高斯混合模型（GMM）

据上次博客已经2周多了，一直没写，惭愧。

一、高斯模型简介

首先介绍一下单高斯模型(GSM)和高斯混合模型(GMM)的大概思想。

1.单高斯模型

如题，就是单个高斯分布模型or正态分布模型。想必大家都知道正态分布，这一分布反映了自然界普遍存在的有关变量的一种统计规律，例如身高，考试成绩等；而且有很好的数学性质，具有各阶导数，变量频数分布由μ、σ完全决定等等，在许多领域得到广泛应用。在这里简单介绍下高斯分布的概率密度分布函数:

$$\phi(y \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

其中θ=(μ,σ^2);

2.高斯混合模型

注：在介绍GMM的时候，注意跟K-means的相似点

K个GSM混合成一个GMM，每个GSM称为GMM的一个component，也就是分为K个类，与K-means一样，K的取值需要事先确定，具体的形式化定义如下：

$$P(y \mid \theta) = \sum_{k=1}^{K} \alpha_k \phi(y \mid \theta_k) \qquad (2)$$

其中，α_k是样本集合中k类被选中的概率：$\alpha_k = P(z = k \mid \theta)$，其中z=k指的是样本属于k类，那么$\phi(y \mid \theta_k)$可以表示为$\phi(y \mid \theta_k) = P(y \mid z = k, \theta)$,很显然$\alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1$，y是观测数据。

这里如果我们事先知道每个样本的分类情况，那么求解GMM的参数非常直观，如下表示：

假设 有K个类，样本数量分别为N_1,N_2,…,N_k且N_1+N_2+…+N_k=N，即有观测数据$y_1, y_2, \ldots y_N$，第k个分类的样本集合表示为S(k)，那么公式（2）中的三个参

数可以表示为：

$$\alpha_k = N_k/N \qquad (3)$$

$$\mu_k = \frac{1}{N_k}\sum_{y\in S(k)} y \qquad (4)$$

$$\sigma_k^2 = \frac{1}{N_k}\sum_{y\in S(k)}(y-\mu_k)^2 \qquad (5)$$

这样是理想情况，例如给你一堆人类的身高的数据，以及对应的性别，那么这个就是估计两个分量的高斯混合模型，需要学习至少5个参数（事实是6个，另外一个可以有1-α得出）。但是如果给你的人类身高的数据，为给出相应的性别的数据，这样进行学习的话就是一个聚类问题，同样可以知道需要聚成两类（注：许多时候连K也是需要事先假设的），直观上可以按照跟K-means算法以致的思路，只是这里的属于某个类是一个概率，而不是一定的。

首先可以先假设聚成K类，然后选择参数的初始值θ_0（总共2K个变量），这里需要引进一个变量γ_{jk}，表示的是第j个观测来自第k个component的概率，即数据j由第k个Component 生成的概率，或者可以说是这个component上生成这个数据的概率，可以根据后验概率计算得到：

$$\gamma_{jk} = P(z=k \mid y_j,\theta) \qquad (6)$$

$$= \frac{P(z=k,y_j\mid\theta)}{\sum_{k=1}^{K}P(z=k,y_j\mid\theta)} \qquad (7)$$

$$= \frac{P(y_j\mid z=k,\theta)P(z=k\mid\theta)}{\sum_{k=1}^{K}P(y_j\mid z=k,\theta)P(z=k\mid\theta)} \qquad (8)$$

$$= \frac{\alpha_k\phi(y_j\mid\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j\mid\theta_k)} \qquad (9)$$

注：这个与α_k的区别，α_k指的是第k个component被选中的概率，需要γ_{jk}对所有的数据j进行累加

公式(6)=>(7)=>(8)=>(9)分别使用了贝叶斯估计，全概率公式以及α_k和$\phi(y_j\mid\theta_k)$的定义就可得出。

上面是根据数据j计算各个component的生成概率，而现在根据每个component生成了1,2,...N点数据，每个component有事一个高斯分布，那么根据α，μ，σ^2的定义又可以直观地得出如下式子：

$$\alpha_k = \frac{\sum_{j=1}^{N}\gamma_{jk}}{N} \qquad (10)$$

$$\mu_k = \frac{\sum_{j=1}^{N}\gamma_{jk}y_j}{\sum_{j=1}^{N}\gamma_{jk}} \qquad (11)$$

$$\sigma_k^2 = \frac{\sum_{j=1}^{N}\gamma_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\gamma_{jk}} \qquad (12)$$

