Expectation Maximization for Mixture Model

dw

September 7, 2021

1 Mixture model

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. It is common to think of probability mixture modeling as a missing data problem. One way to understand this is to assume that the data points under consideration have "membership" in one of the distributions we are using to model the data. When we start, this membership is unknown, or missing. [1]

Assuming N observation (index i) are observed in K distributions(index K). The basis model is:

$$p(x_i|\theta) = \sum_k \alpha_k p(x_i|\theta_k) \tag{1}$$

- θ - parameters including θ_k and α_k.
- θ_k - parameters of k-th distribution.
- α_k - weight or prior, equal to $p(z_{ik})$.

Proof: Hidden variables z_{ik}s indicate which class the i-th observation belongs to, value 1 shows in while 0 shows not, and $\sum_k z_i k = 1$, which satisfy categarical distribution. Likelihood including the hidden variables and observations is:

$$\mathcal{L}(\theta|x_i, z_{ik}) \propto \prod_k p(x_i, z_{ik}|\theta_k) \tag{2}$$

Using formula of total probability:

It just a trick to promise $z_{ik} = 0$ make nonsense in the formula. Using categarical distribution, we have

$$\begin{aligned} \mathcal{L}(\theta|x_i) \propto \sum_k p(x_i|z_{ik}, \theta)p(z_{ik}|\theta) &= \sum_k \prod_k p^{z_{ik}}(x_i, z_{ik}|\theta_k)p(z_{ik}|\theta) \\ &= \sum_k \alpha_k p(x_i, z_{ik}|\theta_k) \end{aligned} \tag{3}$$

Don't be confused that:

- $p(x_i, z_{ik}|\theta) = \{\alpha_k p(x_i|\theta_k)\}^{z_{ik}}$
- $p(x_i|\theta, z_{ik}) = p^{z_{ik}}(x_i|\theta_k)$
- $p(z_{ik}|\theta) = \alpha_k$

2 Expectation Maximization

However, logsum is hard to optimize since all parameters are coupled, using Jensen Inequality we get the lower bound:

$$\log \mathcal{L}_i(\theta|x_i) \propto \sum_i \log p(x_i|\theta) \geq \sum_i \sum_k q(z_{ik}) \log\{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\} = F(q, \theta) \tag{4}$$

Here $q(z_{ik})$ is an arbitrary distribution of z_{ik}, satisfying $\sum_k q(z_{ik}) = 1$ only when

$$q(z_{ik}) = \frac{\alpha_k p(x_i, z_{ik})}{\sum_{k'} p(x_i, z_{ik'})} \tag{5}$$

the equality holds.

Proof: By Jensen's inequality, for the concave function (such as log) we have

$$f(\frac{\sum_i a_i x_i}{\sum_i a_i}) \geq \frac{\sum_i a_i f(x_i)}{\sum_i a_i} \tag{6}$$

When $\sum_i a_i = 1$, it becomes $f(\sum_i a_i x_i) \geq \sum_i a_i f(x_i)$ Replace a_i by $q(z_{ik})$ by $\{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\}$, we have

$$\begin{aligned} \sum_i \log p(x_i|\theta) &= \sum_i \log \sum_k p(x_i, z_{ik}|\theta) \\ &= \sum_i \log \sum_k q(z_{ik}) \{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\} \\ &\geq \sum_i \sum_k q(z_{ik}) \log \{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\} \end{aligned} \tag{7}$$

when $\{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\}$ is nothing to do with k the inequality holds. Notice that

$$p(z_{ik}|x_i, \theta) = \frac{p(x_i, z_{ik}|\theta)}{\sum_{k'} p(x_i, z_{ik'}|\theta)} = \frac{\alpha_k p(x_i|\theta_k)}{\sum_{k'} \alpha_{k'} p(x_i|\theta_{k'})} \tag{8}$$

take z_{ik} as $p(z_{ik}|x_i, \theta)$ and the $\sum_{k'} \alpha_{k'} p(x_i|\theta_{k'})$ is a constant.

3 EM flow chart

- **Initial**: Giving an initial guess of $\theta_{[0]} : \alpha_k^{[0]}, \theta_k^{[0]}$
- **E-step**: $q^{[\text{new}]} \leftarrow \arg\max_q F(q, \theta^{[\text{old}]})$

  $$\mathcal{L}(\theta^{[\text{old}]}|t) \geq F(q, \theta^{[\text{old}]}) = \sum_i \sum_k q(z_{ik}) \log \{\frac{p(t_i, z_{ik}|\theta^{[\text{old}]})}{q(z_{ik})}\} \tag{9}$$

  we get

  $$q(z_{ik}) = \frac{\alpha_k p(x_i|\theta_k^{[\text{old}]})}{\sum_{k'} \alpha_{k'} p(x_i|\theta_{k'}^{[\text{old}]})} \tag{10}$$

- **M-step**: $\theta^{[\text{new}]} \leftarrow \arg\max_\theta F(q^{[\text{new}]}, \theta)$
  By this way, the parameters in different classes can be optimized individually, Notice that:

  $$F(q, \theta) = \sum_i \sum_k q(z_{ik}) \log p(x_i, z_{ik}|\theta) - \sum_i \sum_k q(z_{ik}) \log q(z_{ik}) \tag{11}$$

  Only the former part is dependent on θ, noted as $Q(q, \theta)$

  $$\begin{aligned} Q(q^{[\text{new}]}, \theta) &= \sum_i \sum_k q^{[\text{new}]}(z_{ik}) \log p(x_i, z_{ik}|\theta) \\ &= \sum_k \left\{ \sum_i q^{[\text{new}]}(z_{ik}) \log p(x_i, z_{ik}|\theta) \right\} \end{aligned} \tag{12}$$

  The optimize target is change to single class.

- Repeat until converge.

Proof of Convergence: 1. In E-step, Jensen Inequality promise the convergence.
2. In M-step, the optimize step promise the convergence.
However, EM is not a global minimizer.

4 Alternative

E-step define the lower bounds of $\mathcal{L}(\theta|t) - \mathrm{KL}(q(z_{ik})||p(z_{ik}|x_i, \theta)$
Difference between $\log \mathcal{L}(\theta|x)$ and $F(q, \theta)$ is the KL-divergence of: Proof:

$$\begin{aligned} L(\theta) - F(q,\theta) &= \sum_i^N \sum_{k=1}^K \log\{p(x_i|\theta)\} - \sum_i^N \sum_{k=1}^K q(z_{ik}) \log\{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\} \\ &= \sum_i^N \sum_{k=1}^K q(z_{ik}) \log\{p(x_i|\theta)\} - \sum_i^N \sum_{k=1}^K q(z_{ik}) \log\{\frac{p(x_i, z_{ik}|\theta)}{q(z_{ik})}\} \\ &= \sum_i^N \sum_{k=1}^K q(z_{ik}) \log\{\frac{p(x_i|\theta)/p(x_i, z_{ik}|\theta)}{q(z_{ik})}\} \\ &= \sum_i^N \sum_{k=1}^K q(z_{ik}) \log\{\frac{p(z_{ik}|x_i, \theta)}{q(z_{ik})}\} \\ &= \mathrm{KL}(q(z_{ik})||p(z_{ik}|x_i, \theta)) \end{aligned} \tag{13}$$

The KL-divergence is always positive and have value 0 only if $q(z_{ik}) = p(z_{ik}|x_i, \theta)$.

5 Apply: Gaussian mixture model

Proof