法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，北风网和讲师拥有完全知识产权；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或者机构不得盗版、复制、仿造其中的创意和内容，我们保留一切通过法律手段追究违反者的权利。
- 课程详情请咨询
  - 微信公众号：北风教育
  - 官方网址：http://www.ibeifeng.com/

人工智能之机器学习

期望极大算法（Expectation Maximization）

主讲人：赵翌臣

上海育创网络科技有限公司

北风网 IBEIFENG.COM

期望极大算法（EM）

- 介绍
  - EM算法是一种迭代算法，1977年由 Dempster等人总结提出，用于含有隐变量(hidden variable)的概率模型参数的极大似然估计等
- 组成
  - E步，求期望(expectation) $Q(\theta,\theta^{(i)})=\sum_{Z}P(Z\mid Y,\theta^{(i)})\log P(Y,Z\mid\theta)$
  - M步，求极大(maximization)
- 应用
  - 高斯混合模型GMM、隐式马尔科夫模型HMM等

期望极大算法（EM）

- 概率模型有时既含有观测变量(observable variable)，又含有隐变量或潜在变量(latent variable)。如果概率模型的变量都是观测变量，那么给定数据，可以直接用极大似然估计法估计模型参数。但是，当模型含有隐变量时，就不能简单地使用这些估计方法。EM算法就是含有隐变量的概率模型参数的极大似然估计法

EM算法的引出

- 案例一：在校园里随机找了100个男生和100个女生把自己的性别、身高写在小纸条上，假设男生身高服从某一个正态分布，女生身高服从另一个正态分布。请估计学校男生和女生的身高分布。
- 极大似然估计：事件已经发生了，未知参数为多少时，让该事件发生的概率最大。

以男生为例，100个身高已经得到了，当全校男生的正态分布的参数为多少时，能让出现这100个身高的概率最大

假设全校男生身高分布为p(x|θ)，那么抽到这100个身高的概率（似然函数）为:

$$\mathrm{H}(\theta)=\prod_{i=1}^{100}\mathrm{p}(x_i;\theta)$$

对数似然函数为:

$$H(\theta)=\sum_{i=1}^{100} log\mathrm{p}(x_i;\theta)$$

求导，求极值，反解出让H(θ)最大的θ，即解出了男生的分布，女生同理

EM算法的引出

- 案例二：在校园里随机找了100个男生和100个女生把自己的~~性别~~、身高写在小纸条上，假设男生身高服从某一个正态分布，女生身高服从另一个正态分布。请估计学校男生和女生的身高分布。

这时，就无法将男女分开了，可以使用高斯混合模型来估计整体分布

- EM算法：含有隐变量的事件已经发生了，未知参数（μ1，μ2，σ1，σ2，λ1，λ2）为多少时，让该事件发生的概率最大。

EM算法的引出2

- 假设有3枚硬币，分别记作A，B，C。这些硬币正面出现的概率分别是π，p和q。进行如下掷硬币试验：先掷硬币A，根据其结果选出硬币B或硬币C，正面选硬币B，反面选硬币C；然后掷选出的硬币，掷硬币的结果，出现正面记作1，出现反面记作0；独立地重复n次试验(这里，n=10)，观测结果为 1,1,0,1,0,0,1,0,1,1

- 假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三硬币各自的正面朝上的概率，即三硬币模型的参数。三硬币模型可以写作

$$P(y|\theta)=\sum_z P(y,z|\theta)=\sum_z P(z|\theta)P(y|z,\theta)$$

$$=\pi p^y(1-p)^{1-y}+(1-\pi)q^y(1-q)^{1-y}$$

- 这里，随机变量y是观测变量，表示**一次**试验观测的结果是1或0；随机变量z是隐变量，表示未观测到的掷硬币A的结果；θ=(π，p，q)是模型参数。随机变量y的数据可以观测，随机变量z的数据不可观测。

EM算法的引出2

- 将观测数据表示为$Y=(Y_1, Y_2, \ldots, Y_n)^T$，未观测数据表示为$Z=(Z_1, Z_2, \ldots, Z_n)^T$，则观测数据的似然函数为

$$P(Y \mid \theta) = \sum_Z P(Z \mid \theta) P(Y \mid Z, \theta)$$

- 考虑求模型参数$\theta=(\pi, p, q)$的极大似然估计，即

$$\hat{\theta} = \arg\max_\theta \log P(Y \mid \theta)$$

- 这个问题没有解析解。EM算法是用于求解这种问题的一种迭代算法。EM算法就是含有隐变量的概率模型参数的极大似然估计法

EM算法

- 我们面对一个含有隐变量的概率模型，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

$$L(\theta)=\log P(Y|\theta)=\log\sum_Z P(Y,Z|\theta)=\log\left(\sum_Z P(Y|Z,\theta)P(Z|\theta)\right)$$

