EM Algorithm

Contents

Basic Motivation

suppose we have two random variable x and z, which follows a joint distribution governed by parmameter θ. If you know the information of both the x_i and z_i for each individual i, then the likelihood function for x_i is

$$pr(x = x_i, z = z_i; \theta)$$

Then you can apply the MLE. Sometime, however, some variables may be lost, say ,z_i. We cannot observe the z_i. Then, how to deal with this? notice that we have

$$pr(x = x_i; \theta) = \sum_{z_i} pr(x = x_i, z = z_i'; \theta)$$

Given this, we can construct the MLE as

$$L(\theta) = \sum_{i=1}^{N} \log pr(x = x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z_i} pr(x = x_i, z = z_i'; \theta) \tag{1}$$

Here is a question: It is in general very hard to optimize this, since we have the 'log of sum'. If pr itself has some complicated form, then it is even harder to optimize. Therefore, we may want a more viable way to calculate.

Some Maths

Notice that we have

$$\sum_{i=1}^{N} \log \sum_{z_k} pr(x = x_i, z = z_k; \theta) = \sum_{i=1}^{N} \log \sum_{z_k} Q(z_k) \frac{pr(x = x_i, z = z_k; \theta)}{Q(z_k)}$$

$$\geq \sum_{i=1}^{N} Q(z_k) \sum_{z_k} \log \frac{pr(x = x_i, z = z_k; \theta)}{Q(z_k)} \equiv J(\theta)$$

where we have $\sum_{z_k} Q(z_k) = 1$. The $\geq$ holds due to the Jensen inequality. From the above reasoning we know that $J(\theta)$ is a lower bound of $L(\theta)$.

One important thing to notice: if taking $Q(z)$ as given, $J(\theta)$ is much easier to optimize , since there is only 'sum of log'.Then, how should we find the $Q(z)$?

To answer the question, we need to first make clear how to do the optimization.

1. Given θ, for $L(\theta)$, which is hard to optimize, we want to find out the $Q(z)$ that makes its lower bound, $J(\theta)$, equals $L(\theta)$.
    2. OK, now $L(\theta) = J(\theta)$. Therefore, Optimizing L is equivalent to optimizing J ,and optimizing J is easy. Suppose we have (take Q as given!)worked out the optimal θ that can maximize J. Denote as θ'.
        3. Now, given θ', we have $L(\theta')$. Again , we find out the $Q(z)$ that makes $J(\theta') = L(\theta')$. We then work out the optimal θ' that can maxmize $J(\theta')$ and denote it as $\theta''\ldots$
        We can show that actually this iteration procedure will converge to the optimal point θ^*.
        We have constructed the optimization strategy. The last thing to do is to find out Q that can make $J = L$ at each iteration.This is super-simple. We all know that $\log(\lambda * a + (1-\lambda) * b) \geq \lambda \log a + (1-\lambda) \log b$, and the equality holds only when $a = b$. Taking this back to our case, to make $L = J$ we need to have

$$\frac{pr(x = x_i, z = z_k; \theta)}{Q(z_k)} = \frac{pr(x = x_i, z = z'_k; \theta)}{Q(z'_k)}$$

hold for any k, k'. And since $\sum_{z_k} Q(z_k) = 1$, immediately we have

$$Q(z_k) = \frac{pr(x = x_i, z = z_k; \theta)}{\sum_{z_k} pr(x = x_i, z = z_k; \theta)} \tag{2}$$

This is exactly the conditional probability of z_k given we observe x_i and have parameter θ.

Algorithm Summary

Now we can finally write the the algorithm. First, guess an initial θ.

step 1: given the θ, calculate the $Q(z)$ according to (2)

step 2: maximize the $L(\theta)$ in (1) taking all $Q(z)$ as given. get θ'

step 3: if θ' is not close to θ. then let $\theta = \theta'$, and repeat from step 1.

https://blog.csdn.net/zouxy09/article/details/8537620