PGM - EM for GMM

Caio Corro

1 Gaussian mixture models

Let $\mathcal{Y}$ be a discrete latent variable taking values in $\{1...k\}$ and $\mathcal{X}$ be an observed continuous random variable taking values in $\mathbb{R}^d$. A Gaussian mixture model (GMM) defines a joint probability distribution over $\mathcal{Y}$ and $\mathcal{X}$ as follows:

$$p_\theta(\boldsymbol{x}, y) = p_\theta(y) p_\theta(\boldsymbol{x}|y)$$

where $\theta = \{\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$ are the parameters of the GMM, with $\boldsymbol{\lambda} \in \triangle(k)$, $\boldsymbol{\mu} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{\sigma}^2 \in \mathbb{R}^{k \times d}_{++}$. The two distribution are defined as follows:

$$\begin{aligned} p_\theta(y) &= \lambda_y, \\ p_\theta(\boldsymbol{x}) &= \prod_{i=1}^{d} p_\theta(x_i) = \prod_{i=1}^{d} f(x_i, \mu_{y,i}, \sigma^2_{y,i}), \end{aligned}$$

where f is the PDF of univariate Normal distributions:

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

To generate data from a GMM we can simply rely on ancestral sampling:

1. $y \sim p_\theta(\mathcal{Y})$,
2. $\boldsymbol{x} \sim p_\theta(\mathcal{X}|\mathcal{Y} = y)$,

that is, we generating a point consists of first sampling a "cluster" id from the prior $p_\theta(\mathcal{Y})$ and then sampling a point according to this cluster conditional distribution on observed values $p_\theta(\mathcal{X}|\mathcal{Y} = y)$.

2 Expectation maximization

The parameters of a GMM can be learned via gradient ascent by carefully taking care of the constraints on parameters, for example via reparameterization. Another techniques to learn the GMM parameters is via the Expectation-Maximization (EM) algorithm. One benefit (among others) of EM is that it doesn't require hyper-parameters like a step size.

Given a dataset $\mathcal{D} = \{\boldsymbol{x}\}_{i=1}^n$, the models are learned by maximizing the log-likelihood of the training data:

$$\max_{\theta \in \Theta} \sum_{\boldsymbol{x} \in \mathcal{D}} \log p_\theta(\boldsymbol{x}) = \max_{\theta \in \Theta} \sum_{\boldsymbol{x} \in \mathcal{D}} \log \sum_{y \in \{1...k\}} p_\theta(y) p_\theta(\boldsymbol{x}|y).$$

Let $q_{\boldsymbol{\phi}}(\mathcal{Y}|\mathcal{X})$ be a proposal distribution parameterized by $\boldsymbol{\phi}$ defined as follows:

$$q_{\boldsymbol{\phi}}(y|\boldsymbol{x}) = \phi_y^{(\boldsymbol{x})}$$

where $\boldsymbol{\phi}^{(\boldsymbol{x})} \in \triangle(k)$ are the parameters of the proposal associated with input $\boldsymbol{x}$. The evidence lower bound (ELBO) is a lower bound on the log-likelihood defined as follows:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] + H^{\mathrm{S}}[q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})] \\ &= \textsc{Elbo}(\boldsymbol{x}, \theta, \boldsymbol{\phi})\end{aligned}$$

where $H^{\mathrm{S}}[q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})] = -\sum_{y \in \{1 \dots k\}} q_{\boldsymbol{\phi}}(y|\boldsymbol{x}) \log q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ is the Shannon entropy. The bound can be derived using Jensen's inequality. As such, we can build a surrogate training objective as follows:

$$\max_{\theta \in \Theta} \sum_{\boldsymbol{x} \in \mathcal{D}} \log p_\theta(\boldsymbol{x}) \geq \max_{\theta \in \Theta, \boldsymbol{\phi} \in \Phi} \sum_{\boldsymbol{x} \in \mathcal{D}} \textsc{Elbo}(\boldsymbol{x}, \theta, \boldsymbol{\phi}).$$

where we need to jointly maximize the ELBO over θ and $\boldsymbol{\phi}$.

The EM algorithm is simply an algorith that maximizes the surrogate lower bound using block-coordinate ascent:

- Expectation step: maximize the objective over the proposal parameters $\boldsymbol{\phi}$;
- Maximization step: maximize the objective over the model parameters θ.

Note that the two step must interleave until convergence. For GMMs, both steps enjoy closed-form expressions.

