Understanding EM algorithm

Previously on our paper: Context-aware Location Annotation on Mobility Records through User Grouping, we use *Viterbi* algorithm for the inference of HMM. *Viterbi* algorithm is a specific algorithm under in framework of Expectation-Maximization (EM) algorithm. Recently when I look into the clustering algorithm with the help of JJ and Wenbo, I try to understand EM algorithm again.

Problem formulation

Given a set of observable variables $X = (x_1, x_2, \ldots, x_m)$ and unknown (latent) variables Z, we want to estimate parameters in a model $f(X, Z)$ to the data, where the likelihood is given by

$$\mathcal{L}(\theta) = f(X|\theta) = \prod_{i=1}^{m} p(x_i|\theta) \doteq \sum_{i=1}^{m} log\, p(x_i|\theta)$$

Then our solution is for maximum likelihood estimation (here we use x to replace X for simplicity) :

$$(MLE)\ \theta^* = argmax_\theta f(x|\theta)$$
$$\Leftrightarrow \theta^* = argmax_\theta\ log f(x|\theta)$$

If both of above equation is hard to solve directly, we could:

$$\theta^* = argmax_\theta\ log \int f(x|z,\theta)dz$$
$$= argmax_\theta\ log\ \Sigma_z f(x|z,\theta)$$

We could directly maximize $\Sigma_z log\, p(x, z|\theta)$ using a gradient method (e.g., gradient ascent, conjugate gradient, quasi-Newton) but sometimes the gradient is hard to compute, hard to implement, or we do not want to bother adding in a black-box optimization routine.

Assuming $f(x|z,\theta)$, $f(z|\theta)$ is known and easy to compute, which in most case is true. For example, in Gaussian Mixture Model for clustering, $f(x|z,\theta)$ indicates which cluster x belongs to, $f(z|\theta)$ is the initial Gaussian distribution.

[Our presentation will focus on the maximum likelihood case (ML-EM); the maximum a posteriori case (MAP-EM) is very similar.]

Iteratively solving the θ^*

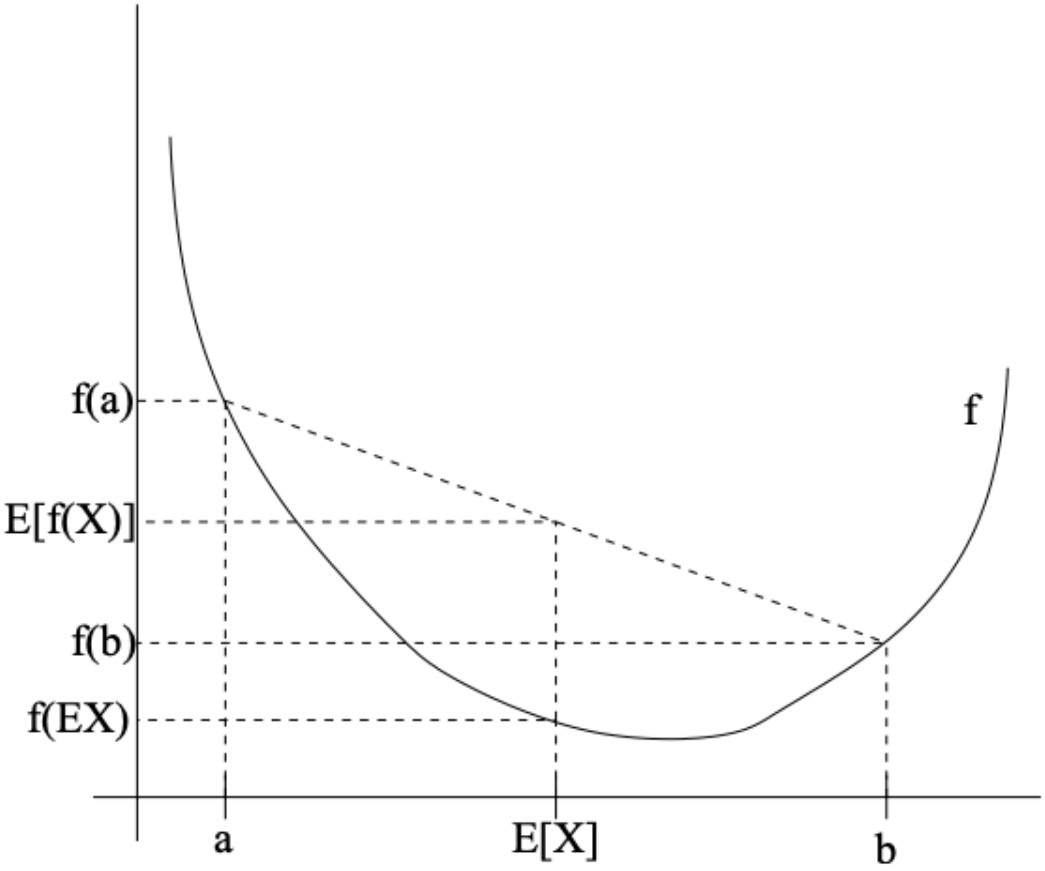
$$
\begin{aligned}
&\theta^{(n+1)} \\
&= argmax_\theta \ log \int f(x|z,\theta)dz \\
&= argmax_\theta \log \int \frac{f(x,z|\theta)}{f(z|x,\theta^{(n)})} f(z|x,\theta^{(n)})dz \\
&= argmax_\theta \ log \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[\frac{f(x,z|\theta)}{f(z|x,\theta^{(n)})}]
\end{aligned} \tag{1}
$$

