作业二、基于 EM 算法的高斯混合模型参数估计

崔多
1479518308@qq.com

一 摘要

高斯混合模型可以看做多个高斯分布的加权和，其中每个高斯分布代表了数据中的一个子群。EM 算法是一种用于无监督学习的迭代算法，常用于解决含有隐变量的概率模型参数估计问题。本文通过 EM 算法，对高斯混合模型的参数进行估算，并对预测效果进行评估。

二 研究内容简介

在高斯混合模型中，假设数据由多个高斯分布组成。每个高斯分布都对应着一个混合系数。可通过 EM 算法解决高斯混合模型的参数估计问题。

三 实验方法

3.1 EM 算法

EM（Expectation-Maximization）算法是一种迭代优化算法，该算法主要是用于在给定数据集的情况下求解包含隐变量的概率模型参数，通过求取关于隐变量的期望并使其最大化来估计模型中的参数。该算法主要包含两个步骤，即 E 步和 M 步。

假设$\{x_1, x_2, \dots, x_n\}$为独立分布的观测数据，其联合分布为$\mathrm{f(x;\theta)}$，θ为参数集合，$\{z_1, z_2, \dots, z_n\}$为分布已知的隐变量。

1）E 步（Expectation）：求解隐变量在当前参数下的期望，计算每个样本的后验概率。即对于第 i 个样本，已知第 k 步估计值θ_k的情况下，求解关于隐函数的对数似然期望，如式（3.1）所示。

$$\mathrm{Q(\theta|X}, \theta_k) = \ E_z \log f(\theta; x, z) \tag{3.1}$$

2） M 步（Maximization）：对于给定的隐变量期望值，求取使其对数似然函数最大化的参数值。基于 E 步骤得到的后验概率，更新模型参数，并计算新的参数。极大值计算公式如式(3.2)所示。

$$\mathrm{Q}(\theta_{k+1}|X, \theta_k) = \ \max_{\theta} Q(\theta|X, \theta_k) \tag{3.2}$$

这样就完成一次迭代，通过交替进行 E 步和 M 步，直到模型收敛，即可得到概率模型的最大似然估计或最大后验概率估计参数值。

3.2 高斯混合模型

高斯混合模型（gaussian mixture model）是一种用于描述数据分布的概率模型。它假设数据由多个高斯分布组成，每个高斯分布称为一个混合成分，而混合系数a_k则表示每个成分在总体中所占比例，且满足$\sum a_k = 1, a_k \geq 0$。其分布一般满足式（3.3）：

$$\mathrm{f}(\mathrm{x};\theta)=\sum_{k=1}^{m}a_k\varphi(x;\theta_k) \tag{3.3}$$

$\varphi(x;\theta_k)$是高斯概率密度函数，$\theta_k=(\mu_k,\sigma_k^2)$，其定义为：

$$\varphi(x;\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right) \tag{3.4}$$

称为第 k 个分模型，μ_k,σ_k^2称为均值与方差。在进行概率密度估计时，GMM 模型可以计算出每个样本属于每个类别的概率，从而可以对样本进行分类。具体地，对于一个新的样本，可以计算它属于每个高斯分布的概率，然后将概率最大的类别作为样本的分类结果。

3.3 EM 算法估计高斯混合模型参数

假定观测数据$\{x_1,\ x_2,\dots,x_n\}$由高斯混合模型生成，即满足公式（3.3），以此利用 EM 算法估计模型参数θ，则估计步骤如下。

1）明确隐变量与似然函数：在高斯混合明星中，观测数据是已知的，观测点隶属于哪一个高斯分布分模型是未知的，故而，隐变量z_{ik}定义如下：

$$z_{ik}=\begin{cases}1, & \textit{第i个感测点隶属于第k个分模型}\\ 0, & else\end{cases} \tag{3.5}$$

