EM算法(Expectation-Maximum)

原理篇

- **本质**：**概率模型**，去估计一个密度函数，**最大化对数似然函数**去估计参数。**无法直接用极大化对数似然函数**得到模型分布的参数，用**启发式跌代法**，**用于求解含有隐变量**的最大似然估计、最大后延概率估计问题。

  > 1. 隐变量：对概率模型有一定的影响，但无法观测；
  >
  >   观测数据X+隐变量 = 完全数据
  >
  > 2. 无法用极大似然的两种情况：数据缺失、含有未知隐变量

- **作用**：经常存在数据缺失或者不可用的的问题，这时候直接处理数据比较困难，而**数据添加**办法有很多种，常用的有神经网络拟合、添补法、卡尔曼滤波法等等，但是EM算法之所以能迅速普及主要源于它**算法简单，稳定上升**的步骤能非常可靠地找到“最优的收敛值”。
- **目标**：是使包含隐变量的数据集的后验概率或似然函数最大化， 进而得到最优的参数估计
- **思想**：我们可以发现我们的算法里已知的是**观察数据**，未知的是隐含数据和模型参数，**在E步**，我们所做的事情是固定模型参数的值，优化隐含数据的分布，而**在M步**，我们所做的事情是固定隐含数据分布，优化模型参数的值。

主要步骤：

已知：概率分布、随机抽取的样本；

未知：分类、模型参数

- E-step：**猜想隐含数据**，更新隐含数据和模型参数
- M-step：**基于观察数据和猜测的隐含数据求极大对数似然**，求解模型K
- E-step：基于前面得到的模型K，继续猜测隐含数据，继续极大化对数似然
- M-step：求模型参数
- 直至模型分布无明显变化，算法收敛

  > EM算法步骤如下：
  >
  > 输入：观测变量数据X，隐变量Z，联合分布p(X,Z|Θ)
  >
  > 输出：模型参数Θ
  >
  > (1) 选择初始参数Θ_0
  >
  > (2)E步：记Θi为第i次迭代参数Θ的估计值，在第i+1次迭代的E步，计算$Q(Θ,Θg)$;
  >
  > (3)M步：确定第i+1次迭代的参数的估计值Θ(i+1),即为：
  >
  > $$\Theta^{(i+1)} = argmax_{\Theta} Q(\Theta, \Theta^{(i)})$$
  >
  > (4)重复(2)和(3)步，直至收敛

EM算法的引入篇

有三枚硬币A、B、C

三枚硬币的模型：

$$p(y|\theta)=\sum_z P(y,z|\theta)=\sum_z p(z|\theta)p(y|z,\theta)$$
$$=\pi p^y(1-p)^{1-y}+(1-\pi)q^y(1-q)^{1-y}$$

N枚硬币的模型：

$$P(Y|\Theta)=\sum_z P(Z|\Theta)P(Y|Z,\Theta)$$
$$=\prod_{j=1}^{n}[\pi p^{y_j}](1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{q-y_j}$$

求模型的参数 Θ= (π,p,q)的极大似然估计，**即目标函数**为：

$$\Theta=argmaxlogp(Y|\Theta)=argmaxlog\sum_z P(Y|Z,\Theta)p(Z,\Theta)$$

面对含隐变量的概率模型，目标是**极大化观测数据（不完全数据）Y关于参数Θ的对数似然函数**，即极大化(取对数方便计算)：

$$L(\Theta)=logP(Y|\Theta)=log\sum_z P(Y,Z|\Theta)$$
$$=log(\sum_z P(Y|Z,\Theta)P(Z|\Theta))$$

EM算法通过迭代逐步近似极大化L(θ)的 + Jensen不等式

$$L(\Theta)-L(\Theta^i)=log(\sum_Z P(Y|Z,\Theta)P(Z|\Theta))-logP(Y|\Theta^{(i)})$$
$$=log[\sum_Z P(Z|Y,\Theta^{(i)})\frac{P(Y|Z)P(Z|\Theta)}{P(Z|Y,\Theta^i)}]-logP(Y|\Theta^{(i)})$$
$$\text{由}Jensen\text{不等式且}\sum_z P(Z|Y,\Theta^{(i)})=1$$
$$\geq\sum_Z P(Z|Y,\Theta^{(i)})log\frac{P(Y|Z,\Theta)P(Z|\Theta)}{P(Z|Y,\Theta^{(i)})}-\frac{logP(Y|\Theta^{(i)})*\sum_Z P(Z|Y,\Theta^{(i)})}{\sum_Z P(Z|Y,\Theta^{(i)})}$$
$$=\sum_Z P(Z|Y,\Theta^{(i)})*log\frac{P(Y|Z,\Theta)P(Z|\Theta)}{P(Z|Y,\Theta^{(i)})P(Y|\Theta^{(i)})}$$

