高斯混合模型

Gaussian Mixture Model

① 介绍

② 极大似然估计

③ EM算法

介绍（Introduction）

K均值聚类：

构建的目标函数：$J=\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n-\mu_k\|^2$

其中 r_{nk} 表示 x_n 属于哪一类，μ_k 表示该类的聚类中心。

显然，目标就是最小化 J。

方法也很简单，先固定 μ_k，更新 r_{nk}；再固定 r_{nk}，更新 μ_k。类似于EM.

有时候 $\|x_n-\mu_k\|^2$ 这种度量会失效，故将其看作一类函数 $\gamma(x_n,\mu_k)$

则引申为 $J=\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\gamma(x_n,\mu_k)$.

同样，现在我们将GMM看作是简单的线性叠加：

$$p(x)=\sum_{k=1}^{K}\pi_k N(x\mid\mu_k,\Sigma_k),\quad \sum_{k=1}^{K}\pi_k=1$$

下面，我们从概率角度看GMM

我们引入一个隐变量 z，z 有 K 个分量，将其看作离散的随机变量。

每个分量取值为0或1，且 $\sum_{k=1}^{K} z^{(k)}=1$.

分布列如下：

z	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$\cdots$	$z^{(k)}$
p	π_1	π_2	π_3	$\cdots$	π_k

即 $p(z^{(k)}=1)=\pi_k$

当 $z^{(k)}=1$ 时，表明当前的数据点属于第 k 个高斯分布，和K均值聚类中的 r_{nk} 类似。

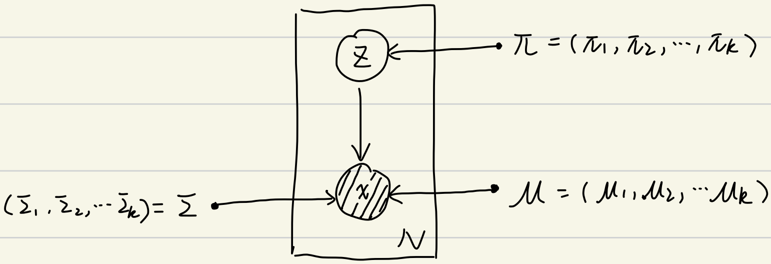
则 $p(z)=\prod_{k=1}^{K}\pi_k^{z^{(k)}}$ $\quad p(x\mid z^{(k)}=1)=N(x\mid\mu_k,\Sigma_k)$

即 $p(x\mid z)=\prod_{k=1}^{K}N(x\mid\mu_k,\Sigma_k)^{z^{(k)}}$

$$p(x,z)=p(z)p(x\mid z)=\prod_{k=1}^{K}\pi_k^{z^{(k)}}N(x\mid\mu_k,\Sigma_k)^{z^{(k)}}$$

则生成过程可以表示为 先从 $p(z)$ 中采样得到 $z=z^{(k)}$.
再从 $p(x|z=z^{(k)})$ 中采样,得到 x.

一组N个独立同分布数据点,的概率图模型表示方法如下.



阴影表示观测变量
实心点表示参数.

同时,由联合概率可以求得 x 的边缘概率.

$$p(x)=\sum_z p(x,z)=\sum_{k=1}^{K}p(x,z=z^{(k)})$$

$$=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k).$$

和线性叠加的表达方式一样.

极大似然估计（MLE）

为了估计参数 π, μ, Σ，我们用MLE.

X: observed data

(X, Z): complete data

θ: π, μ, Σ

$$\log P(X|\theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

我将从下面3个方面来说明直接应用MLE的不可行性.

① 奇异值（singularities）

假设K个高斯分布中，某个高斯分布的 $\Sigma_k = \sigma_k^2 I$，（一般的协方差矩阵也会出现下述现象，但为了方便计算与理解，这里取 $\Sigma_k = \sigma_k^2 I$）

当出现某个数据点，$x_n = \mu_k$，则高斯分布会退化为

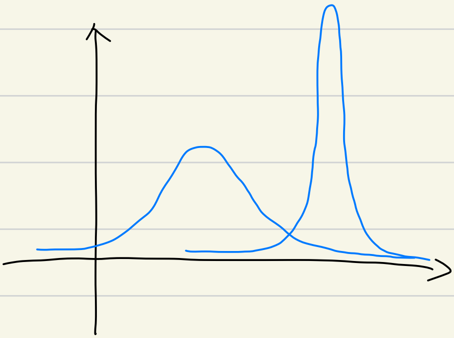
$$N(x_n|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \frac{1}{\sigma_j^D}$$，D为高斯分布的维度.

我们注意到，对于该高斯分布，当 $\sigma_j \to 0$ 时，$N(x_n|\mu_k, \Sigma_k) \to \infty$.

这就导致了 log-likelyhood $\to \infty$

从而，在K个分量中，其中一个分量趋近于无穷，其余的是一个有限的值，导致求解错误.

