生成模型读书笔记三

2021年1月10日　22:35

1. **EM**
   a. 再从分类问题说起

      回顾上面MLE中的式子

$$\theta = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{X} \tilde{p}(X) \log p_\theta(X)$$

      在分类问题中，我们的数据样本包含特征X和标签Y，所以不再只是X，而是(X,Y)，所以把(X,Y)代进上面的式子，再结合p(X,Y)=p(Y|X)*p(X)可以得到

$$\theta = \arg\max_{\theta} \sum_{X,Y} \tilde{p}(X,Y) \log \left[p_\theta(X) p_\theta(Y|X)\right]$$

      因为是分类问题，所以我们感兴趣的是p(Y|X)，$p_\theta(x)$可以认为是个常数(这么想，一个动物有羽毛，求它是鸟的概率，我们建模的是p(鸟|羽毛)，这个建模完全不影响p(羽毛)，所以它跟真实分布是一样的，是个常数)，再把$\tilde{p}(X,Y)$拆成$\tilde{p}(X)*\tilde{p}(Y|X)$，由于$\tilde{p}$指真实分布，所以对于一个确定的样本X，Y也是确定的，所以$\tilde{p}(Y|X)$也是常数，最后就得到了下面分类问题中的最大似然式子，Y_t是目标标签

$$\theta = \arg\max_{\theta} \sum_{X} \tilde{p}(X) \log p_\theta(Y_t|X)$$

$$\theta = \arg\max_{\theta} \mathbb{E}_X\left[\log p_\theta(Y_t|X)\right]$$

   b. 隐变量

      上面的分类问题是个有监督问题，如果是无监督问题，比如聚类，样本数据中不包含Y呢？模型的参数该如何估计？最大似然的式子是

$$\theta = \arg\max_{\theta} \mathbb{E}_{X,Y}\left[\log p_\theta(X,Y)\right]$$

      我们知道有Y的存在，但是具体Y是什么我们不知道，这时候Y就叫做隐变量。

      例子1：K-means，最简单的例子

      回想一下K-means是如何聚类的。首先随机选K个点作为类的质心z_k，然后计算剩下的点和k个质心之间的距离，将每个数据点x_i分配到离它最近的质心所属的类，然后重新计算每个类的质心z，重新计算所属类别，直到收敛为止。这里面，质心就相当于隐变量，数据样本就是观测变量，想学习的模型就是p(y|x)，每个数据点x的标签y

      例子2：高斯混合模型GMM

      GMM是指数据分布是多个高斯分布的混合，我们采样到了数据，但是不知道数据X是从哪个高斯分布里采样出来的，这时候类别c就是一个隐变量

   c. 从极大似然的角度来解决包含隐变量的问题

      i. 思想

         依然是估计模型参数的问题，不过现在情况是存在隐变量，也即数据缺失。

         主要思想为首先根据给出的观测数据，估计出模型参数的值，再根据上一步的估计参数值估计缺失数据(隐变量)的值(分布)，再根据估计出的缺失数据加上已知的观

测数据，估计参数的值，以此类推，迭代直至收敛。

ii. 预备知识：Jensen不等式

如果f是凸函数，X是随机变量，那么： $E[f(X)] \geq f(E[X])$

如果f是凹函数，不等号反向， log是凹函数所以E[log(X)] ≤ log(E[X])

当X是常数时，不等式取等号

iii. 从最大似然分析

当存在未知的隐变量z时，似然函数p(x; θ)就变成了联合分布p(x,z; θ)的x的边缘分布，可以写成

$$\hat{\theta} = argmax \sum_{i=1}^{n} logp(x_i;\theta) = argmax \sum_{i=1}^{n} log \sum_{z_i} p(x_i, z_i;\theta)$$

上面式子是无法直接求解的，因为数据中的z未知（这里注意一个前提，我们预设的模型就是p(x,z; θ)而不再是p(x; θ), 我们知道有z的存在，只是数据观测不到，比如GMM我们知道有k个模型的混合，但是只能观测到x，而不知道x是从哪个模型里采样的）。

