GMM and EM Algorithm

- GMM and EM Algorithm
  - 极大似然估计
  - GMM 混合模型（gaussian mixture model）
  - MLE 参数估计遇到的问题
  - EM 算法（estimate-maximum）

极大似然估计

对于N分类问题 $C_i \in C = \{C_1, C_2, ...\}$，其极大似然分类器为：

$$h^*(X_j) = \max_{C_i \in C} P(X_j | C_i)$$

在参数估计背景下，极大似然参数估计为：

$$\hat{\theta}_{ML} = \max_{\theta} P(\vec{X}|\theta)$$

$$L(\theta) \stackrel{def}{=} \log P(\vec{X}|\theta)$$

下面推导单高斯模型下的极大似然估计。假设一个d维高斯分布，其N样本的概率分布及其对应的对数似然函数为：

$$P(X_i) = |(2\pi)^d \Sigma|^{-\frac{1}{2}} exp(-\frac{1}{2}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu))$$

$$L(\mu, \Sigma) = K + -\frac{N}{2} log|\Sigma| + \sum_{i=1}^{N} -\frac{1}{2}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu)$$

先求解 μ 的最优估计：

$$\frac{\partial L}{\partial \mu} = \frac{\partial \left[\sum_i (X_i - \mu)^T \Sigma^{-1}(X_i - \mu)\right]}{\partial \mu}$$
$$= \frac{\partial \left[\sum_i (X_i^T \Sigma^{-1} X_i - X_i^T \Sigma^{-1} \mu - \mu^T \Sigma^{-1} X_i + \mu^T \Sigma^{-1} \mu)\right]}{\partial \mu}$$
$$= \sum_i -\Sigma^{-1T} X_i - \Sigma^{-1} X_i + (\Sigma^{-1T} \mu + \Sigma^{-1} \mu)$$

令上式在任意 Σ 时取0，得到：

$$\mu = \frac{1}{N} \sum_i X_i$$

下面求解协方差矩阵的估计值：

$$\frac{\partial L}{\partial \Sigma}=\frac{\partial\left[K+-\frac{N}{2}log|\Sigma|+\sum_{i=1}^{N}-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right]}{\partial \Sigma}$$
$$=\frac{\partial\left[-\frac{N}{2}log|\Sigma|+\sum_{i=1}^{N}-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right]}{\partial \Sigma}$$

其中第一项为：

$$\frac{\partial\left[-\frac{N}{2}log|\Sigma|\right]}{\partial \Sigma}=-\frac{1}{2}\frac{\partial\left[log|\Sigma|\right]}{\partial|\Sigma|}\frac{\partial|\Sigma|}{\partial\Sigma}$$
$$=-\frac{N}{2}\frac{1}{|\Sigma|}\Sigma^{*}=-\frac{N}{2}\Sigma^{-1}$$
$$(Hint:\nabla\log\det(\Theta)=\Theta^{-1})$$

第二项为：

$$\frac{\partial\left[\sum_{i=1}^{N}-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right]}{\partial \Sigma}=\sum_{i=1}^{N}-\frac{1}{2}\frac{\partial(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)}{\partial\Sigma}$$
$$=\sum_{i=1}^{N}-\frac{1}{2}\frac{\partial(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)}{\partial\Sigma}$$
$$=\sum_{i=1}^{N}-\frac{1}{2}\left[(X_i-\mu)^T(X_i-\mu)\right]^{-1}$$

综上，令 $\frac{\partial L}{\partial\Sigma}=0$ 后，得：

$$-\frac{N}{2}\Sigma^{-1}+\sum_{i=1}^{N}-\frac{1}{2}\left[(X_i-\mu)(X_i-\mu)^T\right]^{-1}=0$$
$$\Sigma=\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(X_i-\mu)^T$$