即：

$$\Theta^{(i+1)}=argmax[L(\Theta^{(i)})+\sum_Z P(Z|Y,\Theta^{(i)})*log\frac{P(Y|Z,\Theta)P(Z|\Theta)}{P(Z|Y,\Theta^{(i)})P(Y|\Theta^{(i)})}]$$

去掉与Θ无关的变量的式子：

$$\Theta^{(i+1)}=argmax[\sum_Z P(Z|Y,\Theta^{(i)})*logP(Y|Z,\Theta)P(Z|\Theta)]$$
$$=argmaxQ(\Theta,\Theta^{(i)})=E[logP(Z|Y,\Theta^{(i)})]....(1)$$

> 观测数据 + 隐变量 = 完全数据
>
> Jensen不等式：
>
> $$log(\sum\alpha_i\varphi(x_i))>=\sum\alpha_i log\varphi(x_i)$$
> $$\sum\alpha_i=1\text{且}\alpha_i>=0$$

EM算法的收敛性

证明：P(Y|θ)为观测数据的似然函数，且是递增的，即：

$$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)})$$

证明如下：

$$P(Y|\theta) = \frac{P(Y,Z|\theta)}{P(Z|Y,\theta)}......有贝叶斯公式展开得$$
$$logP(Y|\theta) = logP(Y.Z|\theta) - logP(Z|Y,\theta)....取对数$$

$$Q(\theta,\theta^{(i)}) = \sum_z logP(Y,Z|\theta)P(Z|Y,\theta^{(i)})...由（1）得$$

构造下式（因为取对数方便相减和相除，同时构造了贝叶斯公式）：

$$H(\theta,\theta^{(i)}) = \sum_z logP(Z|Y,\theta)P(Z|Y,\theta^{(i)})$$
$$logP(Y|\theta) = Q(\theta,\theta^{(i)}) - H(\theta,\theta^{(i)}).....(2)$$

在式（2）中分别取θ为 θi 和θ i+1，**并相减**：

$$logP(Y|\theta^{(i+1)}) - logP(Y|\theta^{(i)}) =$$
$$[Q(\theta^{(i+1)},\theta^{(i)}) - Q((\theta^{(i)},\theta^{(i)}))] - [H(\theta^{(i+1)},\theta^{(i)}) - H((\theta^{(i)},\theta^{(i)}))]$$

其中对H：

$$[H(\theta^{(i+1)},\theta^{(i)}) - H((\theta^{(i)},\theta^{(i)}))] = \sum_z (logP(Z|Y,\theta^{(i+1)})) =$$
$$\sum_z (log\frac{P(Z|Y,\theta^{(i+1)})}{P(Z|Y,\theta^{(i)})}P(Z|Y,\theta^{(i)}) \leq$$
$$log(\sum_Z \frac{P(Z|Y,\theta^{(i+1)})}{P(Z|Y,\theta^{(i)})}P(Z|Y,\theta^{(i)}))...Jensen不等式得$$
$$log(\sum_Z P(Z|Y,\theta^{(i+1)})) = log1 = 0$$

对Q，由于Q的i+1项已经达到极大，所以有：

$$[Q(\theta^{(i+1)},\theta^{(i)}) - Q((\theta^{(i)},\theta^{(i)}))] \geq 0$$

最后证明得到：

$$logP(Y|\theta^{(i+1)}) \geq logP(Y|\theta^{(i)})$$

应用篇--GMM

EM在GMM中的应用

高斯分布分析

> 图中可知：
>
> 1. 单个高斯拟合效果差，均值应该分布在数据密集处
> 2. 混合高斯模型中的**隐变量**，同时，隐变量在概率模型中不能改变边缘分布，即：