那为何在单一的高斯分布上不会出现该情况呢？



如上，两幅图中，左边是GMM，右边是单一高斯分布。可以看出，在GMM中，两个分量之间是加权求和的操作，也就是说，如果一个数据点导致一个分量为无穷，其他数据点仍然会有一个有限的值，这就导致了总的likelyhood趋近于无穷。

在单一的高斯分布中，若出现x_n，使得当$\sigma\to0$时，当前数据点的似然趋近于无穷，但注意$p(x)=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma^{D}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$。对于其他数据点，e的指数项中$x-\mu\neq0$，也就导致了，当$\sigma\to0$时，其他数据点的似然会以指数速度趋向于0，从而使得整体的likelyhood不会趋近于无穷。

② Indentifiability.

由于总的概率是由K个分量混合而成的，这就导致了，对于任意给定的最大似然解（有K个分量），我们可以对这K个分量进行全排列，则有K!种组合方式，也就是说有K!种解都能得到该似然，这也就违反了统计中 indentifiability 的概念，这个概念在表示模型参数时，是非常重要的。（PRML中介绍的，但我不理解的是K!种解只是换了下位置，为何会对模型参数的表示产生问题）.

③ 解析解.

我们注意到 $\log P(X|\theta)$ 中，log中有对k的求和，从而log不直接作用于高斯分布，当令其导数为0时，无法得到解析解。
（视频和PRML中都下了此结论，但我实际还没算过）

EM算法

EM: $\theta^{(t+1)} = \arg\max\limits_{\theta} E_{z\sim p(z|x,\theta^{(t)})}[\log p(x,z|\theta)]$

注意，K均值算法常被用于EM之前的初始化高斯混合模型的参数。

E-step:

$$Q(\theta,\theta^{(t)}) = E_{z\sim p(z|x,\theta^{(t)})}[\log p(x,z|\theta)]$$

$$= \sum_{k=1}^{K}\left[\sum_{i=1}^{N} p(z_i^{(k)}=1|x_i,\theta^{(t)}) \log p(x_i, z_i^{(k)}=1|\theta)\right]$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{K}\pi_j N(x_i|\mu_j^{(t)},\Sigma_j^{(t)})} \log \pi_k N(x_i|\mu_k,\Sigma_k)$$

M-step:

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{k=1}^{K}\sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{K}\pi_j N(x_i|\mu_j^{(t)},\Sigma_j^{(t)})} \log \pi_k N(x_i|\mu_k,\Sigma_k)$$

以求解π为例.

$$\begin{cases} \pi^{(t+1)} = \arg\max\limits_{\pi} \sum_{k=1}^{K}\sum_{j=1}^{N} \dfrac{\pi_k N(x_i|\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{K}\pi_j N(x_i|\mu_j^{(t)},\Sigma_j^{(t)})} \log \pi_k . \\ s.t. \sum_{k=1}^{K}\pi_k = 1 \end{cases}$$

拉格朗日函数: $\mathcal{L}(\pi,\lambda) = \sum_{k=1}^{K}\sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{K}\pi_j N(x_i|\mu_j^{(t)},\Sigma_j^{(t)})} \log \pi_k + \lambda\left(1-\sum_{k=1}^{K}\pi_k\right)$

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})} \cdot \frac{1}{\pi_k} - \lambda = 0$$

$$\Rightarrow \quad \sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})} - \lambda \pi_k = 0.$$

对π_1到π_K都求偏导，得到K个等式之后，将其相加，得.

$$\underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\pi_k N(x_i|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})}} - \lambda \underbrace{\sum_{k=1}^{K} \pi_k}_{1\,(\text{条件})} = 0.$$

注意：这个分数是$p(z_i^{(k)}=1|x_i, \theta^{(t)})$，求和后为1.（直接算也是1，利用概率的性质也是1）

$$\Rightarrow \quad N - \lambda = 0 \quad \Rightarrow \lambda = N$$

$$\Rightarrow \quad \pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})}$$

另外两个参数结论如下：

$$\begin{cases} \mu_k^{(t+1)} = \dfrac{1}{N_k^{(t)}} \displaystyle\sum_{n=1}^{N} \gamma^{(t)}(z_n^{(k)}) x_n \\[2ex] \Sigma_k^{(t+1)} = \dfrac{1}{N_k^{(t)}} \displaystyle\sum_{n=1}^{N} \gamma^{(t)}(z_n^{(k)}) (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T \end{cases}$$

其中，$\gamma^{(t)}(z_n^{(k)}) = \dfrac{\pi_k N(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j^{(t)}, \Sigma_j^{(t)})}$，$N_k^{(t)} = \sum_{i=1}^{N} \gamma^{(t)}(z_i^{(k)})$

（结论出自于PRML，有空也可以看一下PRML上面有关这一节的介绍）.