- 注意到这一极大化的主要困难是式中有未观测数据并有包含和(或积分)的对数。事实上，EM算法是通过迭代逐步近似极大化L(θ)的。假设在第i次迭代后θ的估计值是$\theta^{(i)}$。我们希望新估计值θ能使L(θ)增加，即L(θ)>L($\theta^{(i)}$)，并逐步达到极大值。为此，考虑两者的差：

$$L(\theta)-L(\theta^{(i)})=\log\left(\sum_Z P(Y|Z,\theta)P(Z|\theta)\right)-\log P(Y|\theta^{(i)})$$

- 利用Jensen不等式(Jensen inequality)得到其下界

$$\log\sum_j \lambda_j y_j \ge \sum_j \lambda_j \log y_j \quad ,\lambda_j \ge 0,\sum_j \lambda_j = 1$$

$$\log\sum_j \lambda_j y_j \ge \sum_j \lambda_j \log y_j \quad ,\lambda_j \ge 0, \sum_j \lambda_j = 1$$

$$L(\theta)-L(\theta^{(i)}) = \log\left(\sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})}\right) - \log P(Y|\theta^{(i)})$$

$$\ge \sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - \log P(Y|\theta^{(i)})$$

$$= \sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}$$

L(θ)的下界为： $B(\theta,\theta^{(i)}) = L(\theta^{(i)}) + \sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}$

因此，任何使B增大的θ，也可以让L(θ)增大，为了使L(θ)有尽可能大的增长，选择$\theta^{(i+1)}$使B达到极大，即

$$\theta^{(i+1)} = \arg\max_\theta B(\theta,\theta^{(i)})$$

EM算法

- 省去对求θ极大化而言是常数的项：

$$\theta^{(i+1)}=\arg\max_{\theta}\left(L(\theta^{(i)})+\sum_{Z}P(Z|Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}\right)$$

$$=\arg\max_{\theta}\left(\sum_{Z}P(Z|Y,\theta^{(i)})\log P(Y|Z,\theta)P(Z|\theta)\right)$$

$$=\arg\max_{\theta}\left(\sum_{Z}P(Z|Y,\theta^{(i)})\log P(Y,Z|\theta)\right)$$

$$=\arg\max_{\theta}Q(\theta,\theta^{(i)})$$

(Q函数)完全数据的对数似然函数logP(Y,Z|θ)关于在给定观测数据Y和当前参数$\theta^{(i)}$下对未观测数据Z的条件概率分布P(Z|Y, $\theta^{(i)}$)的期望称为Q函数