> $$p(x_i) = \int_{z_i} p(x_i|z_i) * p(z_i) dz_i = \alpha_z = \sum_{z_i=1}^{k} \alpha_{z_i} N(x_i|\mu_z, \Sigma_z)$$
>
> 每个数据都有一个隐变量，告诉你在哪个高斯模型中(由两个高斯扩展到n个高斯)
>
> $$P(z_i = z_1|x_i, \Theta^g) = \frac{a}{a+b}$$
>
> 3.
>
> $$P(x) = \sum_{l=1}^{k} \alpha_l * N(X|\mu_l, \Sigma_l) \qquad \sum_{l=1}^{k} \alpha_l = 1$$
>
> $$\Theta = \{\alpha_1, \ldots, \alpha_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_{k-1}\}$$
>
> 4. 目标函数为：
>
> $$\Theta_{MLE} = argmax_{\Theta} * L(\Theta \mid X) = argmax_{\Theta}(\sum_{l=1}^{n} log * \sum_{l=1}^{k} \alpha_l N(X \mid \mu_l, \Sigma_l))$$
>
> 该式子包含和（或积分）的对数，不能像单个高斯模型那样直接求导，再令导数为0来求解。这时我们需要**利用 EM 算法通过迭代逐步近似极大化L(Θ|*X*)来求解。**

EM算法在GMM中的应用：

高斯混合模型的概率分布模型如下：

$$P(Y|\theta) = \sum_{K=1}^{K} \alpha_k \phi(Y|\theta)$$

$$\text{其中：} \sum_{K=1}^{K} \alpha_k = 1, \theta_k = (\mu_k, \sigma_k^2)$$

$$\theta = (\alpha_1, \alpha_2, \ldots, \alpha_k, \theta_1, \theta_2, \ldots, \theta_k)$$

$$\phi(Y|\theta) = \frac{1}{\sqrt{2\pi}\sigma_k} exp(-\frac{(y-\mu_k)^2}{2\sigma_k^2})$$

用 γ 作为隐变量，去确定是哪个模型，γ =1/0.

完整数据的似然函数为：

$$P(y, \gamma|\theta) = \prod_{j=1}^{N} P(y_j, \gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jK}|\theta)$$
$$= \prod_{K=1}^{K} \prod_{j=1}^{N} [\alpha_k \phi(y_j|\theta_k)]^{\gamma^{jK}}$$

完全数据的对数似然为：

$$logP(y, \gamma|\theta) = \sum_{K=1}^{K} [\sum_{j=1}^{N} log\alpha_k + \sum_{j=1}^{N} \gamma_{jk} [log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_K^2}(y_j - \mu_k)^2]]$$

EM算法中的E步：确定Q函数，对每一个隐变量求期望

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

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E[logP(y,\gamma|\theta)|y,\theta^{(i)}] \\
&= E[\sum_{K=1}^{K}[\sum_{j=1}^{N} log\alpha_k + \sum_{j=1}^{N}\gamma_{jk}[log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_K^2}(y_j-\mu_k)^2]]] \\
&= \sum_{K=1}^{K}[\sum_{j=1}^{N}(E_{\gamma_{jk}})log\alpha_K + \sum_{j=1}^{N}(E\gamma_{jk})[log(\frac{1}{\sqrt{2\pi}}) - log\sigma_K -]\frac{1}{2\sigma_k^2}(y_j-\mu_k)2]
\end{aligned}
$$

计算E

$$
\begin{aligned}
\gamma_{jk} &= E(\gamma_{jk}|y,\theta) \\
&= P(\gamma_{jk}=1|y,\theta)*1 + P(\gamma_{jk}=0,\theta)*0 \\
&= \frac{P(\gamma_{jk}=1,y_j|\theta)}{P(y_j|\theta)} = \frac{P(\gamma_{jk}=1,y_j|\theta)}{\sum_{K=1}^{K}P(\gamma=1,y_j|\theta)} \\
&= \ldots = \frac{\alpha_k\phi(y_j|\theta_k)}{\sum_{K=1}^{K}\alpha_k\phi(y_j|\theta_k)}
\end{aligned}
$$

最终求得Q

$$
Q(\theta,\theta^{(i)}) = \sum_{K=1}^{K}[\sum_{j=1}^{N} log\alpha_k + \sum_{j=1}^{N}\gamma_{jk}[log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2]]
$$

M步,进行迭代模型的参数

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

$$
\begin{aligned}
\mu_k &= \frac{\sum_{j=1}^{N}\gamma_{jk}y_j}{\sum_{j=1}^{N}\gamma_{jk}} \\
\sigma_k^2 &= \frac{\sum_{j=1}^{N}\gamma_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\gamma_{jk}} \\
\alpha_k &= \frac{\sum_{j=1}^{N}\gamma_{jk}}{N}
\end{aligned}
$$

参考1：知乎关于EM

参考2：GMM应用

参考3：EM九个境界

参考4：统计学习方法