斯坦福 ML 公开课 13A

本文对应公开课的第 13 个视频，这个视频仍然和 EM 算法非常相关，第 12 个视频讲解了 EM 算法的基础，本视频则是在讲 EM 算法的应用。本视频的主要内容包括混合高斯模型（Mixture of Gaussian, MoG）的 EM 推导、混合贝叶斯模型（Mixture of Naive Bayes，MoNB）的 EM 推导、因子分析模型（Factor Analysis Model）及其 EM 求解。由于本章内容较多，故而分为 AB 两篇，本篇介绍至混合模型的问题。

很久没有写这个系列的笔记了，各种事情加各种懒导致的。虽然慢但是我还是会坚持把它写完的，就像一只打不死的小强，也像古人说的那样，十年可以不将军，但还需日拱一卒。闲话少说，进入正题。

回顾 EM 算法

先回顾一下 EM 算法，我们为什么要使用 EM 算法呢？原因就在于面对如公式 1 中的目标函数时，直接求导是无法解决的或者解决起来比较麻烦，EM 算法通过找下界的方式巧妙的将连加符号移到对数函数之外，使得该问题可解。尤其是当联合概率函数式指数族函数时，EM 的 E-step 和 M-step 都很易求，高斯混合模型就是这样的问题。

$$\ell(\theta)=\sum_{i=1}^{m} logp\left(x^{(i)};\theta\right)=\sum_{i=1}^{m} log\sum_{z^{(i)}} p\left(x^{(i)},z^{(i)};\theta\right) \tag{1}$$

还有另外一种理解 EM 算法的方式，那就是将它看作是坐标上升优化。

$$J(Q,\theta)=\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i\left(z^{(i)}\right)log\frac{p\left(x^{(i)},z^{(i)};\theta\right)}{Q_i\left(z^{(i)}\right)} \tag{2}$$

将 EM 算法要优化的目标看作是以Q,θ为参数的函数，那么 E-step 就是保持θ不变，优化 Q；M-step 就是保持 Q 不变优化θ。

MoG 的 EM 推导

MoG 模型的 E-step 和 M-step 分别按照笔记 12 中的公式 23，公式 24 那样进行推导。对于 E-step：

$$Q_i\left(z^{(i)}=j\right)=p\left(z^{(i)}\middle|x^{(i)};\psi,\mu,\Sigma\right)=\frac{p\left(x^{(i)}\middle|z^{(i)}=j;\mu,\Sigma\right)p\left(z^{(i)}=j\middle|\psi\right)}{\sum_k p\left(x^{(i)}\middle|z^{(i)}=k;\mu,\Sigma\right)p\left(z^{(i)}=k\middle|\psi\right)} \tag{3}$$

其中，$\mathrm{p}\left(x^{(i)}\middle|z^{(i)}\right)\sim\mathrm{N}(\mu_j,\Sigma_j)$，$z^{(i)}\sim Multinomial(\phi)$，代入即能求得$Q_i$。

令：

$$w_j^{(\mathrm{i})}=Q_i\left(z^{(i)}=j\right)$$

对于 M-step，

$$\max_{\psi,\mu,\Sigma}\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)})log\frac{p\left(x^{(i)},z^{(i)};\psi,\mu,\Sigma\right)}{Q_i(z^{(i)})}$$

$$=\max_{\psi,\mu,\Sigma}\sum_{i=1}^{m}\sum_{j=1}^{k} Q_i(z^{(i)}=j)log\frac{p\left(x^{(i)}\middle|z^{(i)}=j;\mu,\Sigma\right)p\left(z^{(i)}=j\middle|\psi\right)}{Q_i(z^{(i)})}$$

$$= \max_{\psi,\mu,\Sigma} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(\mathrm{i})} log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)) * \phi_j}{w_j^{(\mathrm{i})}} \tag{4}$$

在公式 4 中，第一个等号是将 p(x,z)展开为 p(x|z)*p(z)，第二个等号是将 p(x|z)和 p(z)的密度函数展开。

对于公式4的结果，通过求偏导数然后使偏导为0来获得极大值时各参数的取值。例如，对于参数μ，偏导如公式 5：

$$\begin{aligned}
&\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(\mathrm{i})} log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)) * \phi_j}{w_j^{(\mathrm{i})}} \\
&= -\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(\mathrm{i})} \frac{1}{2}\left(x^{(i)}-\mu_j\right)^T \Sigma_j^{-1}(x^{(i)}-\mu_j) \\
&= \frac{1}{2} \sum_{i=1}^{m} w_l^{(\mathrm{i})} \nabla_{\mu_l} 2\mu_l^T \Sigma_j^{-1} x^{(i)} - \mu_l^T \Sigma_j^{-1} \mu_l \\
&= \sum_{i=1}^{m} w_l^{(\mathrm{i})} (\Sigma_j^{-1} x^{(i)} - \Sigma_j^{-1} \mu_l)
\end{aligned} \tag{5}$$

