EM(Expection Maximization)

隐变量

极大似然估计(MLE)

> 极大似然估计适用于参数模型已知的情况，例如，假设男性的身高服从参数为$\Theta = [\ \mu, \sigma^2\ ]$的高斯分布，概率密度为$f(x_j) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_j-\mu)^2}{\sigma^2}}$现在给定一些样本数据$T = (x_1, x_2, \ldots, x_N)$来估计参数$\Theta$。
>
> 首先，假定样本之间独立同分布($i.i.d$)，那么样本x_j发生的概率为$p(x_j) = f(x_j)$，所有样本发生的概率为$P = \prod_{j=1}^N p(x_j) = \prod_{j=1}^N f(x_j)$；为什么偏偏这些样本要发生呢？我们有理由相信他之所以发生是因为他发生的概率最大，这个听起来还是很合理的，那么接下来就是对P求其极大值所对应的参数Θ；对于联乘形式的求导一般化为对对数求导，此时极大值所对应的参数是一样的；即问题转化为：
>
> $$arg\ max_\Theta \sum_{j=1}^N logf(x_j) \tag{1}$$
>
> 接下来对上述多元函数求偏导数并且令其等于0即可。

但是假设这样一种情形：假设男性的身高服从两个参数为$\Theta_1 = [\ \mu_1, \sigma_1^2\ ]$，$\Theta_2 = [\ \mu_2, \sigma_2^2\ ]$的高斯分布，现在给定一些样本数据$T = (x_1, x_2, \ldots, x_N)$但是并没有给出所属的类别，来估计$\Theta_1, \Theta_2$，那么还可以使用极大似然估计吗？显然是不可以的，这时候我们可以引入隐变量帮助我们分析。

隐变量

> 假设给定的K个模型，用随机变量Z表示选择的模型，z表示Z的取值。
>
> 现在假定选择模型也是一个随机事件，即Z服从一个Q分布，那么就有$\sum_{k=1}^K p(Z = z_k) = 1$。
>
> 对于上述假设的情形，我们知道根据2个模型选出的样本集T，但是我们不知道样本集中的样本x_j选自哪个模型，所以不能使用极大似然估计；现在引入了隐变量Z，就可以将选取模型+选取模型中的样本用概率表示出来，然后就可以使用极大似然估计了。

条件概率

条件概率

> 回顾概率论中的条件概率：
>
> $$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{2}$$
>
> 结合贝叶斯公式可以得到：
>
> $$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_X P(Y|X)P(X)} \tag{3}$$
>
> 现在将两个变量扩展到三个变量：

$$
\begin{aligned}
P(x,z|\theta) &= \frac{p(x,z,\theta)}{p(\theta)} \\
p(x|z,\theta) &= \frac{p(x,z,\theta)}{p(z,\theta)} \\
&= \frac{p(x,z,\theta)}{p(\theta)}\frac{p(\theta)}{p(z,\theta)} \\
&= \frac{p(x,z|\theta)}{p(z|\theta)}
\end{aligned}
$$

边缘概率和联合概率

$$
\begin{aligned}
p(x,\theta) &= \sum_z p(x,z,\theta) \\
p(x|\theta) &= \sum_z p(x,z|\theta)
\end{aligned} \tag{4}
$$

EM算法

EM算法用来估计上述含隐变量问题：

假定事先知道K个模型，以及样本集$T=(x_1,x_2,\ldots,x_N)$，现在要估计模型参数θ：

我们知道对于T中的某一个样本x_j，有

$$
p(x_j|\theta) = \sum_z p(x_j,z|\theta) \tag{5}
$$

由于N个样本独立同分布，有：

$$
\begin{aligned}
P(T|\theta) &= \prod_{j=1}^{N} p(x_j,\theta) \\
&= \prod_{j=1}^{N}\sum_z p(x_j,z|\theta)
\end{aligned}
$$

上述式子的意义在于，在参数模型θ给定的条件下，样本集发生的概率为$P(T|\theta)$。

对数似然：

$$
L(T|\theta) = lnP(T|\theta) = \sum_{j=1}^{N} ln\sum_z p(x_j,z|\theta) \tag{6}
$$

我们的目的是极大化$L(T|\theta)$，但是将其求导数会是很复杂的式子不好直接得出解析解，所以将上述式子继续变形

应用隐变量z服从Q分布的条件等价得到：

$$
\begin{aligned}
L(T|\theta) &= \sum_{j=1}^{N} ln\sum_z p(x_j,z|\theta) \\
&= \sum_{j=1}^{N} ln\sum_z Q(z|\theta)\frac{p(x_j,z|\theta)}{Q(z|\theta)}
\end{aligned}
$$

