EM 算法

1 EM算法简单理解

EM算法是求解这个问题的一种迭代算法，它有3步：

E步

E步：计算在当前迭代的模型参数下，观测数据j来自硬币B的概率：

$$\mu^{(i+1)} = \frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j} + (1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}}$$

这个式子也是一目了然的，分子代表选定B并进行一次投掷试验，分母代表选定B或C并进行一次投掷试验，两个一除就得到试验结果来自B的概率。

M步

M步：估算下一个迭代的新的模型估算值：

$$\pi^{(i+1)} = \frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)}$$

$$p^{(i+1)} = \frac{\sum_{j=1}^{n}\mu_j^{(i+1)}y_j}{\sum_{j=1}^{n}\mu_j^{(i+1)}}$$

$$q^{(i+1)} = \frac{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})y_j}{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})}$$

这个也好说，把这n个[试验结果来自B的概率]求和得到期望，平均后，得到B出正面的似然估计，同理有p和q。

重复以上步骤直到收敛

2 EM算法的标准定义

输入：观测数据Y,隐变量，Z参数 θ，联合分布P(Y,Z| θ),条件分布P(Y| Z,θ)
输出： 模型参数θ

(1) 选择参数的初值$\theta^{(0)}$,开始迭代；

(2) E步：记$\theta^{(t)}$为第t次迭代参数θ的估计值，在第t+1次迭代的E步，计算

$$Q(\theta,\theta^{(t)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(t)}] = \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(t)})$$

这里，$P(Z|Y,\theta^{(t)})$是在给定观测数据Y和当前的参数估计$\theta^{(t)}$下隐变量数据Z的条件概率分布；

(3) M步：求使$Q(\theta,\theta^{(t)})$极大化的θ，确定第t+1次迭代的参数的估计值$\theta^{(t+1)}$

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta,\theta^{(t)})$$

(4) 重复第(2)步和第(3)步，直到收敛。

式 $Q(\theta,\theta^{(t)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(t)}] = \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(t)})$ 的函数$Q(\theta,\theta^{(t)})$是EM算法的核心，称为Q函数（Q function）。

说明

下面关于EM算法作几点说明：

步骤（1）参数的初值可以任意选择，但需注意EM算法对初值是敏感的。

步骤(2)E步求$Q(\theta,\theta^{(t)})$。Q函数式中Z是未观测数据，Y是观测数据。注意，$Q(\theta,\theta^{(t)})$的第1个变元表示要极大化的参数，第2个变元表示参数的当前估计值。每次迭代实际在求Q函数及其极大。

步骤(3)M步求$Q(\theta,\theta^{(t)})$的极大化，得到$\theta^{(t+1)}$，完成一次迭代$\theta^{(t)} \to \theta^{(t+1)}$，后面将证明每次迭代使似然函数增大或达到其极值。

步骤(4)给出停止迭代的条件，一般是对较小的正数ε_1、ε_2，若满足

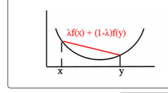
$$\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon_1 \quad 或 \quad \|Q(\theta^{(t+1)},\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)})\| < \varepsilon_2$$

则停止迭代。

3 EM 算法的推导

Jensen’s Inequality

Figure 1: A Convex Function



Jensen’s inequality is a general result in convexity

$$\lambda f(x) + (1-\lambda)f(y) \geq f(\lambda x + (1-\lambda)y$$

We can also generalize the result to expectation: $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

The EM Objective Funciton

Objection funciton

$$\hat{\theta} = \operatorname*{argmax}_\theta \sum_z p(x|\theta)p(x|z,\theta)$$

$$\hat{\theta} = \operatorname*{argmax}_\theta \log p(x,z|\theta) = \operatorname*{argmax}_\theta \log p(z|\theta) + \log p(x|z,\theta).$$

Let D , (x1, ⋯ , xN) be the observed data, and let Z , hidden random variables

where q(z) represents an arbitrary distribution for the random variable Z.

Z , hidden random variables, x , output variables

$$\log p(x|\theta) = \log \sum_z p(z|\theta)p(x|z,\theta)\frac{q(z)}{q(z)}$$
$$= \log \mathbb{E}_q\left[\frac{p(z|\theta)p(x|z,\theta)}{q(z)}\right]$$
$$\geq \mathbb{E}_q\left[\log \frac{p(z|\theta)p(x|z,\theta)}{q(z)}\right]$$
$$\mathcal{L}(\theta;q) = \mathbb{E}_q[\log p(z|\theta)] + \mathbb{E}_q[\log p(x|z,\theta)] - \mathbb{E}_q[\log q(z)].$$

