高斯混合模型

2021-8-14

1. 高斯混合模型
   1. 1. 模型定义
   1. 2. 假设的生成过程
   1. 3. 求后验分布
   1. 4. 聚类
2. 参数求解
   2. 1. 极大似然估计
      2. 1. 1. 求μ_i
      2. 1. 2. 求σ_i
      2. 1. 3. 求α_i
      2. 1. 4. 迭代过程
   2. 2. EM算法
      2. 2. 1. 完全数据的对数似然函数
      2. 2. 2. E步：确定Q函数
      2. 2. 3. M步：Q函数最大化
3. 例子
4. 参考文档

1. 高斯混合模型

1.1. 模型定义

$$P_M(x)=\sum_{i=1}^{k}\alpha_i\cdot\phi(x|\mu_i,\sigma_i^2) \tag{24}$$

其中，各符号解释如下：

1. x为观测变量。
2. 该模型共由k个高斯模型混合组成，$\phi(x|\mu_i,\sigma_i^2)=\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$为高斯模型的概率密度函数。
3. α_i为对应的"混合系数"，随件变量x由哪个高斯模型生成的**先验分布**，满足的约束为$\alpha_i>0,\sum_{i=0}^{k}\alpha_i=1$。

1.2. 假设的生成过程

注意：此过程为我们假设的数据生成过程，基于一定的假设，便于后续的推导。

假设样本的生成过程由高斯混合模型生成：

1. 根据定义的**先验分布**$\alpha_1,\alpha_2,\dots,\alpha_k$选择高斯混合成分（即某个高斯混合模型），其中，对任意一个随机变量x，α_i为选择第i个混合成分的概率。
2. 根据被选择的高斯分布的概率密度函数进行采样，从而生成相应的样本。
3. 重复上述步骤，生成训练集$D=\{x_1,x_2,\dots,x_m\}$。

总结：该过程生成了若干数据，数据来源于多个高斯模型，但是，我们不知道高斯模型的参数，也不知道每个数据来自于哪个高斯模型。而我们要做的，便是解决上述两个"不知道"。

1.3. 求后验分布

该步的目标是解决第二个"不知道"：数据来自于哪个高斯模型。

假设目前已经给定``数据集$D=\{x_1,x_2,\dots,x_m\}$，令随机变量$z_j\in\{1,2,\dots,k\}$表示样本x_j由哪个高斯模型生成，其取值未知。基于**假设**，z_j的先验概率为：

$$P(z_j=i)=\alpha_i,\ i\in\{1,2,\dots,k\} \tag{25}$$

根据贝叶斯定理，z_j的后验分布为：

$$\begin{aligned} P_M(z_j=i|x_j) &= \frac{P(z_j=i)P_M(x_j|z_j=i)}{P_M(x_j)} \\ &= \frac{\alpha_i\cdot\phi(x_j|\mu_i,\sigma_i^2)}{\sum_{l=1}^k\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l^2)} \end{aligned} \tag{26}$$

换言之，$P_M(z_j=i|x_j)$给出了样本x_j由第i个高斯模型生成的后验概率分布，简记为$\gamma_{ji}\ (i=1,2,\ldots,k)$。

由全概率公式可得：

$$\sum_{i=1}^k P_M(z_j=i|x_j)=\sum_{i=1}^k\gamma_{ji}=1 \tag{27}$$

1.4. 聚类

当高斯混合模型（公式一）已知时，高斯混合聚类将把样本集D划分为k个簇$C=\{C_1,C_2,\ldots,C_k\}$。每个样本由一个高斯分布生成，来自于同一个高斯分布的样本被认为一类，每个样本x_j的簇标记λ_j如下确定：

$$\underset{i\in\{1,2,\ldots,k\}}{\mathbf{argmax}}\ \gamma_{ji} \tag{28}$$

因此，从原型聚类的角度来看，高斯混合聚类时采用概率模型（高斯分布）对原型进行刻画，簇划分则由原型对应的后验概率确定。

2. 参数求解

2.1. 极大似然估计

给定样本集D，样本个数为m，可使用极大似然法对模型参数$\{(\alpha_i,\mu_i,\sigma_i^2)|1\le i\le k\}$进行估计。对数似然函数为：

$$\begin{aligned} LL(D) &= ln\left(\prod_{j=1}^m P_M(x_j)\right) \\ &= \sum_{j=1}^m ln\left(P_M(x_j)\right) \\ &= \sum_{j=1}^m ln\left(\sum_{l=1}^k\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l^2)\right) \\ &= \sum_{j=1}^m ln\left(\sum_{l=1}^k\alpha_l\cdot\frac{1}{\sqrt{2\pi}\sigma_l}e^{-\frac{(x_j-\mu_l)^2}{2\sigma_l^2}}\right) \end{aligned} \tag{29}$$

