混合高斯模型和EM算法

menglinwoo

December 2017

1 混合高斯模型和EM

混合模型表达能力更加丰富多样，如多峰分布，而每个分量仍是简单的概率模型，求解上相对简单。但是每个分量的大小是不可观测的，引入了潜在变量，从而导致极大化似然函数无法形式化求解。但是，如果每个分量可以观测，即可形式化求解。EM算法在E步骤去estimate每个分量的大小，在M步骤去形式化求解在评估出的分量的情形下的似然函数的极值。通过反复迭代逼近混合模型的最大化似然概率。

当然，上述EM两个步骤的结果能否收敛到混合模型的似然函数的极值是该方法是否有效的关键。通过Jesen不等式和KL散度等数学证明是可以证明这一点的。

2 混合高斯模型

高斯分布有许多优点并且普遍存在（大数定律），但是单峰函数，所以对于复杂的分布表达能力不足，可以用多个高斯分布的线性组合来逼近这些复杂的分布。GMM的线性组合形式：

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

我们假设存在K个潜在高斯分布, 采用one-hot的方式进行编码，每一个数据点都有一个向量$\boldsymbol{z}$，当某个数据点是由第k个分布产生的，则记元素z_k为1，其余为0。则π_k是第k个分布发生的概率：

$$p(z_k = 1) = \pi_k \tag{2}$$

则$z_k = 1$的后验概率为：

$$\begin{aligned} \gamma(z_k) &\equiv p(z_k = 1|\mathbf{x}) \\ &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \tag{3}$$

考察这个公式，我们可以将π_k看成$z_k = 1$的先验概率，将$\gamma(z_k)$看成观测到x分布之后对应的后验分布，通过公式2可以看出，$\gamma(z_k)$可以看成是分量k对于解释观测到数据的responsibility。

公式2中隐变量$\boldsymbol{z}$概率分布π_k未知,各个分量的高斯分布的参数未知。在给定一组观测的数据集$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$，将这个数据集表示成$N \times D$的矩阵，行方向为样本数，列方向为特征数。如果我们假定数据集是独立同分布地从概率分布抽取得到，则对数似然函数为：

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \Big\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\} \tag{4}$$

在对公式4求极值之前，先考察公式其他的问题。若某个分布坍塌到一个数据点，意味着对于某个分量j,$\sigma_j \to 0$，ln后求和的第j项中$\to \infty$,此时该目标函数已然趋向于∞,这种singularity会导致问题存在病态问题。解决这种问题的方法通常是在求解过程中一旦检测到高斯分量坍塌到一个点时，将它的均值设置成随机一个值，并将其方差设置成一个较大值，然后继续优化。

下面正式对公式4求解极值，由于ln后面是一个求和公式，因此在对各个高斯分布的参数以及各个分量的概率分布求解时不存在一个解析解，推导如下。令公式4关于高斯分量中均值μ_k的导数为零，有

$$0 = \sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{5}$$

故有

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n \tag{6}$$

同理令高斯变量中的方差Σ_k的导数为零，有

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T \tag{7}$$

最后对分量的系数π_k求解最大化4,同时应该注意限制条件:

$$\sum_{k=1}^{K}\pi_k = 1 \tag{8}$$

使用拉格朗日乘子法，有公式:

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) + \lambda(\sum_{k=1}^{K}\pi_k - 1) \tag{9}$$

可得

$$\pi_k = \frac{N_k}{N} \tag{10}$$

从上述推导中，不难可以看出，这些参数的求解结果不存在解析解。但是，如果各个分量的系数π_k确定，则公式5是一个close-formed的解，同样Σ_k的解也是close-formed，甚至连π_k的更新迭代公式也是close-formed。当然既然各个分量的系数π_k都是迭代的，则其一开始就不是确定了的。我们先假定这种反复迭代是可以逼近最大化对数似然函数解的，这个将在后续部分证明。

