Notes on EM for mixture of two Gaussians with known τ and σ

May 31, 2016

Setup

We will assume that our data $x_1, \ldots, x_n$ comes from a mixture of two Gaussians, where $\tau_1 = \tau_2 = 0.5$, $\sigma_1 = \sigma_2 = \sigma$, with σ *known*, and unknown parameters μ_1 and μ_2.

- This means that we are *assuming* that each data point x_i was generated by first tossing a coin with $Pr(\text{Heads}) = \tau_1 = 0.5$, i.e. a fair coin. If it comes up heads, a point is sampled from $N(\mu_1, \sigma^2)$. Alternatively, if it comes up tails, a point is sampled from $N(\mu_2, \sigma^2)$. Let $f_1(x)$ (resp. $f_2(x)$) be the density for the first (respectively second) normal. As you know
$$f_j(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_j)^2/(2\sigma^2)},$$
where j is either 1 or 2. We only get to see the sampled numbers x_i.

- We are *assuming* that we know σ, but not μ_1 or μ_2.

- Given the values $x_1, \ldots, x_n$ and σ, our goal is to find a maximum likelihood estimate $\hat{\theta}_1$ of μ_1 and a maximum likelihood estimate $\hat{\theta}_2$ of μ_2.

We'll see how to use the EM algorithm to do this:

Definition 0.1. Denote by $\theta^t = (\theta_1^t, \theta_2^t)$ our maximum likelihood estimates for (μ_1, μ_2) after the t^{th} iteration of EM.

Definition 0.2. For $j = 1, 2$, define

$$z_{ij} := \begin{cases} 1 & x_i \text{ was sampled from normal distribution } j \text{, i.e. } N(\mu_j, \sigma^2) \\ 0 & \text{otherwise.} \end{cases}$$

EM Algorithm

- Choose some initial value θ^0.
- Repeat until convergence for $t = 0, 1, \ldots$

  E for **Expectation**: Assuming that θ^t are the correct values of the parameters, for each point x_i, determine $E(z_{ij}|x_i) = Pr(z_{ij} = 1|x_i)$.

  M for **Maximization**: Given knowledge of $E(z_{ij}|x_i)$ for each j, and σ, find θ that maximizes
  $$E(\text{LogLikelihood}(x_1, \ldots, x_n, z_{11}, \ldots z_{n1}|\theta)).$$
  Set θ^{t+1} to be this maximizing choice of the parameters θ.

Details of each step

Setup

There are many options for how to choose θ^0. For example:

- Pick two random x_i's as the initial means.
- Pick two random numbers between the minimum of the x_i's and the maximum of the x_i's.
- Sort the x_i's from largest to smallest, and take one mean to be $\lceil n/3 \rceil^{rd}$ largest and the other to be $\lceil 2n/3 \rceil^{rd}$ largest.

Expectation step

As derived in class via Bayes rule, (for $\tau_1 = \tau_2 = 0.5$)

$$E(z_{i1}|x_i) = Pr(z_{i1} = 1|x_i) = \frac{f_1(x_i|\theta^t)}{f_1(x_i|\theta^t) + f_2(x_i|\theta^t)}.$$

($E(z_{i1}|x_i) = Pr(z_{i1} = 1|x_i)$ since z_{i1} is a Bernoulli random variable.) Substituting in, we get

$$Pr(z_{i1} = 1|x_i) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - \theta_1^t)^2/(2\sigma^2)}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - \theta_1^t)^2/(2\sigma^2)} + \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - \theta_2^t)^2/(2\sigma^2)}} = \frac{e^{-(x_i - \theta_1^t)^2/(2\sigma^2)}}{e^{-(x_i - \theta_1^t)^2/(2\sigma^2)} + e^{-(x_i - \theta_2^t)^2/(2\sigma^2)}}$$

Observe that of course $Pr(z_{i1} = 1) + Pr(z_{i2} = 1) = 1$, since $z_{i1} = 1$ means that the point x_i was drawn from first Gaussian, and $z_{i2} = 1$ means x_i was drawn from the second Gaussian (so $z_{i1} = 1$ if and only if $z_{i2} = 0$). Note that z_{i1} and z_{i2} are definitely not independent; each one of them determines the other's value.

Maximization step

First, we assume that z_{i1} and z_{i2} are known (i.e. 0/1 valued) and figure out the loglikelihood function.

To this end, observe that, assuming $\tau_1 = \tau_2 = 0.5$

$$\text{Likelihood}(x_i, z_{ij}|\theta) = L(x_i, z_{ij}|\theta) = \begin{cases} 0.5 f_1(x_i|\theta) & z_{i1} = 1 \\ 0.5 f_2(x_i|\theta) & z_{i2} = 1. \end{cases} \tag{0.1}$$

and, since $x_1, \dots, x_n$ are independent

$$\text{Likelihood}(x_1, \dots, x_n, z_{11}, \dots z_{n1}|\theta) = \prod_{i=1}^{n} L(x_i, z_{ij}|\theta).$$

However, the above form of likelihood function (0.1) is inconvenient to work with. We rewrite it as follows:

$$L(x_i, z_{ij}|\theta) = [0.5 f_1(x_i|\theta)]^{z_{i1}} \cdot [0.5 f_2(x_i|\theta)]^{z_{i2}}.$$

Since exactly one of z_{i1} and z_{i2} is 1 (and the other is 0), this gives us *exactly* the same formula as we got in (0.1).