2.1.1. 求μ_i

令$\frac{\partial LL(D)}{\partial\mu_i}=0$，则：

$$\begin{aligned} \frac{\partial LL(D)}{\partial\mu_i} &= \frac{\partial\sum_{j=1}^m ln\left(\sum_{l=1}^k\alpha_l\cdot\frac{1}{\sqrt{2\pi}\sigma_l}e^{-\frac{(x_j-\mu_l)^2}{2\sigma_l^2}}\right)}{\partial\mu_i} \\ &= \sum_{j=1}^m\frac{\alpha_i\cdot\frac{1}{\sqrt{2\pi}\sigma_i}\cdot 2(x_j-\mu_i)\cdot e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}}}{\sum_{l=1}^k\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l)} \\ &= \sum_{j=1}^m\frac{\alpha_i\cdot 2(x_j-\mu_i)\cdot\phi(x_j|\mu_i,\sigma_i)}{\sum_{l=1}^k\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l)} \end{aligned} \tag{30}$$

$$
\begin{aligned}
&= 2\sum_{j=1}^{m}\gamma_{ji}\cdot(x_j-\mu_i) \\
&= 2\sum_{j=1}^{m}\gamma_{ji}x_j - 2\sum_{j=1}^{m}\gamma_{ji}\mu_i \\
&= 0
\end{aligned}
$$

求解可得：

$$
\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{m}\gamma_{ji}\boldsymbol{x}_j}{\sum_{j=1}^{m}\gamma_{ji}} \tag{31}
$$

2.1.2. 求σ_i

令$\frac{\partial LL(D)}{\partial\sigma_i}=0$，则：

$$
\begin{aligned}
\frac{\partial LL(D)}{\partial\sigma_i} &= \frac{\partial\sum_{j=1}^{m} ln\left(\sum_{l=1}^{k}\alpha_l\cdot\frac{1}{\sqrt{2\pi}\sigma_l}e^{-\frac{(x_j-\mu_l)^2}{2\sigma_l^2}}\right)}{\partial\sigma_i} \\
&= \sum_{j=1}^{m}\frac{\alpha_i\cdot\frac{1}{\sqrt{2\pi}}\left[-1\cdot\frac{1}{\sigma_i^2}\cdot e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}}+\frac{1}{\sigma_i}\cdot(x_j-\mu_i)^2\cdot\frac{1}{\sigma_i^3}\cdot e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}}\right]}{\sum_{l=1}^{k}\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l)} \\
&= \sum_{j=1}^{m}\frac{\alpha_i\cdot\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}}\cdot\left[\frac{-1}{\sigma_i^2}+\frac{1}{\sigma_i^4}(x_j-\mu_i)^2\right]}{\sum_{l=1}^{k}\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l)} \\
&= \sum_{j=1}^{m}\alpha_i\cdot\frac{1}{\sqrt{2\pi}}\cdot\gamma_{ji}\cdot\frac{(x_j-\mu_i)^2-\sigma_i^2}{\sigma_i^4} \\
&= 0
\end{aligned} \tag{32}
$$

求解可得：

$$
\boldsymbol{\sigma}_i^2 = \frac{\sum_{j=1}^{m}\gamma_{ji}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^2}{\sum_{j=1}^{m}\gamma_{ji}} \tag{33}
$$