Why should we divide $f(z|x,\theta^{(n)})$? We will answer this later.

Based on the Jensen's Inequality, since log function is concave, we have:

$$
log\ \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[\frac{f(x,z|\theta)}{f(z|x,\theta^{(n)})}] \geq \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log\ \frac{f(x,z|\theta)}{f(z|x,\theta^{(n)})}] \tag{1.1}
$$

A figurative illuatration about Jensen's Inquality is as follows (in this case it's a convex example):



[Source] (http://cs229.stanford.edu/notes/cs229-notes8.pdf)

Following above equation:

$$
\begin{aligned}
&log\ \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[\frac{f(x.\,z|\theta)}{f(z|x,\theta^{(n)})}] \geq \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log\ \frac{f(x,z|\theta)}{f(z|x,\theta^{(n)})}] \\
&= \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log\ f(x,z|\theta)] - \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log\ f(z|x,\theta^{(n)})]
\end{aligned} \tag{3}
$$

Usually in an iterative fashion, $\theta^{(n)}$ is known, the second term is a constant. Then:

$$\begin{aligned}\theta^{(n+1)} &= argmax_\theta \ log \, \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[\frac{f(x.\,z|\theta)}{f(z|x,\theta^{(n)})}] \\ &\geq argmax_\theta \ \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log \ f(x,z|\theta)] - \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[f(z|x,\theta^{(n)})] \\ &= argmax_\theta \ \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log \ f(x,z|\theta)]\end{aligned}$$

Usually we define:

$$Q(\theta|\theta^{(n)}) = \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log \ f(x,z|\theta)]$$

Therefore, we have

$$\begin{aligned}logf(x|\theta) = \ & log \, \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[\frac{f(x.\,z|\theta)}{f(z|x,\theta^{(n)})}] \\ \geq \ & Q(\theta|\theta^{(n)})\end{aligned} \tag{4}$$

Remark:

$Q(\theta|\theta^{(n)})$ is a lower bound of the objective function $\mathcal{L}(\theta)$

The Classical EM Algorithm

0. GMM as an example

For GMM, we have:

$$\begin{aligned}(x_i|z_k,\theta) &\overset{iid}{\sim} \mathcal{N}(\mu_k,\Sigma_k) \\ (z|\theta) &\sim Cate(\alpha)\end{aligned} \tag{g.1}$$

The goal of GMM is to find a $\theta = (\alpha \in \mathbb{R}^K, \mu \in \mathbb{R}^{K\times P}, \Sigma \in \mathbb{R}^{K\times P\times P})$ that:

$$argmax_\theta \ \prod_{i=1}^{N}\sum_k \alpha_k \cdot f(x_i|z_k,\theta)$$

Our strategy will be to instead repeatedly construct a lower-bound on ` (E-step), and then optimize that lower-bound (M-step).

