STAT 460/560 Class 19: Missing data, latent variables, and the EM algorithm

Ben Bloem-Reddy

Reading: Chapter 9.13 EM Algorithm, [Was04]. Supplemental: Ch. 9.5, [EH21]. Comprehensive: Ch. 9, [Bis06]. Advanced: Ch. 20.2, [Das11].

The censored data we encountered last class is a special type of **missing data**. Today we'll study an algorithm for handling missing data and other types of unobserved variables. The Expectation-Maximization (EM) algorithm is widely used, but we should note that it is not the only approach to these types of problems. We should also note that it works best when the expectation can be computed in closed form, which often is not the case.

Suppose we have random variables $X^n = (X_1, \dots, X_n)$, which we will observe, and $Z^m = (Z_1, \dots, Z_m)$, which we will not observe. Z could represent censored event times or other missing data, or it could represent a collection of **latent variables** that are part of the model. Then the likelihood of x^n is (assuming we have a parametric model)

$$\mathcal{L}_n(\theta) = f(x^n; \theta) = \int f(x^n, z^m; \theta)\, dz^m = \mathbb{E}_\theta[f(X^n \mid Z^m; \theta) | X^n = x^n] \,. \tag{19.1}$$

We'll consider the case of latent variables today, but the same basic ideas apply to other scenarios.

1. Gaussian mixture models

Suppose that each observation X_i is an IID sample from a **mixture of Gaussians** with K components, so that with $\phi(x; \mu, \Sigma)$ the PDF of the multivariate Gaussian distribution,

$$f(x_i; \theta) = \sum_{k=1}^K \pi_k \phi(x_i; \mu_k, \Sigma_k) \,. \tag{19.2}$$

Here, $(\pi_k)_{k=1}^K$ are the **mixture probabilities** so that each $\pi_k \in (0, 1)$ and $\sum_k \pi_k = 1$; μ_k and Σ_k are the mean vector and covariance matrix, respectively, of the k-th mixture component.

This can be recast as a latent variable model, with each observation X_i having a corresponding unobserved component assignment vector $Z_i = (Z_{i1}, \dots, Z_{iK})$ sampled from the categorical distribution with category probabilities $(\pi_k)_{k=1}^K$. (That is, each Z_{ik} is either 0 or 1, and only one entry in the vector is equal to 1.) Then

$$f(x_i \mid z_i; \theta) = \sum_{k=1}^K z_{ik} \phi(x_i; \mu_k, \Sigma_k) = \prod_{k=1}^K \phi(x_i; \mu_k, \Sigma_k)^{z_{ik}} \,, \tag{19.3}$$

so that (19.2) is recovered when we take the expectation with respect to Z_i, as in (19.1). We also see that the "complete data" likelihood, or joint likelihood, is

$$f(x^n, z^n; \theta) = \prod_{i=1}^n \prod_{k=1}^K \phi(x_i; \mu_k, \Sigma_k)^{z_{ik}} \pi_k^{z_{ik}} \,.$$

2. The EM algorithm

The EM algorithm can be used to estimate the entire set of parameters $\theta = ((\pi_k)_{k=1}^K, (\mu_k, \Sigma_k)_{k=1}^K)$. We'll keep things simple today and assume that $X_i \in \mathbb{R}$, and that we know $(\pi_k)_{k=1}^K$ and $(\sigma_k)_{k=1}^K$. A nice exposition of the general case can be found in [Bis06].

Here's the general (abstract) EM algorithm.

Algorithm 19.1.

1. Set an initial value $\theta^{(0)}$.
2. For $j = 1, 2, \ldots,$ until convergence:
   (a) **E-step**: Calculate

$$J(\theta|\theta^{(j-1)}) = \mathbb{E}_{\theta^{(j-1)}}\left[\ln \frac{f(X^n, Z^m; \theta)}{f(X^n, Z^m; \theta^{(j-1)})}\middle| X^n = x^n\right] . \tag{19.4}$$

   (b) **M-step**: Calculate

$$\theta^{(j)} = \arg\max_\theta J(\theta|\theta^{(j-1)}) . \tag{19.5}$$

What does this look like for the Gaussian mixture model? The joint log-likelihood is

$$\ln f(x^n, z^n; \theta) \propto \sum_{i=1}^n \sum_{k=1}^K -z_{ik} \ln(2\pi\sigma_k) - \frac{z_{ik}}{2\sigma_k^2}(x_i - \mu_k)^2 + z_{ik} \ln \pi_k . \tag{19.6}$$