2.1 Expectation step

We can show that, for a given point, the difference between the evidence and the ELBO is equal to the KL divergence between the proposal distribution and the true posterior distribution of the model:

$$\log p_\theta(\boldsymbol{x}) - \textsc{Elbo}(\boldsymbol{x}, \theta, \boldsymbol{\phi}) = KL[q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x}) | p_\theta(\mathcal{Y}|\boldsymbol{x})]$$

where the KL divergence is defined as follows:

$$KL[q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x}) | p_\theta(\mathcal{Y}|\boldsymbol{x})] = \sum_{y \in 1 \dots k} q_{\boldsymbol{\phi}}(y|\boldsymbol{x}) \log \frac{q_{\boldsymbol{\phi}}(y|\boldsymbol{x})}{p_\theta(y|\boldsymbol{x})}$$

The KL divergence is always non negative and is null if and only the two distributions are equal. Therefore, maximizing over the proposal distribution parameters is equal to closing the gap between the evidence and the ELBO, which become exactly null if we set the proposal equal to the posterior distribution of the model, i.e. we want the following equality to hold:

$$\forall y \in \{1 \dots k\} : \quad q_{\boldsymbol{\phi}}(y|\boldsymbol{x}) = p_\theta(y|\boldsymbol{x})$$

By Bayes theorem, the means that we can simply set the parameters $\boldsymbol{\phi}$ as follows:

$$\phi_y^{(\boldsymbol{x})} = \frac{p_\theta(y) p_\theta(\boldsymbol{x}|y)}{\sum_{y' \in \{1 \dots k\}} p_\theta(y) p_\theta(\boldsymbol{x}|y)}$$

2.2 Maximization step

For the maximization step, we can compute the closed form solution expression by solving the equations defined by first-order optimality conditions. For means of the conditional distributions, we have:

$$\frac{\partial}{\partial \mu_{y,j}} \sum_{\boldsymbol{x} \in \mathcal{D}} \textsc{Elbo}(\boldsymbol{x}, \theta, \phi) = 0$$

$$\frac{\partial}{\partial \mu_{y,j}} \left(\sum_{\boldsymbol{x} \in \mathcal{D}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] + H^{\mathrm{S}}[q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})] \right) = 0$$

$$\frac{\partial}{\partial \mu_{y,j}} \sum_{\boldsymbol{x} \in \mathcal{D}} \sum_{y' \in \{1 \dots k\}} \phi_{y'}^{(\boldsymbol{x})} \log \prod_{i=1}^{d} f(x_i, \mu_{y',i}, \sigma_{y',i}^2) = 0$$

Note that the prior and the entropy terms don't depend on the model parameters so their derivatives are null. Therefore we obtain:

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{\partial}{\partial \mu_{y,j}} \log \prod_{i=1}^{d} f(x_i, \mu_{y,i}, \sigma_{y,i}^2) = 0$$

In the sum over y', the derivative of the inner term will be null in all cases except $y' = y$, therefore:

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{\partial}{\partial \mu_{y,j}} \sum_{i=1}^{d} \left(\log 1 - \log(\sigma_{y,i}\sqrt{2\pi}) - \frac{1}{2} \left(\frac{x - \mu_{y,i}}{\sigma_{y,i}} \right)^2 \right) = 0$$

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{x - \mu_{y,j}}{\sigma_{y,j}^2} = 0$$

$$\mu_{y,j} = \frac{\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \mu_{y,j}}{\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})}}$$

which can be interpreted as a weight mean, where the weights are given by the proposal distribution.

Closed form solutions for the variance parameters can be derived in a similar fashion. Note that they are constrained to be strictly positive, but the solution will always satisfy them (except in some "pathological" cases). The details of the computation are as follows:

$$\frac{\partial}{\partial \sigma_{y,j}} \sum_{\boldsymbol{x} \in \mathcal{D}} \textsc{Elbo}(\boldsymbol{x}, \theta, \phi) = 0$$

$$\frac{\partial}{\partial \sigma_{y,j}} \left(\sum_{\boldsymbol{x} \in \mathcal{D}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] + H^{\mathrm{S}}[q_{\boldsymbol{\phi}}(\mathcal{Y}|\boldsymbol{x})] \right) = 0$$

$$\frac{\partial}{\partial \sigma_{y,j}} \sum_{\boldsymbol{x} \in \mathcal{D}} \sum_{y' \in \{0 \dots k\}} \phi_{y'}^{(\boldsymbol{x})} \log \prod_{i=1}^{d} f(x_i, \mu_{y',i}, \sigma_{y',i}^2) = 0$$

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{\partial}{\partial \sigma_{y,j}} \log \prod_{i=1}^{d} f(x_i, \mu_{y,i}, \sigma_{y,i}^2) = 0$$

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{\partial}{\partial \sigma_{y,j}} \sum_{i=1}^{d} \left(\log 1 - \log(\sigma_{y,i}\sqrt{2\pi}) - \frac{1}{2} \left(\frac{x - \mu_{y,i}}{\sigma_{y,i}} \right)^2 \right) = 0$$

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \left(-\frac{1}{\sigma_{y,j}} + \frac{(x - \mu_{y,j})^2}{\sigma_{y,j}^3} \right) = 0$$

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{1}{\sigma_{y,j}} = \sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \frac{(x - \mu_{y,j})^2}{\sigma_{y,j}^3}$$

$$\sigma_{y,j}^2 = \frac{\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})} \left(x - \mu_{y,j}\right)^2}{\sum_{\boldsymbol{x} \in \mathcal{D}} \phi_y^{(\boldsymbol{x})}}$$

Again, this term can be interpreted as a weighted variance.