混合高斯模型（Mixtures of Gaussians）和 EM 算法

JerryLead
csxulijie@gmail.com

这篇讨论使用期望最大化算法（Expectation-Maximization）来进行密度估计（density estimation）。

与 k-means 一样，给定的训练样本是$\{x^{(1)},\ldots,x^{(m)}\}$，我们将隐含类别标签用$z^{(i)}$表示。与 k-means 的硬指定不同，我们首先认为$z^{(i)}$是满足一定的概率分布的，这里我们认为满足多项式分布，$z^{(i)}\sim\text{Multinomial}(\emptyset)$，其中$p(z^{(i)}=j)=\emptyset_j$, $\emptyset_j\geq 0$, $\sum_{j=1}^{k}\emptyset_j=1$，$z^{(i)}$有 k 个值{1,...,k}可以选取。而且我们认为在给定$z^{(i)}$后，$x^{(i)}$满足多值高斯分布，即$(x^{(i)}|z^{(i)}=j)\sim N(\mu_j,\Sigma_j)$。由此可以得到联合分布$p(x^{(i)},z^{(i)})=p(x^{(i)}|z^{(i)})p(z^{(i)})$。

整个模型简单描述为对于每个样例$x^{(i)}$，我们先从 k 个类别中按多项式分布抽取一个$z^{(i)}$，然后根据$z^{(i)}$所对应的 k 个多值高斯分布中的一个生成样例$x^{(i)}$,。整个过程称作混合高斯模型。注意的是这里的$z^{(i)}$仍然是隐含随机变量。模型中还有三个变量$\emptyset$，μ和Σ。最大似然估计为$p(x,z)$。对数化后如下：

$$\begin{aligned}\ell(\phi,\mu,\Sigma) &= \sum_{i=1}^{m}\log p(x^{(i)};\phi,\mu,\Sigma)\\ &= \sum_{i=1}^{m}\log\sum_{z^{(i)}=1}^{k}p(x^{(i)}|z^{(i)};\mu,\Sigma)p(z^{(i)};\phi).\end{aligned}$$

这个式子的最大值是不能通过前面使用的求导数为 0 的方法解决的，因为求的结果不是 close form。但是假设我们知道了每个样例的$z^{(i)}$，那么上式可以简化为：

$$\ell(\phi,\mu,\Sigma)=\sum_{i=1}^{m}\log p(x^{(i)}|z^{(i)};\mu,\Sigma)+\log p(z^{(i)};\phi).$$

这时候我们再来对$\emptyset$，μ和Σ进行求导得到：

$$\begin{aligned}\phi_j &= \frac{1}{m}\sum_{i=1}^{m}1\{z^{(i)}=j\},\\ \mu_j &= \frac{\sum_{i=1}^{m}1\{z^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m}1\{z^{(i)}=j\}},\\ \Sigma_j &= \frac{\sum_{i=1}^{m}1\{z^{(i)}=j\}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{m}1\{z^{(i)}=j\}}.\end{aligned}$$

$\emptyset_j$就是样本类别中$z^{(i)}=j$的比率。μ_j是类别为 j 的样本特征均值，Σ_j是类别为 j 的样例的特征的协方差矩阵。

实际上，当知道$z^{(i)}$后，最大似然估计就近似于高斯判别分析模型（Gaussian discriminant analysis model）了。所不同的是 GDA 中类别 y 是伯努利分布，而这里的 z 是多项式分布，还有这里的每个样例都有不同的协方差矩阵，而 GDA 中认为只有一个。

之前我们是假设给定了$z^{(i)}$，实际上$z^{(i)}$是不知道的。那么怎么办呢？考虑之前提到的 EM 的思想，第一步是猜测隐含类别变量 z，第二步是更新其他参数，以获得最大的最大似然估计。用到这里就是：

循环下面步骤，直到收敛：｛

（E 步）对于每一个 i 和 j，计算

$$w_j^{(i)} \coloneqq p(z^{(i)} = j | x^{(i)}; \Phi, \mu, \Sigma)$$

（M 步），更新参数：

$$\begin{aligned} \phi_j &:= \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}, \\ \mu_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}, \\ \Sigma_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}} \end{aligned}$$

}

在 E 步中，我们将其他参数Φ, μ, Σ看作常量，计算$z^{(i)}$的后验概率，也就是估计隐含类别变量。估计好后，利用上面的公式重新计算其他参数，计算好后发现最大化最大似然估计时，$w_j^{(i)}$值又不对了，需要重新计算，周而复始，直至收敛。

$w_j^{(i)}$的具体计算公式如下：

$$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$

这个式子利用了贝叶斯公式。

这里我们使用$w_j^{(i)}$代替了前面的$1\{z^{(i)} = j\}$，由简单的 0/1 值变成了概率值。

对比 K-means 可以发现，这里使用了“软”指定，为每个样例分配的类别$z^{(i)}$是有一定的概率的，同时计算量也变大了，每个样例 i 都要计算属于每一个类别 j 的概率。与 K-means 相同的是，结果仍然是局部最优解。对其他参数取不同的初始值进行多次计算不失为一种好方法。

虽然之前再 K-means 中定性描述了 EM 的收敛性，仍然没有定量地给出，还有一般化 EM 的推导过程仍然没有给出。下一篇着重介绍这些内容。