Underflow Scaling and Smoothing in EM

The following explanations refer to the mixture of histograms lecture. The notations are defined for documents but are general and suitable to any sequence of events.

- y_t - the t^{th} document
- n_t - the length of document t
- N - number of documents
- n_{t_k} - the frequency of word k in document t
- $|V|$ - vocabulary size

Two steps in EM algorithm

E step:

$$w_{t_i} = P(x_t = i|y_t) = \frac{\alpha_i \Pi_k P_{i_k}^{n_{t_k}}}{\sum_j \alpha_j \Pi_k P_{j_k}^{n_{t_k}}} \tag{1}$$

M step:

$$\alpha_i = P(C_i) = \frac{1}{N} \sum_{t=1}^{N} w_{t_i} \tag{2}$$

$$P_{i_k} = P(w_k|C_i) = \frac{\sum_t w_{t_i} n_{t_k}}{\sum_t w_{t_i} n_t} \tag{3}$$

Underflow in the E step

P_{i_k} are small fractions and therefore $\Pi_k P_{i_k}^{n_{t_k}}$ is numerically not stable (underflow). We will solve this by scaling the values that we are computing (scaling means shifting the range of values we compute to make them numerically stable). Define z_i to be the log of the numerator of w_{t_i}:

$$z_i = \ln(\alpha_i \Pi_k P_{i_k}^{n_{t_k}}) = \ln \alpha_i + \sum_k n_{t_k} \ln P_{i_k}$$

Note: we need to insure that all parameters' values are greater than zero, otherwise the ln will not be defined (this is handled in the smoothing section). z_i is numerically stable and now Equation (1) can be written as:

$$w_{t_i} = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Now we have various e^{z_i} which are unstable. To solve this we define:

$$m = \max_i z_i$$

and now, we multiply the numerator and denominator by e^{-m} and then approximate w_{t_i} as follows:

$$w_{t_i} = \frac{e^{z_i}}{\sum_j e^{z_j}} = \frac{e^{z_i - m}}{\sum_j e^{z_j - m}} = \begin{cases} 0 & z_i - m < -k \\ \frac{e^{z_i - m}}{\sum_{\{j : z_j - m \geq -k\}} e^{z_j - m}} & \text{otherwise} \end{cases} \tag{4}$$

Where k is a parameter, chosen as described below. In the numerator, for very small values of $e^{z_j - m}$, we round the probability for cluster i given document t to zero and therefore we set $w_{t_i} = 0$. In the denominator we ignore values of $e^{z_j - m}$ that are smaller than e^{-k}, since these values stands in e^{-k} proportion to the maximal factor in the sum.

We choose k such that e^{-k} will not be underflow. A reasonable value for k is $k = 10$. Now everything is stable since $0 \geq z_i - m \geq -10$ and therefore $1 \geq e^{z_i - m} \geq e^{-10} \approx 0.000045$ which we can still handle.

Smoothing in the M step

1. We use Equation (2) to calculate α_i. If for some j we will get $\alpha_j = 0$ then in the next E step w_{t_j} will be zero and then $\ln \alpha_i$ is not defined and we cannot calculate z_j. To avoid this we set a threshold ε and if α_j becomes less than it we can do one of two:

   - fix α_j to ε.
   - throw away the cluster j, since no document seems to fall into it, and reduce the number of clusters.

   Either way, after applying this fix we need to make sure that $\sum_i \alpha_i = 1$.
   This can be done by re-computing for all j the normalized value of α_j as follows: $\alpha'_j = \frac{\alpha_j}{\sum_i \alpha_i}$

2. The same problem may occur in Equation (3). The solution here is to smooth P_{i_k}. We can use Lidstone smoothing for instance:

$$P_{i_k} = \frac{\sum_t (w_{t_i} n_{t_k}) + \lambda}{\sum_t (w_{t_i} n_t) + |V|\lambda} \tag{5}$$

Underflow in calculating the Likelihood

After each round of E and M steps we should calculate the Likelihood and verify that it has increased from the previous round. This is a great debugging tool, if in some round we find that the Likelihood decrease - it means that we have a bug in our implementation or that we are smoothing too aggressively. In the previous sections z_i was defined in the scope of a specific document t. In computing the likelihood of all documents we need to define z_i^t as the z_i of document t (and m^t in the same manner).

$$L = P(y_1 \dots y_N) = \Pi_t (\sum_i \alpha_i \Pi_k P_{i_k}^{n_{t_k}}) = \Pi_t (\sum_i e^{\ln(\alpha_i \Pi_k P_{i_k}^{n_{t_k}})}) \tag{6}$$

$$\Downarrow$$

$$\ln L = \sum_t \ln(\sum_i e^{z_i^t})$$

We may encounter underflow in $\ln(\sum_i e^{z_i})$. To avoid this we implement the same solution as in Equation (4):

$$\ln(\sum_i e^{z_i}) = \ln(e^m \sum_i e^{z_i - m}) \approx m + \ln(\sum_{\{i : z_i - m \geq -k\}} e^{z_i - m})$$

$$\Downarrow$$

$$\ln L \approx \sum_t (m^t + \ln(\sum_{\{i: z_i - m^t \geq -k\}} e^{z_i^t - m^t})) \tag{7}$$

and this is numerically stable.