-h_i+\beta_2)}\right). \tag{7}$$

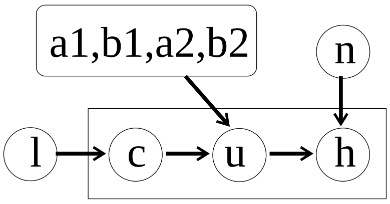

Figure 1: Shown is the generative model for a word. Nodes outside the rectangular box are constants. Nodes within the rectangle are generated once per document.

This is the quantity which we are interested in maximizing. We make some definitions that will make it easier to write down the derivatives. Let

$$G_1 \triangleq G(\alpha_1, \beta_1), \quad H_{1i} \triangleq G(h_i + \alpha_1, n_i - h_i + \beta_1), \tag{8}$$

$$G_2 \triangleq G(\alpha_2, \beta_2), \quad H_{2i} \triangleq G(h_i + \alpha_2, n_i - h_i + \beta_2), \text{ and} \tag{9}$$

$$Z_i \triangleq \lambda G_1/H_{1i} + (1-\lambda) G_2/H_{2i}. \tag{10}$$

Since the logarithm is a strictly monotone, increasing function, we can equivalently maximize the log-likelihood:

$$l = \sum_i \log\left[\lambda G_1/H_{1i} + (1-\lambda) G_2/H_{2i}\right]. \tag{11}$$

We assume that we have a function that, given the current point and gradient at that point returns a new point with higher likelihood. Left is for us to calculate the gradient. First, we give the derivative with respect to λ,

$$\frac{\partial l}{\partial \lambda} = \sum_i \frac{G_1/H_{1i} - G_2/H_{2i}}{Z_i}. \tag{12}$$

Note that $\Gamma'(x) = \Gamma(x)\Psi(x)$, where $\Psi(x)$ is the digamma function[1]. Some useful derivatives are

$$\frac{\partial G_1}{\partial \alpha_1} = G_1\left[\Psi(\alpha_1 + \beta_1) - \Psi(\alpha_1)\right], \tag{13}$$

$$\frac{\partial H_{1i}}{\partial \alpha_1} = H_{1i}\left[\Psi(\alpha_1 + \beta_1 + n_i) - \Psi(\alpha_1 + h_i)\right], \tag{14}$$

$$\frac{\partial G_1}{\partial \beta_1} = G_1\left[\Psi(\alpha_1 + \beta_1) - \Psi(\beta_1)\right], \text{ and} \tag{15}$$

$$\frac{\partial H_{1i}}{\partial \beta_1} = H_{1i}\left[\Psi(\alpha_1 + \beta_1 + n_i) - \Psi(\beta_1 + n_i - h_i)\right]. \tag{16}$$

[1] See, for example, Mathworld's description, http://mathworld.wolfram.com/DigammaFunction.html.

The derivatives with respect to the Beta parameters are all very similar:

$$\frac{\partial l}{\partial \alpha_1} = \sum_i \frac{\lambda}{Z_i} \frac{G_1}{H_{1i}} \left[\Psi(\alpha_1 + \beta_1) - \Psi(\alpha_1) - \Psi(\alpha_1 + \beta_1 + n_i) + \Psi(\alpha_1 + h_i)\right]$$

$$\frac{\partial l}{\partial \alpha_2} = \sum_i \frac{1-\lambda}{Z_i} \frac{G_2}{H_{2i}} \left[\Psi(\alpha_2 + \beta_2) - \Psi(\alpha_2) - \Psi(\alpha_2 + \beta_2 + n_i) + \Psi(\alpha_2 + h_i)\right]$$

$$\frac{\partial l}{\partial \beta_1} = \sum_i \frac{\lambda}{Z_i} \frac{G_1}{H_{1i}} \left[\Psi(\alpha_1 + \beta_1) - \Psi(\beta_1) - \Psi(\alpha_1 + \beta_1 + n_i) + \Psi(\beta_1 + n_i - h_i)\right]$$

$$\frac{\partial l}{\partial \beta_2} = \sum_i \frac{1-\lambda}{Z_i} \frac{G_2}{H_{2i}} \left[\Psi(\alpha_2 + \beta_2) - \Psi(\beta_2) - \Psi(\alpha_2 + \beta_2 + n_i) + \Psi(\beta_2 + n_i - h_i)\right]$$