2.1.3. 求α_i

对于混合系数α_i，除了要最大化$LL(D)$，还需要满足$\alpha_i>0$，$\sum_{i=0}^{k}\alpha_i=1$，构造如下拉格朗日函数：

$$
\mathcal{L} = LL(D)+\lambda\left(\sum_{l=1}^{k}\alpha_l-1\right) \tag{34}
$$

其中，λ为拉格朗日乘子

对α_i求导可得：

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\alpha_i} &= \frac{\partial\left\{\sum_{j=1}^{m} ln\left(\sum_{l=1}^{k}\alpha_l\cdot\frac{1}{\sqrt{2\pi}\sigma_l}e^{-\frac{(x_j-\mu_l)^2}{2\sigma_l^2}}\right)+\lambda\left(\sum_{l=1}^{k}\alpha_l-1\right)\right\}}{\partial\alpha_i} \\
&= \sum_{j=1}^{m}\frac{\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}}}{\sum_{l=1}^{k}\alpha_l\cdot\phi(x_j|\mu_l,\sigma_l^2)}+\lambda
\end{aligned} \tag{35}
$$

$$
\begin{aligned}
&= \sum_{j=1}^{m} \frac{\phi(x_j|u_i,\sigma_i^2)}{\sum_{l=1}^{k}\alpha_l \cdot \phi(x_j|\mu_l,\sigma_l^2)} + \lambda \\
&= 0
\end{aligned}
$$

两边同乘以α_i可得：

$$
\begin{aligned}
&\sum_{j=1}^{m} \frac{\alpha_i \cdot \phi(x_j|u_i,\sigma_i^2)}{\sum_{l=1}^{k}\alpha_l \cdot \Leftrightarrow \phi(x_j|\mu_l,\sigma_l^2)} + \alpha_i\lambda = 0 \\
&\Rightarrow \sum_{j=1}^{m}\gamma_{ji} + \alpha_i\lambda = 0 \\
&\Rightarrow \sum_{l=1}^{k}\sum_{j=1}^{m}\gamma_{ji} + \sum_{l=1}^{k}\alpha_i\lambda = 0 \\
&\Rightarrow \sum_{j=1}^{m}\sum_{l=1}^{k}\gamma_{ji} + \lambda = 0 \\
&\Rightarrow m + \lambda = 0 \\
&\Rightarrow \lambda = -m
\end{aligned} \tag{36}
$$

将λ带入可得：$\sum_{j=1}^{m}\gamma_{ji} - \alpha_i m = 0$

求解可得：

$$
\alpha_i = \frac{1}{m}\sum_{j=1}^{m}\gamma_{ji} \tag{37}
$$

2.1.4. 迭代过程

1. 先根据当前参数计算每个样本属于每个高斯模型的后验概率γ_{ji}
2. 根据参数公式对参数进行更新

2.2. EM算法

2.2.1. 完全数据的对数似然函数

首先需要明确**隐变量**，写出完全数据的对数似然函数。

定义隐变量如下：

$$
\gamma_{jk} = \begin{cases} 1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个分模型} \\ 0, & \text{否则} \end{cases} \tag{38}
$$

那么，完全数据为：$(x_j,\gamma_{j1},\gamma_{j2},\ldots,\gamma_{jk}), \quad j=1,2,\ldots,m$

令$\theta=(\mu,\sigma^2)$，则**完全数据的似然函数**为：

$$
\begin{aligned}
P(x,\gamma \mid \alpha,\theta) &= \prod_{j=1}^{m} P\left(x_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jk} \mid \alpha,\theta\right) \\
&= \prod_{j=1}^{m}\prod_{l=1}^{k}\left[\alpha_l\phi\left(x_j \mid \theta_l\right)\right]^{\gamma_{jl}} \\
&= \prod_{l=1}^{k}\prod_{j=1}^{m}\left[\alpha_l\phi\left(x_j \mid \theta_l\right)\right]^{\gamma_{jl}} \\
&= \prod_{l=1}^{k}\alpha_l^{\sum_{j=1}^{m}\gamma_{jl}}\prod_{j=1}^{m}\left[\phi\left(x_j \mid \theta_l\right)\right]^{\gamma_{jl}} \\
&= \prod_{l=1}^{k}\alpha_l^{\sum_{j=1}^{m}\gamma_{jl}}\prod_{j=1}^{m}\left[\frac{1}{\sqrt{\ }}\exp\left(-\frac{(x_j-\mu_l)^2}{2\sigma^2}\right)\right]^{\gamma_{jl}}
\end{aligned} \tag{39}
$$

$$\prod_{l=1} \quad \prod_{j=1}\left[\frac{1}{\sqrt{2\pi}\sigma_l} \left(\frac{}{2{\sigma_l}^2} \right)\right]$$

