Machine Learning notes

LWang

January 10, 2017

0.1 EM algorithm

Consider the following experiment with coin A has probability of θ_A being head, and coin B has probability θ_B flipping head , we pick one coin at a time, and flip it m times, in total, we have chosed n coins, in other words, we have n and have fliped $n \times m$ times of coin.

Complete information : Suppose we write down which coin we have picked every time, we would have the complete likelihood function $p(x, z|\theta)$. where x is a vector of m flips in one sample, z is the label of coins. For mathematical conveninet, we would like to compute $\log p(x, z|\theta)$, where

$$\log p(x, z|\theta) = \log \prod_{i=1}^{n} p(x_i, z_i|\theta) = \sum_{i=1}^{n} \log p(x_i, z_i|\theta) \tag{0.1.1}$$

the way we solve the parameters is the unsurprising Maximumizing Likelihood function(MLE).

$$\theta := \underset{\theta}{\operatorname{argmax}} \log p(x, z|\theta) \tag{0.1.2}$$

Incomplete information(missing information), we didn't record which coin we have picked, onlying knowing the flipping results. Models with hidden variables are known as latent variable models(LVM), in general, there are K laten variables(z_k, k=1,2,. . . ,K), and m visible variables(x_i, i=1,2,. . . ,m). The incomplete likelihood function is $p(x|\theta)$

$$\begin{aligned}
\log p(x|\theta) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\
&= \sum_{i=1}^{n} \log p(x_i|\theta) \\
&= \sum_{i=1}^{n} \log \sum_{k=1}^{K} p(x_i, z_i = k|\theta)
\end{aligned}$$

This likelihood function is not easy to solve since we now have unkown variable z_i and unknown parameter θ, hence, **EM algorithm ** comes to play. First of all , we introduce the distribution of hidden variable $Q_i(z_i)$.

$$\begin{aligned}
\sum_{i=1}^{n} \log \sum_{k=1}^{K} p(x_i, z_i = k|\theta) &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} Q_i(z_i) \frac{p(x_i, z_i = k|\theta)}{Q_i(z_i)} \\
&= \sum_{i=1}^{n} \log E\big[\frac{p(x_i, z_i = k|\theta)}{Q_i(z_i)}\big] \qquad (0.1.3)
\end{aligned}$$

Again , it is not so elegant, however, we have Jasen inequality to handle this complexity. Recall $f(E[x]) \geq E[f(x)]$ holds for convex functions,and the equale sign exists when $f(x)$ is a constant, now we plug this into the above equation $x \to \frac{p(x_i, z_i=k|\theta)}{Q_i(z_i)}$,

$$\begin{aligned}\sum_{i=1}^{n}\log\sum_{k=1}^{K}p(x_i,z_i=k|\theta) &= \sum_{i=1}^{n}\log E\big[\frac{p(x_i,z_i=k|\theta)}{Q_i(z_i)}\big] \\ &\geq \sum_{i=1}^{n}E\big[\log\frac{p(x_i,z_i=k|\theta)}{Q_i(z_i)}\big] \\ &\geq \sum_{i=1}^{n}\sum_{k=1}^{K}Q_i(z_i)\log\frac{p(x_i,z_i=k|\theta)}{Q_i(z_i)} \end{aligned} \tag{0.1.4}$$

now the goal is to find a solution that the equation has "=" holds.

$$\log p(x|\theta) \geq \sum_{i=1}^{n}\sum_{k=1}^{K}Q_i(z_i)\log\frac{p(x_i,z_i=k|\theta)}{Q_i(z_i)}$$

Remember that we say the equal sign "=" holds if $f(x) = C$.
Define $\frac{p(x_i,z_i=k|\theta)}{Q_i(z_i)} = C$, we know the fact that the summation of the distribution of z_i must be 1, $\sum_{k=1}^{K} Q_i(z_i = k) = 1$, it is not difficult to get $\sum_{k=1}^{K} p(x_i, z_i = k|\theta) = C$.

now, let's see the compact solution of $Q_i(z_i)$

$$\begin{aligned} Q_i(z_i) &= \frac{p(x_i,z_i=k|\theta)}{C} \\ &= \frac{p(x_i,z_i=k|\theta)}{\sum\limits_{k=1}^{K}p(x_i,z_i=k|\theta)} \\ &= \frac{p(x_i|z_i=k,\theta)p(z_i|k,\theta)}{\sum\limits_{k=1}^{K}p(x_i,z_i=k|\theta)} \\ &= \frac{p(x_i|z_i=k,\theta)p(z_i|k,\theta)}{p(x_i|\theta)} \\ &= p(z_i|x_i,\theta) \end{aligned} \tag{0.1.5}$$

With this in mind, we now explain **EM algorithm**

- 1.Initial parameter θ
- 2. E step, compute $Q_i(z_i) := p(z_i|x_i,\theta)$
- 3. M step, $\theta := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{K} Q_i(z_i)\log\frac{p(x_i,z_i=k|\theta)}{Q_i(z_i)}$
- 4. Repeat step 2 and 3 until converged.