At the M-Step, we update the model parameterθ

$$\theta^{(t+1)} = \operatorname*{argmax}_\theta \mathcal{L}(q^{(t+1)},\theta)$$

At the E-Step, we update the posterior value q

$$q^{(t+1)} = \operatorname*{argmax}_q \mathcal{L}(q,\theta^{(t)}) = p(z|x,\theta^{(t)}).$$

as we maximize the objective function, we are also maximizing the log likelihood function.

$$\mathcal{L}(p(z|x,\theta),\theta) = \sum_Z p(z|x,\theta)\log\frac{p(x,z|\theta)}{p(z|x,\theta)}$$
$$= \sum_Z p(z|x,\theta)\log\frac{p(x,z|\theta)p(x|\theta)}{p(x,z|\theta)}$$
$$= \sum_Z p(z|x,\theta)\log p(x|\theta)$$
$$= \log p(x|\theta)\sum_Z p(z|x,\theta)$$
$$= \log p(x|\theta)$$

E-Step

repeat

The E step has the following simple form, which is the same for any mixture model:

$$Q(\theta,\theta^{(t)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(t)}] = \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(t)})$$

M-Step

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta,\theta^{(t)})$$

the bound of EM

It is easy to see that

$$Q(\Theta^{(t)},\Theta^{(t)}) = \sum_{i=1}^{n}\log p(x_i|\Theta^{(t)}) = \ell(\Theta^{(t)}). \quad (18)$$

Since $\Theta^{(t+1)}$ maximizes Q, we have

$$Q(\Theta^{(t+1)},\Theta^{(t)}) \geq Q(\Theta^{(t)},\Theta^{(t)}) = \ell(\Theta^{(t)}). \quad (19)$$

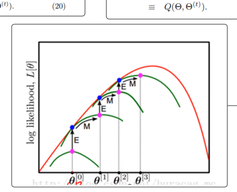
On the other hand, Q lower bounds ℓ. Therefore

$$\ell(\Theta^{(t+1)}) \geq Q(\Theta^{(t+1)},\Theta^{(t)}) \geq Q(\Theta^{(t)},\Theta^{(t)}) = \ell(\Theta^{(t)}). \quad (20)$$

$$\begin{aligned}\ell(\Theta) &= \sum_{i=1}^{n}\log\sum_{y=1}^{K}p(x_i,y|\Theta)\\ &= \sum_{i=1}^{n}\log\sum_{y=1}^{K}p(y|x_i,\Theta^{(t)})\frac{p(x_i,y|\Theta)}{p(y|x_i,\Theta^{(t)})}\\ &\geq \sum_{i=1}^{n}\sum_{y=1}^{K}p(y|x_i,\Theta^{(t)})\log\frac{p(x_i,y|\Theta)}{p(y|x_i,\Theta^{(t)})}\\ &\equiv Q(\Theta,\Theta^{(t)}).\end{aligned}$$

The lower bound is obtained via Jensen’s inequality

$$\log\sum_i p_i f_i \geq \sum_i p_i \log f_i,$$



This shows that Θ(t+1) is indeed a better (or no worse) parameter than Θ(t) in terms of the marginal log likelihood function. By iterating,
we arrive at a local maximum of function

4 EM 算法的推广

EM算法高斯混合模型

GMM使用若干个高斯分布的加权和作为对观测数据集进行建模的基础分布，而由中心极限定理我们知道，大量独立同分布的随机变量的均值在做适当标准化之后会依分布收敛于高斯分布，这使得高斯分布具有普适性的建模能力，继而奠定了使用高斯分布作为主要构成部件的GMM进行数据建模的理论基础.

EM算法通过迭代地构造似然函数下限的方式不断地提升似然函数的取值，从而完成对含有隐变量模型的参数估计，其典型的应用包括GMM、HMM（Hidden Markov Model，隐马尔可夫模型） 的参数估计

建模

GMM模型使用 K个高斯分布的加权和作为其概率密度函数，具体地

$$p(x;\Theta) = \sum_{k=1}^{K}\alpha_k \cdot \mathcal{N}(x;\vec{u}_k,\Sigma_k)$$

其中参数表示为：

$$\Theta = \{\alpha_k,\vec{u}_k,\Sigma_k\}_{k=1}^{K},$$

为保证概率密度为1，应该满足

$$\int_{-\infty}^{+\infty}p(x;\Theta)dx = \sum_{k=1}^{K}\left[\alpha_k\int_{-\infty}^{+\infty}\mathcal{N}(x;\vec{u}_k,\Sigma_k)dx\right] = \sum_{k=1}^{K}\alpha_k = 1 \quad (2)$$

$\alpha_k, k = 1,\cdots,K$ 即为隐变量 Z 所服从的分布 $\mathcal{Q}$ 的参数值, 即
$\alpha_k = \mathcal{Q}(Z=k), k \in \{1,\cdots,K\}$ 通过积可以推出 GMM模型对不定全数据分布的归纳层
通过求对边缘分布 $p(x;\Theta) = \sum_z p(x,z;\Theta) = \sum_z p(x|z;\Theta)p(x|z;\Theta)$ 得到的.