GMM 混合模型（gaussian mixture model）

高斯混合模型是一种生成模型，是多个高斯模型 $C_1,C_2,C_3...$ 构成的（ $N(X|\mu_k,\Sigma_k)$ 代表在对应正态分布中的样本的概率）。它可以看作一个随机变量的值是有 k 的概率由 C_k 对应的高斯分布产生的；当然也可以看作由多个高斯分布拟合成的一个任意的随机变量分布：

$$P(X)=\sum_{k=1}^{K}p_kN(X|\mu_k,\Sigma_k)$$
$$s.t.\sum_{k=1}^{K}p_k=1$$

现在设观察的样本集：$X=(X_1,X_2,...,X_n)$，对应隐变量集：$Z=(Z_1,Z_2,...,Z_n)$，这里的隐变量是用于描述样本附带的某些没有观测到的属性的。例如，在GMM模型里，$Z1$ 表示样本 X_1 是由具体哪一个高斯分布产生的，可以是 $C_1,C_2,C_3...$ 中的任何一个，这是没有观测到的。观测样本的完整集可以表示为：$(X,Z)=((X_1,Z_1),(X_2,Z_2),...,(X_3,Z_3))$

（每个组合完整代表了一个样本，但应当注意样本是有多个特征的，即 X_1,Z_1 都是向量，表示它们取得某个值时用小写代表 $X_1=x_1$）

MLE 参数估计遇到的问题

直接求解上述MLE是不可行的，通过极大似然估计参数列表 $\theta=\{p_1,\mu_1,\Sigma_1,p_2,\mu_2,\Sigma_2...\}$，可以得到：

$$\hat{\theta}_{ML}=arg\max_{\theta}\sum_{i=1}^{N}\log\sum_{k=1}^{K}p_kN(X|\mu_k,\Sigma_k)$$

它的极值点无法得到解析解，因为第二个求和号在 log 内部。

EM 算法（estimate-maximum）

EM算法是一种迭代算法，被用来代替求解上述参数估计问题

$$\theta^{(t+1)}=arg\max_{\theta}E_{z|x,\theta^{(t)}}[\log P(X,Z|\theta)]$$
$$Q(\theta,\theta^{(t)})\overset{def}{=}E_{z|x,\theta^{(t)}}[\log P(X,Z|\theta)]$$

记：

$$\theta^{(t+1)}=arg\max_{\theta}Q(\theta,\theta^{(t)})$$

（E step）下面我们带入GMM模型，得到GMM的迭代公式。简化上述 Q 表达式：

$$\begin{aligned}Q(\theta,\theta^{(t)})&=E_{z|x,\theta^{(t)}}[\log P(X,Z|\theta)]\\&=\sum_{Z}P(Z|X,\theta^{(t)})\log P(X,Z|\theta)\\&=\sum_{Z}\prod_{i=1}^{N}P(Z_i|X_i,\theta^{(t)})\log\prod_{i=1}^{N}P(X_i,Z_i|\theta)\\&=\sum_{Z}\prod_{i=1}^{N}P(Z_i|X_i,\theta^{(t)})\sum_{i=1}^{N}\log P(X_i,Z_i|\theta)\end{aligned}$$

取第二个连加号内的任一项：

$$
\begin{aligned}
R_i &= \sum_{Z} \prod_{i=1}^{N} P(Z_i|X_i,\theta^{(t)}) \cdot \log P(X_1,Z_1|\theta) \\
&= \sum_{Z_1,Z_2,\ldots} \prod_{i=2}^{N} P(Z_i|X_i,\theta^{(t)}) \cdot P(Z_1|X_1,\theta^{(t)}) \log P(X_1,Z_1|\theta) \\
&= \sum_{Z_1} P(Z_1|X_1,\theta^{(t)}) \log P(X_1,Z_1|\theta) \sum_{Z_2,Z_3,\ldots} \prod_{i=2}^{N} P(Z_i|X_i,\theta^{(t)}) \\
&= \sum_{Z_1} P(Z_1|X_1,\theta^{(t)}) \log P(X_1,Z_1|\theta)
\end{aligned}
$$

最后得到：

$$
Q(\theta,\theta^{(t)}) = \sum_{i=1}^{N} \sum_{Z_i} P(Z_i|X_i,\theta^{(t)}) \log P(X_i,Z_i|\theta)
$$