上述步骤可以分成两步：E步骤来计算在已知各个高斯分布的参数Σ_k和μ来求解各个分量的系数π_k，即每个π_k的后验分布，也就是所谓的分量k对观测到变量解释的责任。M步骤是上面最大化对数似然函数得到三个公式。反复使用E步骤和M步骤，直至似然函数的目标函数值收敛。

3 从图模型来审视混合高斯模型

从图模型的角度来讲，变量$\boldsymbol{Z}$是无法观测的隐变量，而变量$\boldsymbol{X}$是观测到的数据集，但整个GMM是一个由潜在变量量$\boldsymbol{Z}$和$\boldsymbol{X}$所构成完整数据集的联合分布。由于我们无法观测隐变量，给出的对数似然函数是:

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln(\sum_{\boldsymbol{Z}} p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})) \tag{11}$$

在这里我们假定完整数据集的对数似然函数的极值求解时非常容易的，如GMM模型中，一旦观测某个数据点的$\boldsymbol{Z}$和$\boldsymbol{X}$，在ln后的求和公式事实上只有一项，即确定为标签的那一项。求和公式不在存在，因此求其极值是非常容易的，但事实上隐变量无法观测。而在EM算法的E步骤，我们是拿到某个数据点的$\boldsymbol{Z}$后验分布，所以求极值非常容易。

4 一般形式的EM算法

根据上面的分析，不难看出，EM算法过程在E步骤使用了得到观测数据$\boldsymbol{X}$的$\boldsymbol{Z}$的后验分布去近似真实的$p(\boldsymbol{Z})$。真实情形下，似然函数为:

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln(\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta})) \tag{12}$$

而近似情形下，我们把近似的$\boldsymbol{Z}$的后验分布记成$q(\boldsymbol{Z})$，那么其似然函数为

$$\ln \widetilde{p}(\boldsymbol{X}|\boldsymbol{\theta}) = \ln(\sum_{\boldsymbol{Z}} q(\boldsymbol{Z})p(\boldsymbol{X}|\boldsymbol{\theta})) \tag{13}$$

显然同样需要评估一下这种概率分布的近似效果，通常使用KL散度衡量。在这里我们并不知道真实分布的$p(\boldsymbol{Z})$，而只知道近似的后验分布的$q(\boldsymbol{Z})$，为了计算上的可能，我们选择的KL散度是$KL(q||p)$，q在前，p在后表示的是表征q近似p的近似误差。

$$KL(q||p) = -\sum_{Z} q(\boldsymbol{Z})\ln\{\frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\} \tag{14}$$

为了照顾公式14中分母那一项，我们需要对公式13右边做一点变动，即添加类似KL散度分布这一项。并重新给公式左边定义，于是有：

$$\mathcal{L}(q,\theta)=-\sum_{Z}q(\boldsymbol{Z})\ln\{\frac{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\} \tag{15}$$

这样就有了：

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta})=\mathcal{L}(q,\theta)+KL(q||p) \tag{16}$$

公式左边是待求的似然函数，右边第一项是近似的似然函数，右边第二项是近似的误差衡量。

由于$KL(q||p)\geq 0$，当且仅当$q(\boldsymbol{Z})=p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$时取等号。故$\ln p(\boldsymbol{X}|\boldsymbol{\theta})\geq\mathcal{L}(q,\boldsymbol{\theta})$，故近似的似然函数值是真实的似然函数的下界。在M步骤，更新参数$\boldsymbol{\theta}$，极大化近似的似然函数，最大化近似的似然函数同时最大化真实的函数的下界。在E步骤，是保持参数$\boldsymbol{\theta}$不动，使得KL散度为零，使得近似的似然函数等于真实的似然函数。简而言之，在E步骤，在现有参数上，近似的似然函数追上了真实的似然函数，在M步骤，更新参数，近似的似然函数以为自己强大了，实际上真实的似然函数更强大了。

大概就是一个兢兢业业的学霸和学神都考了一百分，学神考了一百分是因为满分只有一百分。