上述公式解释：

1. $P\left(x_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jk} \mid \alpha, \theta\right)$

   表示在参数θ下选择某个高斯模型(假设第s个)生成数据x_j的概率，则根据先验知识，该过程分为两个步骤：

   - 依概率α_s选择第s个高斯模型
   - 通过$\phi(x|\theta_s)$生成x_j

   所以，综合来看，如果不知道x_j由哪个模型生成，则该概率为：

$$P\left(x_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jk} \mid \alpha, \theta\right)=\prod_{l=1}^{k}\left[\alpha_l \cdot \phi(x_j|\theta)\right]^{\gamma_{jl}} \tag{40}$$

   由哪个模型生成，则对应的λ为1。

2. $P\left(x_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jk} \mid \alpha, \theta\right)$与$P(x_j)$的区别

   $P(x_j)=\sum_{i=1}^{k}\alpha_i \cdot \phi(x_j|\mu_i, \sigma_i^2)$需要考虑$x_j$可能来源于$k$个高斯模型中的任何一个，而$P\left(x_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jk} \mid \alpha, \theta\right)$则限定了$x_j$只会来源于某一个高斯模型。

3. $\sum_{l=1}^{k}\sum_{j=1}^{m}\gamma_{jl}=1$

完全数据的对数似然函数为：

$$\begin{aligned}\ln P(x, \gamma \mid \alpha, \theta) &= \ln \left\{\prod_{l=1}^{k}\alpha_l^{\sum_{j=1}^{m}\gamma_{jl}}\prod_{j=1}^{m}\left[\frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{(x_j-\mu_l)^2}{2{\sigma_l}^2}\right)\right]^{\gamma_{jl}}\right\} \\ &= \sum_{l=1}^{k}\left\{\sum_{j=1}^{m}\gamma_{jl}\cdot\ln\alpha_l+\sum_{j=1}^{m}\gamma_{jl}\left[\ln\frac{1}{\sqrt{2\pi}}-\ln\sigma_l-\frac{(x_j-\mu_l)^2}{2\sigma_l^2}\right]\right\}\end{aligned} \tag{41}$$

此似然函数中包含隐变量γ_{jl}，因此，不能够直接最大化似然函数。反之，如果知道了γ_{jl}，则最大化似然函数水到渠成。

怎么办呢？基于当前的模型参数(α, θ)，先对隐变量γ_{jl}进行当下的估计。

初始化的时候可以随机给定$(\alpha^{(0)}, \theta^{(0)})$，也就是当前知道了每个高斯分布的参数，而隐变量γ则是指明数据由哪个高斯模型产生，有了每个高斯模型的参数，自然就能够计算数据最可能来自哪个高斯模型，也就知道了隐变量γ。将上述步骤迭代进行之，则是EM算法。

2.2.2. E步：确定Q函数

$$\begin{aligned}&Q\left(\theta, \alpha^{(i)}, \theta^{(i)}\right) \\ &= E\left[\log P(x, \gamma \mid \alpha, \theta) \mid x, \alpha^{(i)}, \theta^{(i)}\right] \\ &= E\left\{\sum_{l=1}^{k}\left\{\left(\sum_{j=1}^{m}\gamma_{jl}|x_j, \alpha^{(i)}, \theta^{(i)}\right)\cdot\ln\alpha_l+\left(\sum_{j=1}^{m}\gamma_{jl}|x_j, \alpha^{(i)}, \theta^{(i)}\right)\left[\ln\frac{1}{\sqrt{2\pi}}-\ln\sigma_l-\frac{(x_j-\mu_l)^2}{2\sigma_l^2}\right]\right\}\right\} \\ &= \sum_{l=1}^{k}\left\{\left(\sum_{j=1}^{m}E\left(\gamma_{jl} \mid x_j, \alpha^{(i)}, \theta^{(i)}\right)\right)\ln\alpha_l+\left(\sum_{j=1}^{m}E\left(\gamma_{jl} \mid x_j, \alpha^{(i)}, \theta^{(i)}\right)\right)\left[\ln\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_l-\frac{1}{2\sigma_l^2}(x_j-\mu_l)^2\right]\right\}\end{aligned} \tag{42}$$