所以要做一些变换。注意，下面推导过程就是EM过程。

首先给p(xi,zi;θ)乘以除以一个分布Q，再利用jensen不等式，

log(E[X])≥E[log(X)]，(1)式log右边的式子是基于Q(z)分布的p/Q的期望，最后找到了最大似然的一个下界

$$\sum_{i=1}^{n} log \sum_{z_i} p(x_i, z_i;\theta) = \sum_{i=1}^{n} log \sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i;\theta)}{Q_i(z_i)} (1)$$

$$\geq \sum_{i=1}^{n} \sum_{z_i} Q_i(z_i) log \frac{p(x_i, z_i;\theta)}{Q_i(z_i)} (2)$$

根据上面Jensen不等式的介绍，如果要让不等式取等号，则log里面的式子需要为常数，也即：

$$\frac{p(x_i, z_i;\theta)}{Q_i(z_i)} = c，\ c为常数$$

变换一下：

P(x_i,z_i; θ) / c = $Q_i(z_i)$, 因为Q是一个分布，所以两边对z_i求和，$\sum_{z_i} Q_i(z_i)$=1, 从而有

$$p(x_i, z_i;\ \theta) = c * Q_i(z_i)$$

$$\sum_{z_i} p(x_i, z_i;\ \theta) = c * \sum_{z_i} Q(z_i) = c$$

将$\sum_{z_i} p(x_i, z_i;\ \theta) = c$代进上式的c，就得到了

$$Q_i(z_i) = \frac{p(x_i, z_i;\theta)}{c}$$

$$Q_i(z_i) = \frac{p(x_i, z_i;\theta)}{\sum_z p(x_i, z_i;\theta)} = \frac{p(x_i, z_i;\theta)}{p(x_i;\theta)} = p(z_i|x_i;\theta)$$

通过上述式子可以算出引入的函数$Q_i(z_i)$就是z的后验概率，这时候不等式取等号，最大化似然就相当于最大化不等式右边的式子

$$argmax \sum_{i=1}^{n} \sum_{z_i} Q_i(z_i) log \frac{p(x_i, z_i;\theta)}{Q_i(z_i)}$$

至此，其实EM算法的整个过程已经出来了，来整理一下我们是怎么做的：

为了最大化似然log$p(x_i, z_i; \theta)$，我们引入Q函数给它构造了一个期望的形式，再通过Jensen不等式找到了这个期望的一个下界。为了让最大化下界等价于最大化似然，不等式要取等号，经过计算，我们知道了当Q函数为z的后验概率时，不等式可以取等号。假设我们能够把后验概率算出来，那么期望就能写出来了(E步)，这时候它就是一个只含有参数的θ的式子，对它求取最大值时的θ(M步)

PS: 一般能用EM算法的，后验分布p(z|x)和联合分布p(x,z)都是比较好求的，比如z的取值只有有限几个，分布形式简单等情况

优化目标的式子还可以进一步简化，式子中的log变成减法展开，可以得到

$$Q_i(z^{(i)})log\frac{P(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})} = Q_i(z^{(i)})logP(x^{(i)}, z^{(i)}|\theta) - Q_i(z^{(i)})logQ_i(z^{(i)})$$

因为优化目标是θ，右边项不含θ，所以相当于常量，可以去掉，就得到了

$$arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)})logP(x^{(i)}, z^{(i)}; \theta)$$

对z求和的部分，可以看成是似然函数logp(x,z; θ)基于分布Q(z)的期望，这也就是算法中E步的含义；

最后对上面式子求解，即优化θ求最大化，也就是算法中M步的含义

所以EM算法流程如下：

输入：观察数据$x = (x^{(1)}, x^{(2)}, \dots x^{(m)})$，联合分布$p(x, z; \theta)$，条件分布$p(z|x; \theta)$，最大迭代次数$J$。

1) 随机初始化模型参数θ的初值θ^0。

2) for j from 1 to J开始EM算法迭代：

a) E步：计算联合分布的条件概率期望：

$$Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}, \theta^j))$$

$$L(\theta, \theta^j) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)})logP(x^{(i)}, z^{(i)}; \theta)$$

b) M步：极大化$L(\theta, \theta^j)$,得到θ^{j+1}：

$$\theta^{j+1} = arg\max_{\theta} L(\theta, \theta^j)$$

c) 如果θ^{j+1}已收敛，则算法结束。否则继续回到步骤a)进行E步迭代。

输出：模型参数θ。

一个具体的用于计算GMM参数值的例子以及代码：
https://zhuanlan.zhihu.com/p/71010421

d. 另一种角度推导：从KL散度推导

如果从KL散度的角度来求呢?