EM算法步骤

$$Q(\theta,\theta^{(i)})=\sum_{Z}P(Z|Y,\theta^{(i)})\log P(Y,Z|\theta)$$

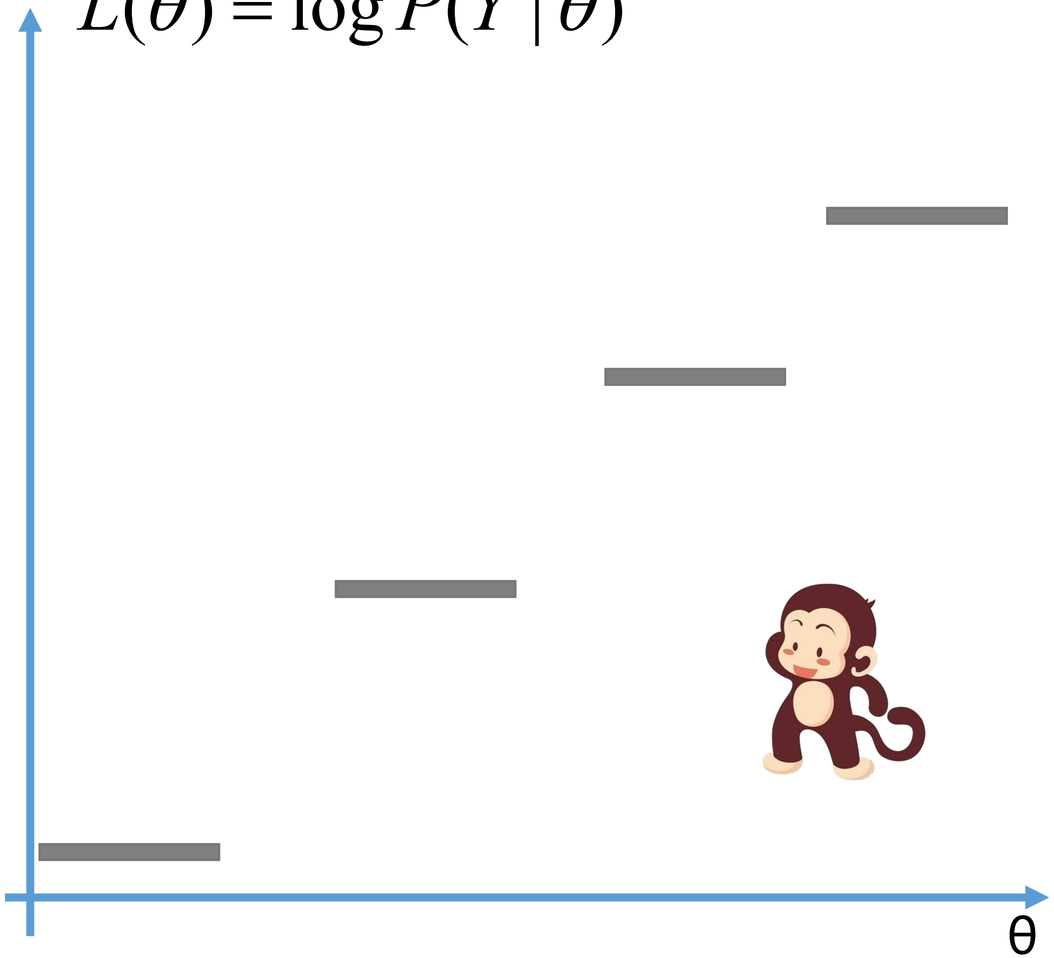
- 步骤一： 参数的初值可以任意选择，但需注意EM算法对初值是敏感的。
- 步骤二： E步求Q(θ,θ⁽ⁱ⁾)。Q函数式中Z是未观测数据，Y是观测数据。注意，Q(θ,θ⁽ⁱ⁾)的第1个变元表示要极大化的参数，第2个变元表示参数的当前估计值。
- 步骤三： M步求Q(θ,θ⁽ⁱ⁾)的极大化，得到θ⁽ⁱ⁺¹⁾。
- 步骤四： 给出停止迭代的条件，一般是对较小的正数ε_1，ε_2，若满足下面条件则停止迭代

$$\left\|\theta^{(i+1)}-\theta^{(i)}\right\|<\varepsilon_1 \quad or \quad \left\|Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\right\|<\varepsilon_2$$

EM算法简易理解

- EM算法就是含有隐变量的概率模型参数的极大似然估计法
- 步骤一： 小猴子先随机找个台阶
- 步骤二（E步）： 求小猴子在当前位置起跳，跳起高度的期望（Q函数，内部有变量θ），这是一个下界（小猴子至少能跳这么高）
- 步骤三（M步）： 对Q函数求极大，让小猴子起跳高度的下界尽可能高，也就间接让小猴子跳得高。
- 步骤四：给出停止迭代的条件（小猴子横向跳的不远了 or 小猴子跳的高度不高了 ）

$$L(\theta)=\log P(Y\mid\theta)$$



THANK YOU

上海育创网络科技有限公司

三硬币模型*

完全数据的
对数似然函数为：
j表示序列索引
zj表示第j个隐状态

$$\log(P(Y,Z|\theta)) = \log(\prod_{j=1}^{n} p(y_j, z_j|\theta))$$
$$= \sum_{j=1}^{n} \log(p(y_j, z_j|\theta))$$

期望为：
i表示EM轮数
θ(i)表示当前轮参数

$$E_{Z|Y,\theta^{(i)}}[\log(P(Y,Z|\theta))]$$
$$= \sum_{j=1}^{n}\sum_{z_j}[p(z_j|y_j,\theta^{(i)})\log(p(y_j,z_j|\theta))]$$
$$= \sum_{j=1}^{n}\left\{\begin{array}{l}[p(z_j=1|y_j,\theta^{(i)})\log(p(y_j,z_j=1|\theta))] \\ +[p(z_j=0|y_j,\theta^{(i)})\log(p(y_j,z_j=0|\theta))]\end{array}\right\}$$