In the simplified setting where we only need to estimate the μ_k's, we can neglect the terms that don't involved μ_k. Thus, the E-step then requires that we find

$$\mathbb{E}_\theta[\ln f(X^n, Z^n; \theta)|X^n = x^n] = -\sum_{i=1}^n \sum_{k=1}^K \mathbb{E}_\theta[Z_{ik}|X^n = x^n] \frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 . \tag{19.7}$$

Activity 19.1. Use Bayes' rule to show that for $k = 1, \ldots, K$,

$$\mathbb{E}_\theta[Z_{ik}|X^n = x^n] = \mathbb{E}_\theta[Z_{ik}|X_i = x_i] = \frac{\pi_k \phi(x_i; \mu_k, \sigma_k^2)}{\sum_{\ell=1}^K \pi_\ell \phi(x_i; \mu_\ell, \sigma_\ell^2)} . \tag{19.8}$$

For convenience, let's denote $\gamma_{ik}(\theta) = \mathbb{E}_\theta[Z_{ik}|X_i = x_i]$. Then

$$J(\theta|\theta^{(j)}) = -\sum_{i=1}^n \sum_{k=1}^K \frac{1}{2\sigma_k^2} \gamma_{ik}(\theta^{(j)}) \left((x_i - \mu_k)^2 - (x_i - \mu_k^{(j)})^2\right) . \tag{19.9}$$

Activity 19.2. Show that

$$\mu_k^{(j+1)} = \frac{\sum_{i=1}^n x_i \gamma_{ik}(\theta^{(j)})}{\sum_{i=1}^n \gamma_{ik}(\theta^{(j)})} . \tag{19.10}$$

That's it (for this example). Since both the E- and M-step can be expressed analytically, implementing the algorithm is straightforward, switcihing between the two.

3. Monotone ascent of EM

The EM algorithm has a nice property: at each step j, the (log-)likelihood is non-decreasing. To see this, recall that the Kullback–Leibler (KL) divergence between two probability densities is

$$D(p \mid\mid q) = \int p(x) \ln \frac{p(x)}{q(x)} dx \, . \tag{19.11}$$

This can also be applied to conditional probability densities.

Activity 19.3. Show that if $\theta^{(j)}$, $\theta^{(j+1)}$ are iterates of the EM algorithm, then

$$\ln \frac{f(x^n; \theta^{(j+1)})}{f(x^n; \theta^{(j)})} = J(\theta^{(j+1)} | \theta^{(j)}) + D(f(z^m | x^n; \theta^{(j)}) \mid\mid f(z^m | x^n; \theta^{(j+1)})) \, . \tag{19.12}$$

Hence, since $\theta^{(j+1)}$ was chosen to maximize $J(\theta | \theta^{(j)})$, it must be that $J(\theta^{(j+1)} | \theta^{(j)}) \geq J(\theta^{(j)} | \theta^{(j)}) = 0$. Moreover, the KL divergence is non-negative, so the log-likelihood ratio on the left-hand side of (19.12) must be non-negative, which implies that the likelihood ratio is at least 1. Hence, the likelihood is non-decreasing.

If the log-likelihood is concave then this guarantees convergence to the global maximum. If it is not then the best we can say is that the EM algorithm will converge to a stationary point (local maximum, local minimum, or saddle point); which stationary point depends on the initial value $\theta^{(0)}$.

4. Extensions of EM

The Gaussian mixture example was an especially nice setting: both the E-step and the M-step could be computed in closed form. (This will generally be possible with mixtures of exponential family distributions and other nice cases.) If the M-step can't be computed in closed form then numerical optimization can be used for each M-step. Given the effectiveness of numerical optimization, this is typically not a bad situation to be in.

If the E-step cannot be computed in closed form then we have to resort to approximating the associated integral. The need to estimate/approximate integrals is a never-ending source of statistics research problems. The simplest approach is Monte Carlo EM, which (you guessed it) approximates the expectation with a Monte Carlo average. However, this requires a lot of sampling (since the M-step requires the expectation as a function of θ), at which point we might be inclined to try more sophisticated sampling methods like the Gibbs sampler (which is structurally very similar to EM).

References

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

[Das11] A. DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*. Springer New York, 2011.

[EH21] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2021.

[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.