EM algorithm

Paul Bouyé

The **Intuition** section is based on a video by ritvikmath [2]. The rest is based on the lecture notes of CS229 [1].

Intuition

1. We have [1, 2, x] draws from a $\mathcal{N}(1, 1)$ $\rightarrow$ Best guess for x ?
   $\Rightarrow$ we take the distribution mean $\mu = 1$
2. We have [0, 1, 2] draws from a $\mathcal{N}(\mu, 1)$ $\rightarrow$ Best guess for μ ?
   $\Rightarrow$ we take the mean of the data $\mu = \frac{0+1+2}{3} = 1$
3. Now, we have [1, 2, x] draws from a $\mathcal{N}(\mu, 1)$ $\rightarrow$ Best guess for (x, μ) ?
   Game: assume $\mu_0 = 0$, then $x_0 = 0$ (like case 1)
   We have [1, 2, $x_0 = 0$], $\Rightarrow \mu_1 = \frac{1+2+0}{3} = 1$ (like case 2)
   We update x by setting $x_1 = \mu_1 = 1$, then $\mu_2 = \frac{4}{3}$
   $\Rightarrow x_2 = \frac{4}{3} \Rightarrow \mu_3 = \frac{10}{3} \Rightarrow \ldots$ it converges to $x^* = \mu^* = 1.5$ $\rightarrow$ consistent

EM algorithm

We consider an estimation problem with a training set of n independant samples $(x_1, \ldots, x_n)$ and a latent variable model $p(x, z \mid \theta)$ with z being the latent variable (which for simplicity is assumed to take a finite number of values). The density for x can be obtained by marginalizing the latent variable z:

$$p(x \mid \theta) = \sum_z p(x, z \mid \theta)$$

We define the log-likelihood:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log p(x_i \mid \theta) \\ &= \sum_{i=1}^{n} \log \left(\sum_{z_i} p(x_i, z_i \mid \theta) \right) \end{aligned}$$

$\rightarrow$ this gives a non-convex optimization problem

We will first consider optimizing the likelihood for a single example x. We optimize:

$$\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$$

Let q be a distribution over the possible values of z. We have $\sum_z q(z) = 1$.

$$\log p(x \mid \theta) = \boxed{\log \sum_z q(z) \frac{p(x, z \mid \theta)}{q(z)} \geq \sum_z q(z) \log \frac{p(x, z \mid \theta)}{q(z)}}$$

according to Jensen's inequality.

To make the bound tight for a particular value of θ, we want to approach equality. We want the expectation to be taken at a "constant"-valued random variable, i.e.:

$$\frac{p(x, z \mid \theta)}{q(z)} = c$$

Since $\sum_z q(z) = 1$, we get $c = \sum_z p(x, z \mid \theta)$ and thus $q(z) = \frac{p(x,z|\theta)}{\sum_z p(x,z|\theta)} = \frac{p(x,z|\theta)}{p(x|\theta)}$.

$$\Rightarrow q(z) = p(z \mid x, \theta) \qquad (q \text{ is the posterior distribution of } z \text{ given } x \text{ and } \theta)$$

When $q(z) = p(z \mid x, \theta)$, the equation (1) is an equality (we let this proof as an exercise).

For convenience, we define the evidence lower bound (ELBO):

$$\mathrm{ELBO}(x \mid q, \theta) = \sum_z q(z) \log \frac{p(x, z \mid \theta)}{q(z)}$$

Thus, we can re-write (1) as:

$$\boxed{\forall q, \theta, x, \quad \log p(x \mid \theta) \geq \mathrm{ELBO}(x \mid q, \theta)}$$

Intuitively, the EM algorithm alternatively updates q and θ by: a) setting $q(z) = p(z \mid x, \theta^{(t)})$ so that $\mathrm{ELBO}(x \mid q, \theta) = \log p(x \mid \theta)$ for x and the current θ b) maximizing $\mathrm{ELBO}(x \mid q, \theta)$ with respect to θ while fixing q

For multiple training examples, we simply sum the ELBOs:

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta) \geq \sum_{i=1}^{n} \mathrm{ELBO}(x_i \mid q_i, \theta) = \sum_{i=1}^{n} \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i \mid \theta)}{q_i(z_i)}$$

Algorithm

Repeat until convergence:

1. E-step: $\forall i$, set $q_i(z_i) = p(z_i \mid x_i, \theta)$.

2. M-step: Set:

$$\begin{aligned}\theta &= \arg\max_\theta \sum_{i=1}^n \mathrm{ELBO}(x_i \mid q_i, \theta) \\ &= \arg\max_\theta \sum_{i=1}^n \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i \mid \theta)}{q_i(z_i)}\end{aligned}$$

Mixture of Gaussians

We define the responsibilities $r_{ik} = q_i(z_i = k) = p(z_i = k \mid x_i, \theta)$ with $\theta = \{\pi, \mu, \Sigma\}$.