这样其实只是把原本样本一定属于某一类改成了一个样本属于某类的概率，而k类样本数量N_k变成了概率相加，$N_k = \sum_{j=1}^{N}\gamma_{jk}$，就可以直接得出(10),(11),(12)的公式。

不知不觉就把EM算法的两步迭代给写完了，即将公式(9)和公式(10),(11),(12)进行相互迭代，直到收敛，高斯混合模型就聚类好了。

下面给出较为清晰的训练高斯混合模型的算法步骤：

算法1

选取初始值θ^0初始化θ，

repeat{

(1)估计每个数据的每个component生成概率，即γ_{jk}：

$$\gamma_{jk} = \frac{\alpha_k \phi(y_j \mid \theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k)}$$

(2)根据γ_{jk}，估计每个component的参数，得：

公式(10),(11),(12)

}直到收敛

这样，高斯混合模型已经介绍完，当然上面只是直观介绍，具体的与EM算法思想对应关系说明会放在后面一节。

再算法推导之前，我们先看下K-means和高斯混合模型的异同点。

3.高斯混合模型与K-means异同点

相同点：（1）需要指定K值

（2）需要指定初始值，例如K-means的中心点，GMM的各个参数

（3）都是含有EM算法思想

不同点：（1）优化目标函数不同，K-means:最短距离，GMM：最大化log似然估计

（2）E步的指标不同，K-means:点到中心的距离（硬指标），GMM：求解每个观测数据　　的每个component的概率（软指标）

二、高斯混合模型参数估计说明（EM算法）

下面将EM算法与高斯混合模型的参数估计对应起来，如果不清楚地或者已经忘了部分内容的，可以参照上篇博客EM算法学习。

1.明确影变量，写出完全数据的对数似然函数

从上节，我们可以看出，就可以作为隐变量，那么完全数据的对数似然函数为：

$$
\begin{aligned}
l(\theta) &= \sum_{j=1}^{N} logP(y_j \mid \theta) \\
&= \sum_{j=1}^{N} log \sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k) \\
&= \sum_{j=1}^{N} log \sum_{k=1}^{K} \gamma_{jk} \frac{\alpha_k \phi(y_j \mid \theta_k)}{\gamma_{jk}} \\
&\geq \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} log \frac{\alpha_k \phi(y_j \mid \theta_k)}{\gamma_{jk}} \qquad (13)
\end{aligned}
$$

可以发现这就是很简单的EM算法利用Jensen不等式的推导。

2.EM算法E步

见上篇博客EM算法学习，这部分就是求解期望H函数，也就是求解隐含参数概率γ_{jk}即可，那么根据Jensen不等式，等式成立的约束条件，即可得出公式(7),(8),(9)，这里不做多复述。

在李航老师的《统计学习方法》中，给了γ_{jk}比较正规的说明，即在当前模型参数下第j个观测数据来自第k个分模型的概率，称为分模型k对观测数据y_j的响应度。

3.EM算法M步

根据EM算法M步，可得，这里就是在知道γ_{jk}的情况下，求解θ使得$\sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} log \frac{\alpha_k \phi(y_j|\theta_k)}{\gamma_{jk}}$取到最大值，而这个式子中log上没有叠加式子，就可以求偏导为零，求得参数值。

这里所遇到的问题，怎样详细地推导出公式(10),(11),(12)。

首先将公式(1)代入(13)，得:

$$
\begin{aligned}
H(\theta, \theta^t) &= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} log \frac{\alpha_k \phi(y_j \mid \theta_k)}{\gamma_{jk}} \\
&= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} log \frac{\frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}{\gamma_{jk}} \\
&= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} \left[\log \alpha_k + \log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{(y-\mu)^2}{2\sigma_k^2} - \log \gamma_{jk}\right] \qquad (14
\end{aligned}
$$

那么我们只要对公式（14）分别对μ，σ^2偏导为零即可，而α在$\alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1$的条件下求偏导为0，需要使用拉格朗日定理