再次回顾前面讲最大似然的部分，最大似然和最小化KL散度是相通的，上面的最大似然就是让含有未知参数以及隐变量的分布函数似然最大，如果从KL散度来推导就是要让含

有未知参数和隐变量的函数距离真实分布的距离最小，那么下面尝试从KL散度角度来推导一下：

假设p̃是真实分布，X是观测变量，Y是隐变量，p_θ是带有未知参数θ的预设分布函数，以GMM为例

$$
\begin{aligned}
&KL\left(\tilde{p}(X,Y)\|p_\theta(X,Y)\right)\\
=&\sum_{X,Y}\tilde{p}(X,Y)\log\frac{\tilde{p}(X,Y)}{p_\theta(X,Y)}\\
=&\sum_X\tilde{p}(X)\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)\tilde{p}(X)}{p_\theta(X|Y)p_\theta(Y)}\\
=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)\tilde{p}(X)}{p_\theta(X|Y)p_\theta(Y)}\right]\\
=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}+\sum_Y\tilde{p}(Y|X)\log\tilde{p}(X)\right]\\
=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}\right]+\mathbb{E}\left[\log\tilde{p}(X)\right]\\
=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}\right]+\text{常数}
\end{aligned}
$$

算出来的式子中有两个未知的，一是p~(Y|X), 二是模型中的参数θ，所以现在来运用EM算法的思想交替计算。首先给定θ，推导一下怎么算p~(Y|X)。

$$
\begin{aligned}
&\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_{\theta^{(r)}}(X,Y)}\\
=&\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_{\theta^{(r)}}(Y|X)}-\sum_Y\tilde{p}(Y|X)\log p_{\theta^{(r)}}(X)\\
=&KL\left(\tilde{p}(Y|X)\|p_{\theta^{(r)}}(Y|X)\right)-\text{常数}
\end{aligned}
$$

所以要求整个式子的最小值，相当于要求这个KL散度的最小值，KL≥0，最小值就是0，并且在p~(Y|X)=p_θ(Y|X)时取得最小值0，所以

$$
\tilde{p}(Y|X)=p_{\theta^{(r)}}(Y|X)=\frac{p_{\theta^{(r)}}(Y)p_{\theta^{(r)}}(X|Y)}{\sum_Y p_{\theta^{(r)}}(Y)p_{\theta^{(r)}}(X|Y)}
$$

回忆一下从最大似然角度推导的Q函数：

$$
Q_i(z_i)=\frac{p(x_i,z_i;\theta)}{\sum_z p(x_i,z_i;\theta)}=\frac{p(x_i,z_i;\theta)}{p(x_i;\theta)}=p(z_i|x_i;\theta)
$$

是不是一样，所以给定θ求p~(Y|X)就是前面介绍的EM算法里的E步，并且p_θ(Y|X)就是当中的Q函数，得到了与上文一致的结论；

得到了当前p~(Y|X)之后就可以计算原KL散度基于θ的最小值了

我们再一步展开，将最后一行写成log的减法，即

$E_x[\sum_Y \tilde{p}(Y|X)\log\tilde{p}(Y|X) - \sum_Y \tilde{p}(Y|X)\log p_\theta(X|Y)p_\theta(Y)]$

因为优化的目标是找到θ，使得这个KL散度最小，第一项不含θ，所以与优化目标无关，可以看做常量，因此要优化的KL散度最终表达式为

$-E_x\sum_Y \tilde{p}(Y|X)\log p_\theta(X|Y)p_\theta(Y)$

$\text{argmin}_\theta\ -E_x[\sum_Y \tilde{p}(Y|X)\log p_\theta(X|Y)p_\theta(Y)]$

$\text{argmax}_\theta\ E_x[\sum_Y \tilde{p}(Y|X)\log p_\theta(X|Y)p_\theta(Y)]$

这个式子中括号里的式子就是似然函数logp(X,Y; θ)基于分布Y后验分布的期望

与在最大似然分析Q函数的时候，"似然函数logp(x,z; θ)基于分布Q(z)的期望"一致
最终最小化原KL散度相当于最大化这个期望值，这就对应EM算法的M步，至此，推导结束，并且可以惊喜地发现推导结果与从最大似然角度来推是一样的

e. 总结

EM算法，就是对复杂目标函数的交替训练方法。E步，就是求似然函数基于隐变量后验分布的期望，对于不同的后验分布似然函数会有不同的期望，M步，就是求这个期望的最大值。