其中，$E\left(\gamma_{jl} \mid x_j, \alpha^{(i)}, \theta^{(i)}\right)$则是基于当前参数对隐变量$\gamma_{jl}$的估计：

$$\begin{aligned}E\left(\gamma_{jl} \mid x_j, \alpha^{(i)}, \theta^{(i)}\right) &= P\left(\gamma_{jl}=1 \mid x_j, \alpha^{(i)}, \theta^{(i)}\right) \\ &= \frac{P\left(\gamma_{jl}=1, x_j \mid \alpha^{(i)}, \theta^{(i)}\right)}{\sum_{t=1}^{k}P\left(\gamma_{jt}=1, x_j \mid \alpha^{(i)}, \theta^{(i)}\right)} \\ &= \frac{P\left(\gamma_{jl}=1 \mid \alpha^{(i)}, \theta^{(i)}\right)\cdot P\left(x_j \mid \gamma_{jl}=1, \alpha^{(i)}, \theta^{(i)}\right)}{\sum_{t=1}^{k}P\left(x_j \mid \gamma_{jt}=1, \alpha^{(i)}, \theta^{(i)}\right)P\left(\gamma_{jt}=1 \mid \alpha^{(i)}, \theta^{(i)}\right)} \\ &= \frac{\alpha_l^{(i)}\cdot\phi\left(x_j \mid \theta_l^{(i)}\right)}{\sum_{t=1}^{k}\alpha_t^{(i)}\cdot\phi\left(x_j \mid \theta_t^{(i)}\right)} \\ &= \hat{\gamma}_{jl}^{(i)}, \quad j=1,2,\cdots,m; \quad l=1,2,\cdots,k\end{aligned} \tag{43}$$

将$E\left(\gamma_{jl} \mid x_j, \alpha^{(i)}, \theta^{(i)}\right)$带回$Q$函数可得：

$$Q\left(\theta, \alpha^{(i)}, \theta^{(i)}\right)=\sum_{l=1}^{k}\left\{\left(\sum_{j=1}^{m} \hat{\gamma}_{jl}^{(i)}\right) \ln \alpha_l+\left(\sum_{j=1}^{m} \hat{\gamma}_{jl}^{(i)}\right)\left[\ln \left(\frac{1}{\sqrt{2 \pi}}\right)-\log \sigma_l-\frac{1}{2 \sigma_l^2}\left(x_j-\mu_l\right)^2\right]\right\} \tag{44}$$

2.2.3. M步：Q函数最大化

有了Q函数，就可以对Q函数进行最大化，得到下一次迭代的模型参数了，即：

$$\alpha^{(i+1)}, \theta^{(i+1)}=\arg \max _\theta Q\left(\theta, \alpha^{(i)}, \theta^{(i)}\right) \tag{45}$$

采用求导及拉格朗日乘子法，可得第$i+1$轮参数为：

$$\begin{aligned}
\hat{\mu}_l &= \frac{\sum_{j=1}^m \hat{\gamma}_{jl}^{(i)} x_j}{\sum_{j=1}^m \hat{\gamma}_{jl}^{(i)}} \\
\hat{\sigma}_l^2 &= \frac{\sum_{j=1}^m \hat{\gamma}_{jl}^{(i)}\left(x_j-\mu_l^{(i)}\right)^2}{\sum_{j=1}^N \hat{\gamma}_{jl}^{(i)}} \\
\hat{\alpha}_l &= \frac{\sum_{j=1}^m \hat{\gamma}_{jl}^{(i)}}{m} \\
&\text{其中，} l=1,2, \cdots, k
\end{aligned} \tag{46}$$

3. 例子

见《西瓜书》P210页。

4. 参考文档

1. [(48条消息) 详解EM算法与混合高斯模型(Gaussian mixture model, GMM)林立民爱洗澡-*CSDN*博客混合高斯模型]()
2. 《西瓜书》 P206