后验概率　　联合概率

三硬币模型*

联合概率

$$p(y_j, z_j = 1|\theta) = p(y_j|z_j = 1, \theta)p(z_j = 1|\theta)$$
$$= \pi p^{y_j}(1 - p)^{1-y_j}$$
$$p(y_j, z_j = 0|\theta) = p(y_j|z_j = 0, \theta)p(z_j = 0|\theta)$$
$$= (1 - \pi)\, q^{y_j}(1 - q)^{1-y_j}$$

后验概率

$$\mu_j^{(i+1)} = p(z_j = 1|y_j; \theta^{(i)})$$
$$= \frac{p(y_j,z_j=1)}{p(y_j)} = \frac{p(y_j|z_j=1)p(z_j=1)}{\sum_{z_j} p(y_j,z_j)}$$
$$= \frac{p(y_j|z_j=1)p(z_j=1)}{p(y_j,z_j=1)+p(y_j,z_j=0)}$$
$$= \frac{(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j} * \pi^{(i)}}{(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j} * \pi^{(i)} + (q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j} * (1-\pi^{(i)})}$$

1.这里计算了一个在模型参数下，观测数据yi来自掷硬币B的概率
2.来自硬币C的概率可以用1-它

三硬币模型*

最终结果：

$$
\begin{aligned}
&E_{Z|Y,\theta^{(i)}}[\log(P(Y,Z|\theta))] \\
&= \sum_{j=1}^{n}\left\{\begin{array}{l}[p(z_j=1|y_j,\theta^{(i)})\log(p(y_j,z_j=1|\theta))] \\ +[p(z_j=0|y_j,\theta^{(i)})\log(p(y_j,z_j=1|\theta))]\end{array}\right\} \\
&= \sum_{j=1}^{n}\left\{\begin{array}{l}\mu_j^{(i+1)} * \log\left(\pi p^{y_j}(1-p)^{1-y_j}\right) \\ +\left(1-\mu_j^{(i+1)}\right) * \log\left((1-\pi)\,q^{y_j}(1-q)^{1-y_j}\right)\end{array}\right\}
\end{aligned}
$$

E-Step Done!

三硬币模型*

对E-Step的式子求极值，对参数求偏导，令其为0即可

$$\frac{\partial f}{\partial \pi}=\sum_{j=1}^{n}\left\{\mu_j^{(i+1)}*\frac{1}{\pi}-\left(1-\mu_j^{(i+1)}\right)*\frac{1}{1-\pi}\right\}$$

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

$$=\frac{n\pi-\sum_{j=1}^{n}\mu_j^{(i+1)}}{\pi(1-\pi)}=0$$

$$\frac{\partial f}{\partial p}=\sum_{j=1}^{n}\mu_j^{(i+1)}*\frac{\pi\left\{y_j p^{y_j-1}(1-p)^{1-y_j}+p^{y_j}\left[-(1-y_j)(1-p)^{-y_j}\right]\right\}}{\pi p^{y_j}(1-p)^{1-y_j}}$$

$$=\sum_{j=1}^{n}\mu_j^{(i+1)}*\frac{\left\{y_j p^{y_j-1}(1-p)^{-y_j}*(1-p)+p^{y_j-1}*p\left[(y_j-1)(1-p)^{-y_j}\right]\right\}}{p^{y_j}(1-p)^{1-y_j}}$$

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

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

$$=\sum_{j=1}^{n}\mu_j^{(i+1)}*\frac{\{y_j-p\}}{p(1-p)}=0$$

$$\frac{\partial f}{\partial q}=\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right)*\frac{(1-\pi)\left\{y_j q^{y_j-1}(1-q)^{1-y_j}+q^{y_j}\left[-(1-y_j)(1-q)^{-y_j}\right]\right\}}{(1-\pi)q^{y_j}(1-q)^{1-y_j}}$$

$$=\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right)*\frac{\left\{y_j q^{y_j-1}(1-q)^{-y_j}*(1-q)+q^{y_j-1}*q\left[(y_j-1)(1-q)^{-y_j}\right]\right\}}{q^{y_j}(1-q)^{1-y_j}}$$

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

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

$$=\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right)*\frac{\{y_j-q\}}{p(1-q)}=0$$

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

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

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

M-Step Done!