将公式 5 设为 0，可以得到μ_l的迭代公式，如公式 6：

$$\mu_{\mathrm{l}} := \frac{\sum_{i=1}^{m} w_l^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_l^{(i)}} \tag{6}$$

再进一步，求解参数ϕ_j的更新规则。去除和ϕ_j无关的项后，要求偏导的函数为：

$$\sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(\mathrm{i})} log\phi_j \tag{7}$$

但是，因为ϕ_j是多项分布的概率值，所以ϕ_j比μ_l多一个约束条件，即$\sum_j \phi_j = 1$。这时，就要对其拉格朗日函数进行求导了，拉格朗日函数如下：

$$\mathcal{L}(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(\mathrm{i})} log\phi_j + \beta(\sum_{j=1}^{k} \varphi_j - 1) \tag{8}$$

其中，β是拉格朗日乘子。虽然ϕ_j还有不小于 0 的约束条件，但发现只是用上述约束得到的结果往往都是满足这个条件，所以在拉格朗日函数中没有添加该条件。

对拉格朗日函数求偏导：

$$\frac{\partial}{\partial \phi_j} \mathcal{L}(\phi) = \sum_{\mathrm{i}=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \beta \tag{9}$$

令偏导等于 0：

$$\phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)}}{-\beta} \tag{10}$$

所以，$\phi_j \propto \sum_{i=1}^{m} w_j^{(i)}$，又因为$\sum_j \phi_j = 1$。所以可得：

$$-\beta = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} = m \tag{11}$$

所以：

$$\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)} \tag{12}$$

MoNB 的 EM 推导

设想一个文本聚类过程，该过程可以应用于新闻消息的聚类，将相同事件或者相同主题的新闻进行聚合。

文本聚类可以看成是一个混合贝叶斯模型，简单起见，假设只有两个类，且采用伯努利事件模型。

假设有 m 个样本$\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$，其中每个样本都是 N 维向量，每个分量的值都只有 0、1 两个，即$x^{(i)} \in \{0,1\}^n$，且$x_j^{(i)} = I\{$词语 j 在文档 i 中是否出现$\}$。类别$z^{(i)} \in \{0,1\}$。

在开始的时候，我们并不知道$z^{(i)}$的值。

对于本模型来说，有如下几个基本参数，参数 1 在公式 13：

$$\phi_z = p(z = 1) = p(z^{(i)} = 1) \tag{13}$$

其中，$z^{(i)} \sim Bernoulli(\phi)$，所以，z 服从两点分布。之所以不把参数分开设为$\phi_{z=1}$和$\phi_{z=0}$两个，是因为它们之和为 1，一个确定后另一个也就固定了。本模型中由于类别和分量都是二值的，所以参数都这样设计。

还有两个参数在公式 14、15：

$$\varphi_{j|z=1} = p(x_j^{(i)} = 1 | z^{(i)} = 1) \tag{14}$$

$$\varphi_{j|z=0} = p(x_j^{(i)} = 1 | z^{(i)} = 0) \tag{15}$$

同理，之所以让$x_j^{(i)} = 1$也是因为$x_j^{(i)}$的二值性。

又由于贝叶斯模型的独立性假设，我们还有如下计算公式：

$$p\left(x^{(i)} \middle| z^{(i)}\right) = \prod_{j=1}^{n} p(x_j^{(i)} | z^{(i)}) \tag{16}$$

参数定义完后，就可以将其代入到 EM 计算框架中了。对于 E-step，有：

$$w^{(i)} = p\left(z^{(i)} = 1 \middle| x^{(i)}; \varphi_{j|z}, \phi_z\right) = \frac{p\left(x^{(i)} \middle| z^{(i)} = 1\right) p(z^{(i)} = 1)}{\sum_{j=0}^{1} p(x^{(i)} | z^{(i)} = j) p(z^{(i)} = j)} \tag{17}$$

其中，$p\left(x^{(i)} \middle| z^{(i)} = 1\right)$和$p(z^{(i)} = 1)$都可以由公式 13、14、15 定义的参数得到。注意公式 17 中$w^{(i)}$与 MoG 中的$w_j^{(i)}$的区别，之所以公式 17 中的 w 没有下标，是因为 z 是二值的，另一个参数可以由$1 - w^{(i)}$直接得到。

而对于 M-step，先写出其需要最大化的函数：

$$l\left(\varphi_{j|z}, \phi_z\right) = l(\theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i(z^{(i)})}$$

$$=\sum_{i=1}^{m}\left[w^{(i)}log\frac{p\left(x^{(i)},z^{(i)}=1;\varphi_{j|z},\phi_z\right)}{w^{(i)}}+\left(1-w^{(i)}\right)log\frac{p\left(x^{(i)},z^{(i)}=0;\varphi_{j|z},\phi_z\right)}{1-w^{(i)}}\right] \tag{18}$$

