1 Introduction to MMM and EM

1.1 Model definition

n - number of topics $\{1, ..., n\}$, m - number of words $\{1, ..., m\}$. Denote the model as $\theta = (\pi, e)$, where $\pi \in \mathbb{R}^n$ and $e = e_1, ...e_n \in \mathbb{R}^m$.
We sample from the model the following way: we pick a topic Y from distribution π, then sample a word X from distribution e_Y. To sample a sequence of T draws, we repeat the process T times.

- $\Pr[Y = i] = \pi_i$
- $\Pr[X = j | Y = i] = e_{ij}$

Reminder - $\Pr[A|B] = \frac{\Pr[A,B]}{\Pr[B]}$, $\Pr[A] = \sum \Pr[A, B]$.
Using this we can derive the following:

- $\Pr[X = j, Y = i] = \Pr[Y = i] \Pr[X = j | Y = i] = \pi_i e_{ij}$
- $\Pr[X = j] = \sum_{i=1}^{n} \Pr[X = j, Y = i] = \sum_{i=1}^{n} \pi_i e_{ij}$
- $\Pr[Y = i | X = j] = \frac{\Pr[X=j,Y=i]}{\Pr[X=j]} = \frac{\pi_i e_{ij}}{\sum_{k=1}^{n} \pi_k e_{kj}}$

1.2 Estimating the model

1.2.1 Easy - learning with visible data

Given a sequence of T draws $Y = y_1...y_T$ and $X = x_1...x_T$ we wish to find parameters θ such that the complete likelihood $\Pr[X, Y|\theta]$ is maximized. If we denote $A_i = |\{1 \le t \le T | y_t = i\}|$ and $E_{ij} = |\{1 \le t \le T | y_t = i, x_t = j\}|$

$$\Pr[X, Y|\theta] = \prod_{i=1}^{n} \pi_i^{A_i} \cdot \prod_{i=1}^{n} \prod_{j=1}^{m} e_{ij}^{E_{ij}}$$

. To be more efficient we can try looking into the log likelihood which looks much nicer:

$$\log \Pr[X, Y|\theta] = \sum_{i=1}^{n} A_i \log(\pi_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} E_{ij} \log(e_{ij})$$

Our goal is to maximize this subjected to $\sum_{i=1}^{n} \pi_i = 1$ and $\sum_{j=1}^{m} e_{ij} = 1$ for any $1 \le i \le n$. In other words (using lagrange multiplier) we want to maximize:

$$f(\theta, \delta_\pi, \delta_{e_i}, ..., \delta_{e_n}) = \sum_{i=1}^{n} A_i \log(\pi_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} E_{ij} \log(e_{ij}) - \delta_\pi (\sum_{i=1}^{n} \pi_i - 1) - \sum_{i=1}^{n} \delta_{e_i} (\sum_{j=1}^{m} e_{ij} - 1)$$

We will derive by each of the variables and compare to 0:

$$\frac{\partial f}{\partial \pi_i} = \frac{A_i}{\pi_i} - \delta_\pi = 0 \Rightarrow \pi_i = \frac{A_i}{\delta_\pi}$$
$$\frac{\partial f}{\partial e_{ij}} = \frac{E_{ij}}{e_{ij}} - \delta_{e_i} = 0 \Rightarrow e_{ij} = \frac{E_{ij}}{\delta_{e_i}}$$
$$\frac{\partial f}{\delta_\pi} = \sum_{i=1}^{n} \pi_i - 1 = 0 \Rightarrow \sum_{i=1}^{n} \frac{A_i}{\delta_\pi} = 1 \Rightarrow \delta_\pi = \sum_{i=1}^{n} A_i$$
$$\frac{\partial f}{\delta_{e_i}} = \sum_{j=1}^{m} e_{ij} - 1 = 0 \Rightarrow \sum_{j=1}^{m} \frac{E_{ij}}{\delta_{e_i}} = 1 \Rightarrow \delta_{e_i} = \sum_{j=1}^{m} E_{ij}$$

In total we get

$$\pi_i = \frac{A_i}{\sum_{i=1}^{n} A_i} \quad e_{ij} = \frac{E_{ij}}{\sum_{j=1}^{m} E_{ij}}$$

.