- π : cluster weight, prior probability of z
- μ : cluster mean
- Σ : cluster covariance matrix

(E-step) compute the responsibilities r_{ik}
(M-step) maximize the quantity:

$$\begin{aligned}\sum_{i=1}^n \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i \mid \pi, \mu, \Sigma)}{q_i(z_i)} &= \sum_{i=1}^n \sum_{k=1}^K q_i(z_i = k) \log \frac{p(x_i \mid z_i = k, \mu, \Sigma) p(z_i = k \mid \pi)}{q_i(z_i = k)} \\ &= \sum_{i=1}^n \sum_{k=1}^K r_{ik} \log \Big[\frac{(2\pi)^{-D/2} |\Sigma_k|^{-1} e^{-\frac{1}{2}(x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k)} \pi_k}{r_{ik}} \Big]\end{aligned}$$

We maximize this over the parameters π_m, μ_m, Σ_m.

- Maximizing over μ_m:

$$\nabla_{\mu_m} \Big(\sum_{i=1}^n \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i \mid \pi, \mu, \Sigma)}{q_i(z_i)} \Big) = 0$$

$$\begin{aligned}&\Leftrightarrow \sum_{i=1}^n r_{im} \nabla_{\mu_m} \Big[\log \Big(\frac{(2\pi)^{-D/2} |\Sigma_m|^{-1} e^{-\frac{1}{2}(x_i - \mu_m)^\top \Sigma_m^{-1} (x_i - \mu_m)} \pi_m}{r_{im}} \Big) \Big] = 0 \\ &\Leftrightarrow \sum_{i=1}^n r_{im} \nabla_{\mu_m} \Big[\mu_m^\top \Sigma_m^{-1} \mu_m - 2 x_i^\top \Sigma_m^{-1} \mu_m \Big] = 0 \\ &\Leftrightarrow \sum_{i=1}^n r_{im} \Big[2\Sigma_m^{-1} \mu_m - 2\Sigma_m^{-1} x_i \Big] = 0 \\ &\Leftrightarrow \sum_{i=1}^n r_{im} \Big[\mu_m - x_i \Big] = 0 \\ &\Leftrightarrow \boxed{\mu_m = \frac{\sum_{i=1}^n r_{im} x_i}{\sum_{i=1}^n r_{im}}}\end{aligned}$$

- Maximizing over Σ_m^{-1}:

$$\nabla_{\Sigma_m^{-1}} \sum_{i=1}^{n} -\frac{1}{2} r_{im} \Big[\log |\Sigma_m| + (x_i - \mu_m)^\top \Sigma_m^{-1} (x_i - \mu_m) \Big] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} r_{im} \Big[\Sigma_m - (x_i - \mu_m)^\top (x_i - \mu_m) \Big] = 0$$

$$\Leftrightarrow \boxed{\Sigma_m = \frac{\sum_{i=1}^{n} r_{im} (x_i - \mu_m)^\top (x_i - \mu_m)}{\sum_{i=1}^{n} r_{im}}}$$

- Maximizing over π_m:

$\nabla_{\pi_m} \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \log \pi_k = 0$ doesn't work since π_k is a probability distribution. We can use the Lagrange multipliers method to solve this problem. We define the Lagrangian:

$$\mathcal{L}(\pi, \beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \log \pi_k + \beta \Big(1 - \sum_{k=1}^{K} \pi_k \Big)$$

where β is the Lagrange multiplier. Taking the derivatives, we find:

$$\frac{\partial \mathcal{L}}{\partial \pi_m} = 0 \quad \Leftrightarrow \quad \frac{\sum_{i=1}^{n} r_{im}}{\pi_m} - \beta = 0 \quad \Leftrightarrow \quad \pi_m = \frac{1}{\beta} \sum_{i=1}^{n} r_{im}$$

Since $\sum_{m=1}^{K} \pi_m = 1$, we have:

$$\begin{aligned} \sum_{m=1}^{K} \pi_m &= \frac{1}{\beta} \sum_{m=1}^{K} \sum_{i=1}^{n} r_{im} \\ &= \frac{1}{\beta} \sum_{i=1}^{n} \sum_{m=1}^{K} r_{im} \\ &= \frac{1}{\beta} \sum_{i=1}^{n} 1 \\ &= \frac{n}{\beta} = 1 \end{aligned}$$

Thus, we have:

$$\forall\, m, \; \boxed{\pi_m = \frac{1}{n} \sum_{i=1}^{n} r_{im}}$$

References

[1] Tengyu Ma and Andrew Ng. *CS229 Lecture Notes, Part IX: The EM algorithm*. `https://cs229.stanford.edu/summer2023/cs229-notes8.pdf`. Stanford Machine Learning Course.

[2] ritvikmath. *EM Algorithm : Data Science Concepts*. `https://www.youtube.com/watch?v=xy96ArOpntA`, 2022. YouTube Video.