In practice, Gaussian mixture model is widely used, and parameters $\theta = \mu, \Sigma$, the prior distribution of k is denoted as $\pi_k = p(z_i = k|\mu_k, \Sigma_k)$

$$p(\vec{x}|\mu,\Sigma) = \prod_{i=1}^{m}\sum_{k=1}^{K}\pi_k\mathcal{N}(x_i|z_i=k,\mu_k,\Sigma_k) \tag{0.1.6}$$

For mathemetical convenience, we would like to compute $\operatorname{argmax}_\theta \log[p(\vec{x}|\theta)]$

$$\log[p(\vec{x}|\mu,\Sigma)] = \sum_{i=1}^{m}\log\big[\sum_{k=1}^{K}p(z_i=k|\mu_k,\Sigma_k)\mathcal{N}(x_i|z_i=k,\mu_k,\Sigma_k)\big] \tag{0.1.7}$$

also, we define $\gamma_{ik} = p(z_i = k|x_i, \mu_k, \Sigma_k)$, the posterior distribution that point i belongs cluster k, it is knowns as the reponsibility of cluster k for point i. According to Bayes' rule

$$\begin{aligned}\gamma_{ik} = p(z_i = k|x_i, \mu_k, \Sigma_k) &= \frac{p(z_i = k|\mu_k, \Sigma_k)p(x_i|z_i = k, \mu_k, \Sigma_k)}{p(x_i|\mu_k, \Sigma_k)} \\ &= \frac{p(z_i = k|\mu_k, \Sigma_k)p(x_i|z_i = k, \mu_k, \Sigma_k)}{\sum\limits_{k'=1}^{K} p(z_i = k'|\mu_k', \Sigma_k')p(x_i|z_i = k', \mu_k', \Sigma_k')}\end{aligned} \tag{0.1.8}$$

The EM algorithm for Gaussian Mixture Models work as follows:

1. Initialize the different parameters π, μ and Σ.
2. E step, Compute the responsibilities $\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|z_i=k,\mu_k,\Sigma_k)}{\sum\limits_{k'=1}^{K} \pi_k' \mathcal{N}(x_i|z_i=k',\mu_k',\Sigma_k')}$
3. M step, set partial derivative eq(0.1.7) wrt $\mu_k$andΣ_k.

$$\begin{aligned}\frac{\partial \log[p(\vec{x}|\mu, \Sigma)]}{\partial \mu_k} &= \sum_{i=1}^{m} \frac{\pi_k \mathcal{N}(x_i|z_i = k, \mu_k, \Sigma_k)}{\sum\limits_{k'=1}^{K} \pi_k' \mathcal{N}(x_i|z_i = k', \mu_k', \Sigma_k')} \Sigma_k^{-1}(x_i - \mu_k) \\ &= \sum_{i=1}^{m} \gamma_{ik} \Sigma_k^{-1}(x_i - \mu_k) \\ &= 0\end{aligned}$$

there, we obtain

$$\mu_k = \frac{\sum\limits_{i=1}^{m} \gamma_{ik} x_i}{\sum\limits_{i=1}^{m} \gamma_{ik}} = \frac{1}{m_k} \sum_{i=1}^{m} \gamma_{ik} x_i$$

Similarly,

$$\begin{aligned}\frac{\partial \log[p(\vec{x}|\mu, \Sigma)]}{\partial \Sigma_k} &= \sum_{i=1}^{m} \gamma_{ik} \Big[\exp[(x_i - \mu_k)\Sigma^{-1}(x_i - \mu_k)^T] + \Sigma_k \exp[(x_i - \mu_k)\Sigma^{-1}(x_i - \mu_k)^T](-\Sigma_k^{-2}) \Big] \\ &= \sum_{i=1}^{m} \gamma_{ik} \Big[1 - (x_i - \mu_k)(x_i - \mu_k)^T \Sigma^{-1} \Big] \exp[(x_i - \mu_k)\Sigma^{-1}(x_i - \mu_k)^T] \\ &= 0\end{aligned}$$

Since $\exp[(x_i - \mu_k)\Sigma^{-1}(x_i - \mu_k)^T] \neq 0$, the solution of μ_k is

$$\Sigma_k = \frac{1}{m_k} \sum_{i=1}^{m} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T \tag{0.1.9}$$

Next, we introduce Largange multiplier to solve π_k. the constrains on π_k is $\sum\limits_{k=1}^{K} \pi_k = 1$

$$\log p(x|\mu, \Sigma) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

maximizing the above wrt π_k

$$\frac{\partial \log[p(\vec{x}|\mu, \Sigma)]}{\partial \pi_k} = \sum_{i=1}^{m} \frac{\mathcal{N}(x_i|z_i = k, \mu_k, \Sigma_k)}{\sum\limits_{k=1}^{K} \pi_k \mathcal{N}(x_i|z_i = k, \mu_k, \Sigma_k)} + \lambda = 0$$

Multipling both part with π_k and summing over k, we have $\lambda = -N$, again multipling both sides with π_k yields

$$0 = \sum_{i=1}^{m} \gamma_{ik} - \pi_k m$$

we have

$$\pi_k = \frac{m_k}{m} \tag{0.1.10}$$

4. Repeat step 2 and 3 until convergence.