1.3 Harder - learning with hidden data

In most real life problems Y is not given and only $X = x_1...x_T$ ($Y = y_1...y_T$ is hidden) so instead of maximizing the complete likelihood we want to maximize the probability of what we see -

$$\Pr[X|\theta] = \prod_{t=1}^{T} \Pr[x_t|\theta] = \prod_{t=1}^{T} \sum_{i=1}^{n} \Pr[x_t, y_t = i] = \prod_{t=1}^{T} \sum_{i=1}^{n} \pi_i e_{ix_t}$$

The problem - this is hard. Instead we use Expectation-Maximization algorithm (EM). This is an iterative method which promise us to get to a local maximum. We start from a random start θ_0 do the following until convergence:

1. Expectation step - $Q(\theta;\theta_{t-1}) = \mathop{\mathbb{E}}_{Y|X,\theta_{t-1}} \Big[\log \Pr[X, Y|\theta] \Big]$
2. Maximization step - $\theta_t = \operatorname*{argmax}_{\theta} Q(\theta;\theta_{t-1})$

Define for $Y = y_1...y_T$ - $E_{ij}(Y) = |\{1 \le t \le T | y_t = i, x_t = j\}|$, $A_i(Y) = |\{1 \le t \le T | y_t = i\}|$. Note $A_i(Y) = \sum_{j=1}^{m} E_{ij}$. Let us compute the Q function:

$$Q(\theta;\theta_0) = \mathop{\mathbb{E}}_{Y|X,\theta_0} \Big[\log \Pr[X, Y|\theta] \Big] = \mathop{\mathbb{E}}_{Y|X,\theta_0} \left[\sum_{i=1}^{n} A_i(Y) \log(\pi_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} E_{ij}(Y) \log(e_{ij}) \right]$$
$$= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{Y|X,\theta_0} [A_i(Y)] \log(\pi_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} \mathop{\mathbb{E}}_{Y|X,\theta_0} [E_{ij}(Y)] \log(e_{ij})$$
$$= \sum_{i=1}^{n} A_i \log(\pi_i) + \sum_{i=1}^{n} \sum_{j=1}^{m} E_{ij} \log(e_{ij})$$

Where we now define $A_i = \underset{Y|X,\theta_0}{\mathbb{E}}[A_i(Y)]$, $E_{ij} = \underset{Y|X,\theta_0}{\mathbb{E}}[E_{ij}(Y)]$. We are only left to compute $\underset{Y|X,\theta_0}{\mathbb{E}}[E_{ij}(Y)]$. Let us define for any Y, t $E_{ij}(Y,t) = \begin{cases} 1 & y_t = i, x_t = j \\ 0 & else \end{cases}$. Also for convenience $B_j = |\{1 \le t \le T | x_t = j\}|$. also By this:

$$\underset{Y|X,\theta_0}{\mathbb{E}}[E_{ij}(Y)] = \underset{Y|X,\theta_0}{\mathbb{E}}\left[\sum_{t=1}^{T} E_{ij}(Y,t)\right] = \sum_{t=1}^{T} \underset{Y|X,\theta_0}{\mathbb{E}}[E_{ij}(Y,t)] = \sum_{t=1}^{T} \underset{y_t|x_t,\theta_0}{\mathbb{E}}[(Y_t = i, x_t = j)]$$
$$= \sum_{t|x_t=j} \Pr[y_t = i | x_t = j, \theta_0] = B_j \cdot \Pr[y_t = i | x_t = j, \theta_0] = B_j \cdot \frac{\pi_{0_i} e_{0_{ij}}}{\sum_{k=1}^{n} \pi_{0_k} e_{0_{kj}}}$$

So for the E-step We will compute E_{ij} for all pairs i, j, with this we can also get A_i. And for the M-step We want to maximize the Q function, under the conditions $\sum_i \pi_i = 1, \sum_j e_{ij} = 1$, which is exactly what we did in the easy case! To wrap everything up:

initialization - Given $X = x_1 ... x_T$ and ε, create $B_j = |\{1 \le t \le T | x_t = j\}|$ and initialize θ_0 randomly.

E-step - For each i, j, $E_{ij} = B_j \Pr[Y = i | X = j, \theta_0]$, for each i $A_i = \sum_{j=1}^{m} E_{ij}$.

M-step - For each i, j, $e_{ij} = \frac{E_{ij}}{\sum_{k=1}^{m} E_{kj}}$, for each i $A_i = \frac{A_i}{\sum_{j=1}^{n} A_j} = \frac{A_i}{T}$. Set $\theta_1 = (\pi, e)$

convergence - check convergence:

- If $\log(\Pr[X|\theta_1]) - \log(\Pr[X|\theta_0]) < \varepsilon$ set $\theta = \theta_1$ and finish.
- Else set $\theta_0 = \theta_1$ return to E-step.

Note all computations should be numerically stable and should be done with log probabilities. Denote $\tilde{\pi} = \log \pi$ and $\tilde{e} = \log e$:

Multiplication -

$$\Pr[X = j, Y = i] = \pi_i e_{ij} \rightarrow \log \Pr[X = j, Y = i] = \tilde{\pi_i} + \tilde{e_{ij}}$$

Summation -

$$\Pr[X = j] = \sum_{i=1}^{n} \pi_i e_{ij} \rightarrow \log \Pr[X = j] = \text{logsumexp}_{i=1}^{n}(\tilde{\pi_i} + \tilde{e_{ij}})$$

(note logsumexp is a premade scipy/numpy function)