样本 Xi生成的概率可表示为

$$\begin{aligned}p(x_i;\Theta) &= \sum_{z_i}p(x_i,z_i;\Theta)\\ &= \sum_{k=1}^{K}\mathcal{N}(x_i|Z=k;\vec{u}_k,\Sigma_k)\mathcal{Q}(Z=k)\\ &= \sum_{k=1}^{K}\alpha_k\cdot\mathcal{N}(x_i|Z=k;\vec{u}_k,\Sigma_k)\end{aligned}$$

但是在实际问题中，常常存在隐变量，上式不能完全使用，变换后为：

$$\begin{aligned}\mathcal{NLL}(\Theta) &= -\log\prod_{i=1}^{n}p(x_i;\Theta)\\ &= -\sum_{i=1}^{n}\log p(x_i;\Theta)\\ &= -\sum_{i=1}^{n}\log\left[\sum_{z_i}p(x_i,z_i;\Theta)\right]\\ &= -\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}\alpha_k\cdot p(x_i|z_i=k;\Theta)\right]\end{aligned} \qquad \begin{aligned}&\min \quad \mathcal{NLL}(\Theta)\\ &s.t. \quad \sum_{k=1}^{K}\alpha_k = 1\end{aligned}$$

需要注意的是，由于对数函数的凹凸性正好相反，所以此处Jensen’s inequality中的符号应该取反.，我们借助在第 t-1次迭代中得到的参数估计值 θ 来获得 xi 关于zi 的后验分布

$$p(z_i|x_i;\Theta_{t-1}) = \frac{p(x_i|z_i;\Theta_{t-1})\cdot p(z_i;\Theta_{t-1})}{\sum_{z_i}p(x_i|z_i;\Theta_{t-1})\cdot p(z_i;\Theta_{t-1})}$$

收敛性证明

受Jensen’s inequality的启发，如果我们能在对数函数的连加操作里面构造出一个关于 Z 的期望，那我们就能将连加操作移至对数函数的外面了，具体地，对于对数似然函数我们有

$$\begin{aligned}\mathcal{LL}(\Theta) &= \sum_{i=1}^{n}\log\left[\sum_{z_i}p(x_i,z_i;\Theta)\right]\\ &= \sum_{i=1}^{n}\log\left[\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right]\\ &\geq \sum_{i=1}^{n}\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\log\left[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right]\end{aligned}$$

需要注意的是，由于对数函数的凹凸性正好相反，所以此处Jensen’s inequality中的符号应该取反.，我们借助在第 t-1次迭代中得到的参数估计值 θ 来获得 xi 关于zi 的后验分布

$$p(z_i|x_i;\Theta_{t-1}) = \frac{p(x_i|z_i;\Theta_{t-1})\cdot p(z_i;\Theta_{t-1})}{\sum_{z_i}p(x_i|z_i;\Theta_{t-1})\cdot p(z_i;\Theta_{t-1})}$$

再使用这个关于 zi 的后验分布构造期望

$$\mathbf{E}_{z_i}\left[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right] = \sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}$$

而后由Jensen’s inequality即可得到

$$\log\mathbf{E}_{z_i}\left[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right] \geq \mathbf{E}_{z_i}\left[\log\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right]$$
$$\Downarrow$$
$$\log\left[\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right] \geq \sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\log\left[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right] \quad (16)$$

我们记下届函数为：

$$\begin{aligned}B(\Theta,\Theta_{t-1}) &= \sum_{i=1}^{n}\mathbf{E}_{z_i}\left[\log\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right]\\ &= \sum_{i=1}^{n}\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\log\left[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\right]\end{aligned}$$

即对数似然函数的下界函数有

$$\mathcal{LL}(\Theta) \geq B(\Theta,\Theta_{t-1})$$

对于(15), 当 $\Theta = \Theta_{t-1}$ 时有
$\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})} = \frac{p(z_i|x_i;\Theta_{t-1})\cdot p(x_i;\Theta_{t-1})}{p(z_i|x_i;\Theta_{t-1})} = p(x_i|\Theta_{t-1})$ 于是 x_i 取值数值为常数, 此时Jensen's inequality等号成立, 即有

$$\sum_{i=1}^{n}\log\left[\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\frac{p(x_i,z_i;\Theta_{t-1})}{p(z_i|x_i;\Theta_{t-1})}\right] = \sum_{i=1}^{n}\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\log\left[\frac{p(x_i,z_i;\Theta_{t-1})}{p(z_i|x_i;\Theta_{t-1})}\right]$$
$$\Downarrow$$
$$B(\Theta_{t-1},\Theta_{t-1}) = \mathcal{LL}(\Theta_{t-1}) \quad (16)$$