Now we are good to go:

$$\text{Likelihood}(x_1, \dots, x_n, z_{11}, \dots z_{n1}|\theta) = \prod_{i=1}^{n} L(x_i, z_{ij}|\theta) = \prod_{i=1}^{n} \Big([0.5 f_1(x_i|\theta)]^{z_{i1}} \cdot [0.5 f_2(x_i|\theta)]^{z_{i2}} \Big).$$

An immediate observation is that, since $z_{i1} + z_{i2} = 1$ **always**, this simplifies to

$$\prod_{i=1}^{n} 0.5^{z_{i1}+z_{i2}} \Big([f_1(x_i|\theta)]^{z_{i1}} \cdot [f_2(x_i|\theta)]^{z_{i2}} \Big) = \prod_{i=1}^{n} 0.5 \Big([f_1(x_i|\theta)]^{z_{i1}} \cdot [f_2(x_i|\theta)]^{z_{i2}} \Big).$$

Now we compute the loglikelihood function $LL(x_1, \dots, x_n, z_{11}, \dots z_{n1}|\theta)$, which is

$$LL(\vec{x}, \vec{z_{i1}}|\theta) = n \ln 0.5 + \sum_{i=1}^{n} z_{i1} \ln f_1(x_i|\theta) + \sum_{i=1}^{n} z_{i2} \ln f_2(x_i|\theta)$$

Since

$$\ln(f_j(x_i|\theta)) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\theta_j)^2/(2\sigma^2)} \right) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \theta_j)^2}{2\sigma^2}$$

the loglikelihood function is

$$\begin{aligned} LL(\vec{x}, \vec{z_{ij}}|\theta) &= n \ln 0.5 - \sum_{i=1}^{n} \left(\frac{z_{i1}}{2} \ln(2\pi\sigma^2) + \frac{z_{i1}(x_i - \theta_1)^2}{2\sigma^2} \right) - \sum_{i=1}^{n} \left(\frac{z_{i2}}{2} \ln(2\pi\sigma^2) + \frac{z_{i2}(x_i - \theta_2)^2}{2\sigma^2} \right) \\ &= n \ln 0.5 - \frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{n} \left(\frac{z_{i1}(x_i - \theta_1)^2}{2\sigma^2} \right) - \sum_{i=1}^{n} \left(\frac{z_{i2}(x_i - \theta_2)^2}{2\sigma^2} \right). \end{aligned}$$

This last step follows again from the fact that $z_{i1} + z_{i2} = 1$ deterministically.

Finally, we are ready to actually do the M step, which is to find the maximum of $E(LL(\vec{x}, \vec{z_{i1}}|\theta)$, where the expectation is with respect to the z_{ij}'s conditioned on $\vec{x}$.

Thus our goal is to maximize

$$E(LL(\vec{x}, \vec{z_{i1}}|\theta)) = E\left[n \ln 0.5 - \frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\left(\frac{z_{i1}(x_i-\theta_1)^2}{2\sigma^2}\right) - \sum_{i=1}^{n}\left(\frac{z_{i2}(x_i-\theta_2)^2}{2\sigma^2}\right)\right].$$

which by linearity of expectation this is simply

$$E(LL(\vec{x}, \vec{z_{i1}}|\theta)) = n \ln 0.5 - \frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\left(\frac{E(z_{i1}|x_i)(x_i-\theta_1)^2}{2\sigma^2}\right) - \sum_{i=1}^{n}\left(\frac{E(z_{i2}|x_i)(x_i-\theta_2)^2}{2\sigma^2}\right).$$

Conveniently, we computed $E(z_{i1}|x_i)$ and $E(z_{i2}|x_i)$ in the expectation step.

So we merely solve the equations

$$\frac{\partial E(LL(\vec{x}, \vec{z_{i1}}|\theta))}{\partial\theta_1} = 0 \quad \text{and} \quad \frac{\partial E(LL(\vec{x}, \vec{z_{i1}}|\theta))}{\partial\theta_2} = 0$$

and check that we have found maxima. Everything separates nicely when you take the derivatives etc and it turns out that the solution is

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} E(z_{i1}|x_i)x_i}{\sum_{i=1}^{n} E(z_{i1}|x_i)} \quad \text{and} \quad \hat{\theta}_2 = \frac{\sum_{i=1}^{n} E(z_{i2}|x_i)x_i}{\sum_{i=1}^{n} E(z_{i2}|x_i)}.$$

Convergence

A common way to decide if your algorithm has converged is to pick some threshold value, and declare victory as soon as the differences in the log likelihoods between two iterations is lower than that threshold.