其中，

$$p\left(x^{(i)},z^{(i)}=1;\varphi_{j|z},\phi_z\right)=\prod_{j=1}^{n}p(x_j^{(i)}|z^{(i)}=1)\phi_z \tag{19}$$

$$p\left(x^{(i)},z^{(i)}=0;\varphi_{j|z},\phi_z\right)=\prod_{j=1}^{n}p(x_j^{(i)}|z^{(i)}=0)(1-\phi_z) \tag{20}$$

在公式 18 中，对ϕ_z求偏导：

$$\frac{\partial}{\partial\phi_z}l\left(\varphi_{j|z},\phi_z\right)=\frac{\partial}{\partial\phi_z}\sum_{i=1}^{m}[w^{(i)}log\phi_z+(1-w^{(i)})\log(1-\phi_z)]$$

$$=\sum_{i=1}^{m}[\frac{w^{(i)}}{\phi_z}-\frac{1-w^{(i)}}{1-\phi_z}] \tag{21}$$

令偏导为 0，可得：

$$\phi_z=\frac{\sum_{i=1}^{m}w^{(i)}}{m} \tag{22}$$

对$\varphi_{j|z=1}$求偏导：

$$\frac{\partial}{\partial\varphi_{j|z=1}}l\left(\varphi_{j|z},\phi_z\right)=\frac{\partial}{\partial\varphi_{j|z=1}}\sum_{i=1}^{m}w^{(i)}logp(x_j^{(i)}|z^{(i)}=1)$$

$$=\frac{\partial}{\partial\varphi_{j|z=1}}\sum_{i=1}^{m}w^{(i)}\log\left[(\varphi_{j|z=1})^{I\{x_j^{(i)}=1\}}(1-\varphi_{j|z=1})^{(1-I\left\{x_j^{(i)}=1\right\})}\right]$$

$$=\sum_{i=1}^{m}\left[\frac{w^{(i)}I\{x_j^{(i)}=1\}}{\varphi_{j|z=1}}-\frac{w^{(i)}(1-I\left\{x_j^{(i)}=1\right\})}{1-\varphi_{j|z=1}}\right] \tag{23}$$

令偏导为 0，可得：

$$\varphi_{j|z=1}=\frac{\sum_{i=1}^{m}w^{(i)}I\{x_j^{(i)}=1\}}{\sum_{i=1}^{m}w^{(i)}} \tag{24}$$

同理，对$\varphi_{j|z=0}$求偏导，然后令偏导为 0，可得公式 25，过程类似公式 23，故忽略。

$$\varphi_{j|z=0}=\frac{\sum_{i=1}^{m}(1-w^{(i)})I\{x_j^{(i)}=1\}}{\sum_{i=1}^{m}(1-w^{(i)})} \tag{25}$$

从公式 22、24、25 来看，可见混合贝叶斯模型的 EM 推导结果与朴素贝叶斯的极大似然估计十分相似。区别在于在朴素贝叶斯中，类别属性是已知的，而在混合贝叶斯中，类别属性是未知的，因而其多了一个概率表示w。

混合模型的问题

虽然 EM 算法很好很强大，可以很好的拟合混合模型，但是在上面的 MoG 和 MoNB 模

型中，要想得到一个较好的结果，需要有一个前提条件，即足够的数据量，m>>n，m 为样本数目，n 为每个样本的维度。当$m \approx n$或者$m \ll n$时，再应用 MoG 模型甚至是 Gaussian 模型，就会出现问题。

以数据符合高斯分布为例，那么使用极大似然估计，可以得到参数：

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

因为样本的数目小于维度，所以得到的方差Σ是奇异矩阵，所谓奇异矩阵，即特征值为 0，不满秩的矩阵。所以$|\Sigma|^{1/2} = 0$，因而不能写出其概率密度函数。为什么会这样呢？这相当于线性方程组求解，未知数的个数比方程的数目多，因而不能完全的求出所有未知数。

如何解决这个问题呢？可以对方差添加一些限制，比如将Σ设为对角矩阵，这样得到的概率分布的图形的轴是与坐标轴平行的。限制再严格一些，将对角矩阵的各个元素的值都设为相同的，这样得到的概率分布的图形轮廓都是圆形的。

用添加限制的方法确实能解决$|\Sigma|^{1/2} = 0$的问题，但是这样的建模会导致不同维度之间的相关性丢失。而因子分析模型就是来解决这个问题，它使用更多的参数来对Σ建模，能反映出一些维度之间的关联性信息，但是，它仍然不能拟合出一个完全的协方差矩阵。