1 EM Algorithm for Multinomial Mixture Models

Let K be the number of topics, D be the number of documents, and V be the dimensionality of the term vocabulary. We can represent a document in terms of a V-dimensional vector of term counts $\mathbf{x}_d = (x_{d,1}, \dots, x_{d,V})$. Then we can express

$$\Pr\left[\,\mathbf{x}_d \mid z_d = k\,\right] = \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}$$

and

$$\Pr\left[\,\mathbf{x}_d\,\right] = \sum_k \Pr\left[\,z_d = k\,\right] \Pr\left[\,\mathbf{x}_d \mid z_d = k\,\right] = \sum_k \rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}.$$

The overall probability of the data, given independence of z_d across documents, is

$$\Pr\left[\,\mathbf{x}\,\right] = \prod_d \sum_k \rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}$$

so that the log-likelihood is

$$\log\left(\Pr\left[\,\mathbf{x}\,\right]\right) = \sum_d \log\left[\sum_k \rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}\right].$$

The typical approach in maximum likelihood estimation is to directly maximize $\log\left(\Pr\left[\,\mathbf{x}\,\right]\right)$ with respect to the ρ_k and $\beta_{k,v}$ parameters. Unfortunately the summation over latent topics enters inside the logarithm so this is not a tractable problem. Instead the EM algorithm provides an alternative way of recovering parameter estimates.

We begin by writing down the joint likelihood $p\left(\mathbf{x}_d, z_d\right)$ of document d. Note that

$$\Pr\left[\,\mathbf{x}_d, z_d = k\,\right] = \Pr\left[\,\mathbf{x}_d \mid z_d = k\,\right] \Pr\left[\,z_d = k\,\right] = \rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}$$

so we can express the joint likelihood of d as

$$p\left(\mathbf{x}_d, z_d\right) = \left[\rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}\right]^{\mathbb{1}(z_d = k)}$$

and, again given independence, the overall joint likelihood of observing all documents as

$$p\left(\mathbf{x}, \mathbf{z}\right) = \prod_d \left[\rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}\right]^{\mathbb{1}(z_d = k)}$$

leading to a log probability of

$$\log\left[p\left(\mathbf{x},\mathbf{z}\right)\right] = \sum_d \mathbb{1}(z_d = k)\left[\log(\rho_k) + \sum_v x_{d,v}\log\left(\beta_{k,v}\right)\right].$$

The idea of the EM algorithm is to iteratively compute the expectation of $\log\left[p\left(\mathbf{x},\mathbf{z}\right)\right]$ with respect to the posterior distribution of the latent variables given the parameters, and then to maximize the resulting expectation with respect to the parameters. We consider both steps in turn.

1.1 E-step

Clearly $\mathbb{E}\left[\,\mathbb{1}(z_d = k) \mid \rho, \beta, \mathbf{x}_d\,\right] = \Pr\left[\,z_d = k \mid \rho, \beta, \mathbf{x}_d\,\right]$. Moreover by Bayes' rule we obtain that

$$\hat{z}_{d,k} \equiv \Pr\left[\,z_d = k \mid \rho, \beta, \mathbf{x}_d\,\right] = \frac{\Pr\left[\,\mathbf{x}_d \mid \rho, \beta, z_d = k\,\right]\Pr\left[\,z_d = k\,\right]}{\sum_k \Pr\left[\,\mathbf{x}_d \mid \rho, \beta, z_d = k\,\right]\Pr\left[\,z_d = k\,\right]} = \frac{\rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}}{\sum_k \rho_k \prod_v \left(\beta_{k,v}\right)^{x_{d,v}}}.$$

So taking expectations of $\log\left[p\left(\mathbf{x},\mathbf{z}\right)\right]$ with respect this latent variable distribution yields

$$Q(\rho,\beta) = \sum_d \hat{z}_{d,k}\left[\log(\rho_k) + \sum_v x_{d,v}\log\left(\beta_{k,v}\right)\right].$$

1.2 M-step

We want to maximize $Q(\rho,\beta)$, but have to consider the constraints that ρ and β_k must be probability vectors. We can form the following Lagrangean to solve this problem:

$$\mathcal{L}(\rho,\beta,\lambda,\nu_k) = Q(\rho,\beta) + \lambda\left(1 - \sum_k \rho_k\right) + \sum_k \nu_k\left(1 - \sum_v \beta_{k,v}\right).$$

Differentiating with respect to ρ_k gives

$$\frac{\sum_d \hat{z}_{d,k}}{\rho_k^*} - \lambda = 0 \rightarrow \rho_k^* = \frac{\sum_d \hat{z}_{d,k}}{\lambda}$$

Moreover, the sum-to-one constraint gives $\frac{\sum_k \sum_d \hat{z}_{d,k}}{\lambda} = 1$ so that the optimal update is

$$\rho_k^* = \frac{\sum_d \hat{z}_{d,k}}{\sum_k \sum_d \hat{z}_{d,k}} \tag{M-step 1}$$

This is the average probability that documents have topic k, which makes sense.

Differentiating with respect to $\beta_{k,v}$ gives

$$\frac{\sum_d \hat{z}_{d,k} x_{d,v}}{\beta^*_{k,v}} - \nu_k = 0$$

which after again using the sum-to-one constraint gives

$$\beta^*_{k,v} = \frac{\sum_d \hat{z}_{d,k} x_{d,v}}{\sum_d \hat{z}_{d,k} \sum_v x_{d,v}}. \tag{M-step 2}$$

This also makes sense. It's the expected number of times documents of type k generate term v over the expected number of words generated by type k documents.

1.3 Algorithm

Putting all this together, we have the following algorithm:

1. Seed $\rho_k^0 = 1/k$ and draw β_k^0 randomly, for example from a uniform distribution (Dirichlet with parameters $\mathbf{1}$).

2. At iteration j:

   (a) Form $Q(\rho^{j-1}, \beta^{j-1})$ (E-step)

   (b) Update ρ^j and β^j according to (M-step 1) and (M-step 2) (M-step)

   (c) Check whether the difference between $\log\left(\Pr\left[\,\mathbf{x}\,\right]\right)$ when computed at ρ^j and β^j and at ρ^{j-1} and β^{j-1} is sufficiently small. If yes stop, if no proceed to iteration $j+1$. (Can also set maximum number of iterations).