具体推导如下：

- 公式（10）

令$L(\theta)=H(\theta,\theta^t)+\lambda(\sum_{k=1}^{K}\alpha_k-1)$，那么对L(θ)分别对α和λ求偏导为零，得(这里谢谢网友@紫梦lan提醒，有个公式推导小错误，现在已经修改)：

$$\because \frac{\partial L(\theta)}{\partial \alpha_k}=\sum_{j=1}^{N}\gamma_{jk}/\alpha_k-\lambda=0$$

$$\therefore \alpha_k=\frac{\sum_{j=1}^{N}\gamma_{jk}}{\lambda} \qquad (15)$$

$$\because \frac{\partial L(\theta)}{\partial \lambda}=\sum_{k=1}^{K}\alpha_k-1=0 \qquad (16)$$

$$take\ (15)\ into\ (16)$$

$$\therefore \frac{\sum_{k=1}^{K}\sum_{j=1}^{N}\gamma_{jk}}{\lambda}-1=\frac{\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}}{\lambda}-1=0$$

$$According\ to\ the\ definition\ of\ \gamma_{jk}:$$

$$\gamma_{jk}=\frac{\alpha_k\phi(y_j\mid\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j\mid\theta_k)}$$

$$\therefore \sum_{k=1}^{K}\gamma_{jk}=1 \quad \therefore \frac{\sum_{j=1}^{N}1}{\lambda}-1=0$$

$$\therefore \lambda=N. \quad \therefore \alpha_k=\frac{\sum_{j=1}^{N}\gamma_{jk}}{N}$$

(我这边latex公式中不能输入中文，所以上式推导用英文说明了，抱歉，应该推导清楚了。)

- 公式（11）

$$\because \frac{\partial H(\theta,\theta^t)}{\partial \mu_k}=\sum_{j=1}^{N}\gamma_{jk}\frac{y_k-\mu_k}{\sigma_k^2}=0$$

$$\therefore \sum_{j=1}^{N}\gamma_{jk}(y_k-\mu_k)=0$$

$$\therefore \sum_{j=1}^{N}\gamma_{jk}y_k=\sum_{j=1}^{N}\gamma_{jk}\mu_k$$

$$\therefore \mu_k=\frac{\sum_{j=1}^{N}\gamma_{jk}y_k}{\sum_{j=1}^{N}\gamma_{jk}}$$

- 公式（12）

对σ^2求偏导为零得到的σ^2值跟对σ求偏导为零得到的σ^2值是一样的，所以这里我对σ求偏导为零得到的σ^2

$$\because \frac{\partial H(\theta,\theta^t)}{\partial \sigma_k} = \sum_{j=1}^{N} \gamma_{jk} \left[-\frac{1}{\sigma_k} + \frac{(y_k-\mu_k)^2}{\sigma_k^3} \right] = 0$$

. $two\ sides \times \sigma_k^3\ so$

. $$\sum_{j=1}^{N} \gamma_{jk} \left[-\sigma_k^2 + (y_k-\mu_k)^2 \right] = 0$$

. $$\therefore \sigma_k^2 = \frac{\sum_{j=1}^{N} \gamma_{jk}(y_k-\mu_k)^2}{\sum_{j=1}^{N} \gamma_{jk}}$$

到现在公式(10),(11),(12)都已推出，其实这部分的推导只是简单的应用了最大似然估计得出。

三、总结

其他的混合模型，例如朴素贝叶斯混合模型都可以使用EM算法推出使用，这里不再次推导。我个人觉得EM算法就是相互迭代，求出一个稳定值，而这种相互迭代的方法用的范围挺广的，例如混合模型，k-means,HMM中的Baum-welch算法等。

思考：相互迭代，比较试用的场景：某个事件大概可以分为两部分，一个是整一个类发生的概率，一个每个小个体发生的概率，而这个两个部分可以相互求解计算，迭代到一个稳定值或者类似于hits算法一样。例：记得去年WWW会议中有以一篇恶意评论用户的查找，就是给每个用户一个概率，给每个恶意评论组(包括几个一起恶意评论的用户）一个概率，而这两个概率就可以相互迭代达到稳定值，得到的结果也较好。是否可以利用相同思想应用到其他场景。

四、主要参考资料

[1]李航.统计学习方法.北京：清华大学出版社，2012

[2]http://www.cnblogs.com/zhangchaoyang/articles/2624882.html

[3]http://blog.pluskid.org/?p=39, pluskid的

[4]http://www.cnblogs.com/jerrylead/archive/2011/04/06/2006936.html JerryLead

[5]http://cs229.stanford.edu/materials.html, Andrew NG教授的讲义

分类: 统计学习方法

标签: EM，GMM