2. GMM

a. 概率模型的一个具体例子

上面提到过高斯混合模型，这里详细说说，也是一种很常见的生成模型。其定义如下：
混合模型：数据总体分布由k个子分布组成
混合高斯模型：每个子分布都是一个高斯分布
有如下几点：

i. GMM模型共由k个高斯分布组成；
ii. 观测数据属于第k个分布的概率是α_k，且$\sum\alpha_k=1$
iii. 第k个分布的均值是μ_k，方差是σ_k

因此，如果将高斯分布记作φ(x|θ)，那么GMM概率分布可以写作：

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

对于这个模型，它的参数为

$$\theta=(\tilde{\mu_k},\tilde{\sigma_k},\tilde{\alpha_k})$$

隐变量为k，表示x是从第k个高斯分布中采样的，它的似然函数可以写成：

$$logL(\theta)=\sum_{j=1}^{N}logP(x_j|\theta)=\sum_{j=1}^{N}log(\sum_{k=1}^{K}\alpha_k\phi(x|\theta_k))$$

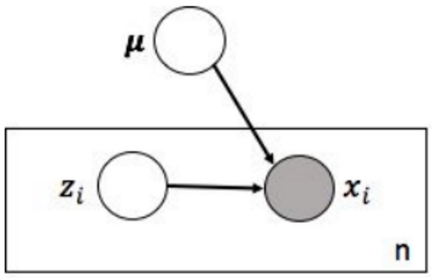
隐变量k的后验分布可以记为$\omega_{i,k}$，表示第i个数据是从第k个子分布采样出来的概率，通过下面式子可以算出来

$$\omega_{i,k}=p(z=k|x_i,\mu_k,\sigma_k)=\frac{\alpha_k N(x_i|\mu_k,\sigma_k)}{\sum_k\alpha_k N(x_i|\mu_k,\sigma_k)}$$

于是就可以用前面介绍的EM算法求解GMM中的参数

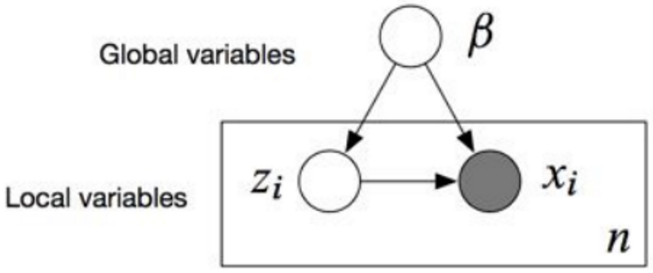
b. 概率图

如果用一个图将GMM中各个变量的关系表示出来，可以画成下面这个样子。图中μ和z都是未知的，对应于模型中的θ和k，数据集有n个样本，每个样本都有观测到的数据x_i，无法观测的隐变量z_i，数据的概率密度函数带有参数μ。箭头表示依赖关系，参数变量μ和隐变量z决定了x的值，即x的分布是以μ，z为条件的条件分布。



读懂盘子图很简单，一看变量，白圈是隐含变量，盘里的是局部变量，盘外的是全局变量，灰圈是观测值；二看盘子，盘子表示里面的变量zi和xi独立重复n次；三看依赖，箭头表示生成谁需要谁。

推广到一般的模型，有



$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

全局变量：模型参数(在贝叶斯学派中，一切都是随机的，所以参数也不是单独一个值，也是服从某个概率分布的变量。不过这些不重要，暂时用不上，这里主要说明一下为什么这么画)；

局部变量：分为观测到的数据x_i和未观测到的隐变量z_i，每个样本(x_i, z_i)都是独立同分布的