- E-step

  The E-step of the EM algorithm computes $Q(\theta|\theta^{(n)}) = \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[\frac{f(x.z|\theta)}{f(z|x,\theta^{(n)})}]$ and repeatedly construct a lower-bound on $\mathcal{L}(\theta)$

  - given the observed data X , and the current parameter estimate $\theta^{(n)}$.
- M-step:

- The M-step consists of maximizing over θ the expectation computed above. That is, we set

$$\theta^{(n+1)} := argmax_\theta \; Q(\theta; \theta^{(n)})$$

1. Solving Expectation in E-step

Remember that from Equation 4 we know $Q(\theta|\theta^{(n)})$ us a lower bound of $\mathcal{L}(\theta)$. Then if we consistently push up the $Q(\theta|\theta^{(n)})$, then we will some how get a good estimation on $\mathcal{L}(\theta)$.

Then we will show how to compute $Q(\theta|\theta^{(n)})$.

$$\begin{aligned} Q(\theta|\theta^{(n)}) &= \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log\ f(x|z^{(n)},\theta^{(n)}) \cdot f(z^{(n)}|\theta^{(n)})] \\ &= \sum_k f(z_k^{(n)}|x,\theta^{(n)}) \cdot log[f(x|z_k^{(n)},\theta^{(n)}) \cdot f(z_k^{(n)}|\theta^{(n)})]) \text{ (this will be used for calculation)} \\ &= \sum_k \sum_{i=1}^N r_{ik}^{(n)}\ log\ [f(x_i|z_k^{(n)},\theta^{(n)}) \cdot f(z_k^{(n)}|\theta^{(n)})] \\ &= \sum_k \sum_{i=1}^N r_{ik}^{(n)}\ log\ f(x_i|z_k^{(n)},\theta^{(n)}) + \sum_k \sum_{i=1}^N r_{ik}^{(n)}\ log\ [f(z_k^{(n)}|\theta^{(n)})] \end{aligned} \tag{6}$$

Equation 6 can be estimated easily.

$$f(z_k|x_i,\theta) = \frac{f(x_i|z_k,\theta)f(z_k,\theta)}{f(x_i,\theta)} = \frac{f(x_i|z_k,\theta)f(z_k|\theta)}{f(x_i|\theta)} = \frac{f(x_i|z_k,\theta)f(z_k|\theta)}{\sum_k f(x_i|z_k,\theta)f(z_k|\theta)} \triangleq r_{ik} \tag{5}$$

For GMM, calculating r_{ik} is easy, $r_{ik} = \frac{\mathcal{N}(\mu_k,\Sigma_k)\cdot\alpha_k}{\sum_k[\mathcal{N}(\mu_k,\Sigma_k)\cdot\alpha_k]}$, which indicates the probability of x_i belongs to cluster z_k.

Here, we have $\sum_k r_{ik} = 1$ and $r_{ik} \geq 0, \forall k$, this will be used later.

Since we can compute r_{ik} by summing weights, and $f(x_i|z_k^{(n)},\theta^{(n)})$ is a drawn from a known distribution as in $g.1$, the same goes with $f(z_k^{(n)}|\theta^{(n)})$.

2. Maximizing in M-step

Maximize the M-step is equal to solving the following optimization problem:

$$\text{find a theta } \theta \text{ that satisfies:} \tag{p.1}$$

$$max_\theta Q(\theta|\theta^{(n)}) \triangleq max_\theta \sum_k \sum_{i=1}^N r_{ik}^{(n)}\ log\ f(x_i|z_k^{(n)}, \theta^{(n)}),$$

$$s.t.\ \sum_{k=1}^K \alpha_k = 1, \alpha_k \geq 0, \forall k = 1, \dots, K$$

Using K.T.T condition, we can solve its conjecuate problem:

$$min_\theta \mathcal{L}'(\theta) = min_\theta - \sum_k \sum_{i=1}^N r_{ik}^{(n)}\ log\ f(x_i|z_k^{(n)}, \theta^{(n)}) + \lambda(\sum_{k=1}^K \alpha_k - 1) - \sum_{j=1}^K \eta_j \alpha_j$$

$$s.t.$$

$$\frac{\partial \mathcal{L}'}{\partial \theta} = 0$$

$$\lambda \neq 0$$

$$\eta_j \geq 0$$

$$\eta_j \alpha_j(\theta^*) = 0$$

$$\sum_{k=1}^K \alpha_k(\theta^*) - 1 = 0, j = 1, \dots, K$$

$$\alpha_k(\theta^*) \geq 0$$

For GMM, solving above equation leads to:

$$\mu_k = \frac{\sum_{i=1}^N r_{ik} x_i}{\sum_{i=1}^N r_{ik}}$$

$$\Sigma_k = \frac{1}{\sum_{i=1}^N r_{ik}} \sum_{i=1}^N r_{ik}(x_i - \mu_k)(x_i - \mu_k)^T,$$

$$k = 1, \dots, K$$

Why will EM converge?

Now we need to prove:

$$logf(x|\theta^{(n+1)}) \geq logf(x|\theta^{(n)})$$

which shows EM always monotonically improves the log-likelihood.

The key to showing this result lies in our choice of the $f(z|x, \theta^{(n)})$.

By combining Equation 1 and 1.1 :

$$
\begin{aligned}
log f(x|\theta^{(n+1)}) &\geq Q(\theta^{(n+1)}|\theta^{(n)}) \text{ (this is because Jensen's Inequality)} \\
&= \mathbb{E}_{z\sim f(z|x,\theta^{(n)})}[log\ \frac{f(x,z|\theta^{(n+1)})}{f(z|x,\theta^{(n)})}] \\
&= \sum_i^N \sum_k^K f(z_k^{(n)}|x_i,\theta^{(n)})[log\ \frac{f(x_i,z_k^{(n)}|\theta^{(n+1)})}{f(z_k^{(n)}|x_i,\theta^{(n)})}] \\
&\geq \sum_i^N \sum_k^K f(z_k^{(n)}|x_i,\theta^{(n)})[log\ \frac{f(x_i,z_k^{(n)}|\theta^{(n)})}{f(z_k^{(n)}|x_i,\theta^{(n)})}] \text{ (this is because } \theta^{(n+1)} = argmax_\theta Q(\theta|\theta^{(n)}) \\
&= \sum_i^N \sum_k^K f(z_k^{(n)}|x_i,\theta^{(n)})[log\ \frac{f(x_i,z_k^{(n)}|\theta^{(n)})\cdot f(\theta^{(n)})\cdot f(x_i|\theta^{(n)})}{f(z_k^{(n)}|x_i,\theta^{(n)})\cdot f(\theta^{(n)})\cdot f(x_i|\theta^{(n)})}] \\
&= \sum_i^N \sum_k^K f(z_k^{(n)}|x_i,\theta^{(n)})[log\ f(x_i|\theta^{(n)})] \\
&= \sum_i^N log\ f(x_i|\theta^{(n)}) \cdot \sum_k^K f(z_k^{(n)}|x_i,\theta^{(n)}) \\
&= \sum_i^N log\ f(x_i|\theta^{(n)}) \text{ (this is because } \textstyle\sum_k^K f(z_k^{(n)}|x_i,\theta^{(n)}) = 1 \text{)} \\
&= log\ f(x|\theta^{(n)})
\end{aligned}
$$

Hence, EM causes the likelihood to converge monotonically. In our description of the EM algorithm, we said we'd run it until convergence. Given the result that we just showed, one reasonable convergence test would be to check if the increase in θ between successive iterations is smaller than some tolerance parameter, and to declare convergence if EM is improving θ too slowly.

References

[1] CS229 Lecture notes

[2] The EM Algorithm for Gaussian Mixtures, CS 274A

[3] The EM Algorithm

[4] IEOR E4570: Machine Learning for OR&FE