其中，根据高斯混合模型可以很容易得到下式。并且里面的 $p_{Z_i}, \mu_{z_i}, \Sigma_{z_i}$ 都是每次迭代中等待优化的参数，即 $\theta^{(t)} = \{p_1,\mu_1,\Sigma_1,p_2,\mu_2,\Sigma_2\ldots\}$

$$
P(X_i,Z_i|\theta) = p_{z_i} N(X|\mu_{z_i},\Sigma_{z_i})
$$

（M step）下面我们计算 $P(Z_i|X_i,\theta^{(t)})$ ，它是在**给定混合模型参数**下的条件概率。在GMM模型中使用贝叶斯定理：

$$
\begin{gathered}
P(X_i|\theta^{(t)}) = \sum_{k=1}^{K} p_k^{(t)} N(X_i|\mu_k^{(t)},\Sigma_k^{(t)}) \\
P(X_i,Z_i|\theta^{(t)}) = P(Z_i|\theta^{(t)}) P(X_i|Z_i,\theta^{(t)}) = p_{Z_i}^{(t)} N(X_i|\mu_{Z_i}^{(t)},\Sigma_{Z_i}^{(t)}) \\
P(Z_i|X_i,\theta^{(t)}) = \frac{P(X_i,Z_i|\theta^{(t)})}{P(X_i|\theta^{(t)})} = \frac{p_{Z_i}^{(t)} N(X_i|\mu_{Z_i}^{(t)},\Sigma_{Z_i}^{(t)})}{\sum_{k=1}^{K} p_k^{(t)} N(X_i|\mu_k^{(t)},\Sigma_k^{(t)})}
\end{gathered}
$$

将上面得到的两个条件概率带入到 Q 中

$$
\begin{aligned}
Q(\theta,\theta^{(t)}) &= \sum_{i=1}^{N} \sum_{Z_i} P(Z_i|X_i,\theta^{(t)}) \log P(X_i,Z_i|\theta) \\
&= \sum_{i=1}^{N} \sum_{Z_i} \frac{p_{Z_i}^{(t)} N(X_i|\mu_{Z_i}^{(t)},\Sigma_{Z_i}^{(t)})}{\sum_{k=1}^{K} p_k^{(t)} N(X_i|\mu_k^{(t)},\Sigma_k^{(t)})} \log[p_{z_i} N(X|\mu_{z_i},\Sigma_{z_i})]
\end{aligned}
$$

下面开始求解迭代参数的估计值，注意上式中分数的一项是不带任何待预测参数的，即之前被记为 $P(Z_i|X_i,\theta^{(t)})$ 的项。使用拉格朗日乘数法优化约束下的 p_k ，对 μ_{z_i}, Σ_{z_i} 求极值。由带约束的拉格朗日优化得到：

$$
p_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} P(Z_i = C_k|X_i,\theta^{(t)})
$$

其中 $P(Z_i=C_k|X_i,\theta^{(t)})$ 使用贝叶斯定理求解：

$$P(Z_i=C_k|X_i,\theta^{(t)})=\frac{P(Z_i=C_k|\theta^{(t)})P(X_i|Z_i=C_k,\theta^{(t)})}{P(X_i|\theta^{(t)})}$$

以上几个概率都是可求的，$P(X_i|\theta^{(t)})$ 是在当前迭代参数下产生样本的概率，$P(Z_i=C_k|\theta^{(t)})$ 可以直接从当前迭代参数 $\theta^{(t)}$ 中得到，即 $p_k^{(t)}=P(Z_i=C_k|\theta^{(t)})$。$P(X_i|Z_i=C_k,\theta^{(t)})$ 为第k个高斯子分布下 X_i 的分布，由 $\mu_k^{(t)},\Sigma_k^{(t)}$ 确定。

同理，对 Q 求导可以得到均值和协方差的局部最优解。

$$\mu_k^{(t+1)}=\frac{\sum_{i=1}^N X_i P(Z_i=C_k|X_i,\theta^{(t)})}{\sum_{i=1}^N P(Z_i=C_k|X_i,\theta^{(t)})}$$

$$\Sigma_k^{(i+1)}=\frac{\sum_{i=1}^N (X_i-\mu_k^{(i+1)})(X_i-\mu_k^{(i+1)})^T P(Z_i=C_k|X_i,\theta^{(t)})}{\sum_{i=1}^N P(Z_i=C_k|X_i,\theta^{(t)})}$$