接下来利用ln函数的凹凸性应用$Jensen$不等式得到：

$$
\begin{aligned}
L(T|\theta) &= \sum_{j=1}^{N} ln\sum_z Q(z|\theta)\frac{p(x_j,z|\theta)}{Q(z|\theta)} \\
&\geq \sum_{j=1}^{N}\sum_z Q(z|\theta)ln\frac{p(x_j,z|\theta)}{Q(z|\theta)}
\end{aligned}
$$

所以可以极大化$L(T|\theta)$的上述下界来达到极大化$L(T|\theta)$的目的。

分布$Q(z|\theta)$是我们假设出来的，我们仅仅知道$\sum_z Q(z|\theta)=1$，但是可以根据$Jensen$不等式成立的条件得到另一个结论：

$Jensen$不等式成立的条件为：

$$\begin{aligned}\frac{p(x_j,z|\theta)}{Q(z|\theta)}&=c\\ p(x_j,z|\theta)&=cQ(z|\theta)\\ \sum_z p(x_j,z|\theta)&=c\sum_z Q(z|\theta)\\ c&=\sum_z p(x_j,z|\theta)=p(x_j|\theta)\end{aligned}$$

由此得到：

$$Q(z|\theta)=\frac{p(x_j,z|\theta)}{p(x_j|\theta)}=p(z|x_j,\theta) \tag{7}$$

所以所求极大化问题转化为：

$$arg\ max_\theta \sum_{j=1}^N\sum_z p(z|x_j,\theta)lnp(x_j,z|\theta) \tag{8}$$

EM算法流程：

输入：观测数据集$T=(x_1,x_2,...,x_N)$，联合分布$p(x,z|\theta)$，条件分布$p(z|x,\theta)$，迭代最大次数M

1. 随机初始化参数θ为θ^0
2. 在最大迭代次数范围内($i\ \ from\ \ 1\ \ to\ \ M$)：
   - $E—step$：

$$Q(z|\theta_i)=p(z|x_j,\theta_{i-1}) \tag{9}$$

- $M—step$：

$$arg\ max_{\theta_i}\sum_{j=1}^N\sum_z p(z|x_j,\theta_{i-1})lnp(x_j,z|\theta_{i-1}) \tag{10}$$

   - 重复上述$E—step$和$M—step$直至收敛

输出：参数θ。

高斯混合模型(Gaussian Mixture Model,GMM)

高斯混合模型是指具有如下形式的概率分布模型：

$$P(y|\theta)=\sum_{k=1}^K a_k\phi(y|\theta_k) \tag{11}$$

其中$\sum_{k=1}^K a_k=1$，$\phi(y|\theta_k)$是高斯分布密度，$\theta_k=(\mu_k,\Sigma_k)$，

$$\phi(y|\theta_k)=\frac{1}{(2\pi)^{\frac{d}{2}}}\frac{1}{|\Sigma_k|^{\frac{1}{2}}}exp[-\frac{1}{2}(\bar{x}-\bar{\mu}_k)^T\Sigma_k^{-1}(\bar{x}-\bar{\mu}_k)] \tag{12}$$

d为维度。

> Q :**为什么**$\sum_{k=1}^K a_k=1$**?**
>
> **高斯混合模型为概率模型，须保证规范性，所以**$\sum_{k=1}^K a_k=1$。
>
> Q :a_k**的意义?**
>
> **高斯混合模型为概率模型，**a_k**可以认为选取第**k**个模型的概率，即**$a_k=p(z_k|\theta_k)$或者$a_k=p(z_k|\Theta)$。

现在假设样本集$T=(x_1,x_2,\dots,x_N)$取自GMM模型，现用EM算法来估计参数θ_k：

按照上述EM算法流程，首先需要找到Q函数：

$$Q(z|\Theta)=p(z|x_j,\Theta) \tag{13}$$

其中$\Theta=[\theta_1,\theta_2,\dots,\theta_K]$

根据贝叶斯公式以及全概率公式可以得到：

$$\begin{aligned} Q(z_k|\Theta) &= \frac{p(z_k|\Theta)p(x_j|z_k,\Theta)}{\sum_{i=1}^K p(z_i|\Theta)p(x_j|z_i,\Theta)} \\ &= \frac{a_k\phi(x_j|\theta_k)}{\sum_{i=1}^K a_i\phi(x_j|\theta_i)} \end{aligned}$$