而对于 $\Theta = \Theta_t$ 由(16)有

$$\begin{aligned}\mathcal{LL}(\Theta_t) &= \sum_{i=1}^{n}\log\left[\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\frac{p(x_i,z_i;\Theta_t)}{p(z_i|x_i;\Theta_{t-1})}\right]\\ &\geq \sum_{i=1}^{n}\sum_{z_i}p(z_i|x_i;\Theta_{t-1})\cdot\log\left[\frac{p(x_i,z_i;\Theta_t)}{p(z_i|x_i;\Theta_{t-1})}\right]\\ &= B(\Theta_t,\Theta_{t-1}) \quad (17)\end{aligned}$$

由(16)(17)可得

$$\mathcal{LL}(\Theta_t) \geq B(\Theta_t,\Theta_{t-1}) \geq B(\Theta_{t-1},\Theta_{t-1}) = \mathcal{LL}(\Theta_{t-1}) \quad (18)$$

算法

输入：观测数据 $y_1,y_2,\cdots,y_N$，高斯混合模型；
输出：高斯混合模型参数；
（1）取参数的初始值开始迭代
（2）E 步：根据当前模型参数，计算分模型 k 对观测数据 y_j 的响应度

$$\hat{\gamma}_{jk} = \frac{\alpha_k\phi(y_j|\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j|\theta_k)}, \quad j=1,2,\cdots,N;\ k=1,2,\cdots,K$$

（3）M 步：计算新一轮迭代的模型参数

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}y_j}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}, \quad k=1,2,\cdots,K$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}(y_j-\hat{\mu}_k)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}, \quad k=1,2,\cdots,K$$

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}}{N}, \quad k=1,2,\cdots,K$$

（4）重复第（2）步和第（3）步，直到收敛.

GEM算法

1

输入：观测数据，F 函数；
输出：模型参数.
（1）初始化参数 $\theta^{(0)}$，开始迭代
（2）第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，$F^{(i)}$ 为函数 F 的估计，求 $F^{(i)}$ 极子极大化 $F(F,\theta^{(i)})$
（3）第 2 步，求 $\theta^{(i+1)}$ 使 $F(F^{(i)})$ 的极大化
（4）重复 (2) 和 (3)，直到收敛。 ■
在 GEM 算法 1 中，有时求 $Q(\theta,\theta^{(i)})$ 的极大化是很困难的，下面介绍的 GEM 算法 2 和 GEM 算法 3 并不是直接求 $\theta^{(i+1)}$ 使 $Q(\theta,\theta^{(i)})$ 达到极大的 θ，而是找一个 $\theta^{(i+1)}$ 使得 $Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$。

2

输入：观测数据，Q 函数；
输出：模型参数.
（1）初始化参数 $\theta^{(0)}$，开始迭代
（2）第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，计算

$$Q(\theta,\theta^{(i)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}] = \sum_Z P(Z|Y,\theta^{(i)})\log P(Y,Z|\theta)$$

（3）第 2 步，求 $\theta^{(i+1)}$ 使

$$Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$$

（4）重复 (2) 和 (3)，直到收敛。 ■
当参数 θ 的维数为 d（$d \geq 2$）时，可采用一种特殊的 GEM 算法，它将 EM 算法的 M 步分解为 d 次条件极大化，每次只改变参数向量的一个分量，其余分量不改变。

3

输入：观测数据，Q 函数；
输出：模型参数.
（1）初始化参数 $\theta^{(0)} = (\theta_1^{(0)},\theta_2^{(0)},\cdots,\theta_d^{(0)})$，开始迭代
（2）第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)} = (\theta_1^{(i)},\theta_2^{(i)},\cdots,\theta_d^{(i)})$ 为参数 $\theta = (\theta_1,\theta_2,\cdots,\theta_d)$ 的估计值，计算

$$Q(\theta,\theta^{(i)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}] = \sum_Z P(Z|Y,\theta^{(i)})\log P(Y,Z|\theta)$$

（3）第 2 步，进行 d 次条件极大化：
首先，在 $\theta_2^{(i)},\cdots,\theta_d^{(i)}$ 保持不变的条件下求使 $Q(\theta,\theta^{(i)})$ 达到极大的 $\theta_1^{(i+1)}$，
然后，在 $\theta_1 = \theta_1^{(i+1)}$，$\theta_j = \theta_j^{(i)}$，$j = 3,4,\cdots,k$ 的条件下求使 $Q(\theta,\theta^{(i)})$ 达到极大的 $\theta_2^{(i+1)}$；
如此继续，经过 d 次条件极大化，得到 $\theta^{(i+1)} = (\theta_1^{(i+1)},\theta_2^{(i+1)},\cdots,\theta_d^{(i+1)})$ 使得

$$Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$$

（4）重复 (2) 和 (3)，直到收敛。 ■