那么对数似然函数就为：

$$\begin{aligned} L &= \sum_{j=1}^N\sum_z p(z|x_j,\Theta)lnp(x_j,z|\Theta) \\ L &= \sum_{j=1}^N\sum_z \frac{a_k\phi(x_j|\theta_k)}{\sum_{i=1}^K a_i\phi(x_j|\theta_i)}lnp(x_j|z,\Theta)p(z|\Theta) \\ L &= \sum_{j=1}^N\sum_{k=1}^K \frac{a_k\phi(x_j|\theta_k)}{\sum_{i=1}^K a_i\phi(x_j|\theta_i)}ln\phi(x_j|\theta_k)a_k \\ L(a,\mu,\Sigma) &= \sum_{j=1}^N\sum_{k=1}^K \lambda_{jk}[lna_k-\frac{1}{2}ln|\Sigma_k|-\frac{1}{2}(x_j-\bar{\mu}_k)^T\Sigma_k^{-1}(x_j-\bar{\mu}_k)] \end{aligned}$$

上述式子令$\frac{a_k\phi(x_j|\theta_k)}{\sum_{i=1}^K a_i\phi(x_j|\theta_i)}=\lambda_{jk}$，它是一个常数概率。在化简过程中还省略了一些常数。

矩阵求导公式

此题中用到的矩阵求导公式：

$$\begin{aligned} \frac{\partial x^TAx}{\partial x} &= (A+A^T)x \\ \frac{\partial x^T\Sigma^{-1}y}{\partial\Sigma} &= -(\Sigma^{-1})^Txy^T(\Sigma^{-1})^T \\ \frac{\partial|\Sigma|}{\partial\Sigma} &= |\Sigma|(\Sigma^{-1})^T \end{aligned} \tag{14}$$

极大似然

$$\frac{\partial L}{\partial\mu_k}=\sum_{j=1}^N\lambda_{jk}\frac{1}{2}(\Sigma_k^{-1}+(\Sigma_k^{-1})^T)(x_j-\mu_k)=0 \tag{15}$$

因为Σ_k是对称矩阵，所以上市左乘Σ_k得到：

$$\mu_k=\frac{\sum_{j=1}^N\lambda_{jk}x_j}{\sum_{j=1}^N\lambda_{jk}} \tag{16}$$

对Σ_k求偏导数：

$$\frac{\partial L}{\partial\Sigma_k}=\sum_{j=1}^N\lambda_{jk}[-\frac{1}{2}\frac{|\Sigma_k|(\Sigma^{-1})^T}{|\Sigma_k|}+\frac{1}{2}(\Sigma^{-1})^T(x_j-\bar{\mu}_k)(x_j-\bar{\mu}_k)^T(\Sigma^{-1})^T]=0 \tag{17}$$

现左乘Σ_k，再右乘Σ_k得到：

$$\sum_{j=1}^N\lambda_{jk}[-\Sigma_k+(x_j-\bar{\mu}_k)(x_j-\bar{\mu}_k)^T]=0 \tag{18}$$

所以：

$$\Sigma_k=\frac{\sum_{j=1}^N\lambda_{jk}(x_j-\bar{\mu}_k)(x_j-\bar{\mu}_k)^T}{\sum_{j=1}^N\lambda_{jk}} \tag{19}$$

对于a_k，还有限制条件$\sum_{k=1}^{K} a_k = 1$：

对此构造新的约束函数为：

$$H = L + \eta(\sum_{k=1}^{N} a_k - 1) \tag{20}$$

求偏导数得到：

$$\begin{aligned} \frac{\partial L}{\partial a_k} &= \sum_{j=1}^{N} \lambda_{jk} \frac{1}{a_k} + \eta && = 0 \\ \frac{\partial L}{\partial \eta} &= \sum_{k=1}^{K} a_k - 1 && = 0 \end{aligned}$$

进而得到

$$a_k = \frac{1}{N} \sum_{j=1}^{N} \lambda_{jk} \tag{21}$$

GMM模型EM算法流程

输入：观测数据集$T = (x_1, x_2, \ldots, x_N)$，高斯模型个数$K$，最大迭代次数

1. 初始化高斯混合分布的参数$(a_k, u_k, \Sigma_k), k = 1, 2, \ldots, K$
2. 在最大迭代次数范围内($i \ \ from \ \ 1 \ \ to \ \ M$)：
   - $E — step$：

     计算λ_{jk}，$j = 1, 2, \ldots, N$
   - $M — step$：

     更新参数：

$$\begin{aligned} \mu_k &= \frac{\sum_{j=1}^{N} \lambda_{jk} x_j}{\sum_{j=1}^{N} \lambda_{jk}} \\ \Sigma_k &= \frac{\sum_{j=1}^{N} \lambda_{jk} (x_j - \bar{\mu}_k)(x_j - \bar{\mu}_k)^T}{\sum_{j=1}^{N} \lambda_{jk}} \\ a_k &= \frac{1}{N} \sum_{j=1}^{N} \lambda_{jk} \end{aligned}$$

   - 重复上述$E — step$和$M — step$直至收敛

输出：参数