由此可得似然函数

$$\mathrm{f}(\mathrm{x},\mathrm{z},\theta)=\prod_{i=1}^{n}\prod_{k=1}^{m}(a_k\varphi(x;\theta_k))^{z_{ik}}=\prod_{k=1}^{m}a^{\sum_{k=1}^{n}z_{ik}}\prod_{i=1}^{n}\left(\varphi(x;\theta_k)\right)^{z_{ik}} \tag{3.6}$$

结合公式（3.4），可以得到完全数据的似然函数为：

$$\mathrm{f}(\mathrm{x},\mathrm{z},\theta)=\prod_{k=1}^{m}a_k^{n_k}\ \prod_{i=1}^{n}\left(\ \frac{1}{\sqrt{2\pi}\sigma_k}exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)\right)^{z_{ik}} \tag{3.7}$$

由此可求得其对数似然函数。

$$\mathrm{L}(\mathrm{x},\mathrm{z},\theta)=\ \log f(x,z,\theta) \tag{3.8}$$

2）求解隐变量期望：易知隐变量期望为：

$$E(z_{ik})=1*P(z_{ik}=1|x_i)+0*P(z_{ik}=0|x_i)=P(z_{ik}=1|x_i) \tag{3.9}$$

根据贝叶斯公式，可知

$$P(z_{ik}=1|x_i)=\ \frac{P(z_{ik}=1,x_i)}{P(x_i)}=\ \frac{a_k\varphi(x_i;\theta_k)}{\sum_{k=1}^{m}a_k\varphi(x_i;\theta_k)} \tag{3.10}$$

根据公式 3.8 以及公式 3.1，关于隐蔽那辆的对数似然函数的期望

3）更新参数使隐变量期望最大：根据公式 3.2 可知，使得目标函数达到极大值点，也就是函数的导数为 0 时：

$$\frac{\partial Q}{\partial\mu_k}=\sum_{i=1}^{n}z_{ik}\left(-\frac{(x-\mu_k)}{2\sigma_k^2}\right)*(-1)=\sum_{i=1}^{n}z_{ik}(x-\mu_k)=0 \tag{3.11}$$

同理可得：

$$\frac{\partial Q}{\partial\sigma_k^2}=\sum_{i=1}^{n}z_{ik}\left(-\frac{1}{2\sigma_k^2}+\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)=\sum_{i=1}^{n}z_{ik}[\sigma_k^2+\sum_{i=1}^{n}z_{ik}(x-\mu_k)^2]=0 \tag{3.12}$$

$$\frac{\partial L(a_k)}{\partial a_k} = \sum_{i=1}^{n} \frac{z_{ik}}{a_k} + \gamma = 0 \tag{3.13}$$

其中γ为拉格朗日算子，根据$\sum a_k = 1$可知:

$$a_k = \frac{\sum_{i=1}^{n} z_{ik}}{n} \tag{3.14}$$

通过设定函数收敛值特定范围或控制迭代次数可以求得相关参数。

四 实验步骤

本文基于 python 语言，建立 EM 模型步骤如下:

1）设定参数θ，本文基于生成文件代码，高斯函数混合模型数据分布为 1：3 以及相关参数，设定初始参数为：$a_1 = 0.3, \mu_1 = 160, \sigma_1 = 3$，$a_2 = 0.7, \mu_2 = 180, \sigma_2 = 5$，依照此初始数据进行迭代。

2) 计算隐变量z_{ik}的数学期望，其中隐变量可以视为观测点归属于某一高斯分布的概率，观测点可视为隶属于隐变量最高的高斯模型。根据公式 3.10 可求得隐变量期望为:

$$E(z_{ik}) = \frac{a_k \varphi(x_i; \theta_k)}{\sum_{k=1}^{m} a_k \varphi(x_i; \theta_k)} \tag{4.1}$$

3)计算并更新迭代参数值，根据公式 3.11-3.14，可以获得参数值为:

$$\mu_k = \frac{\sum_{i=1}^{n} z_{ik} x_i}{\sum_{i=1}^{n} z_{ik}} \tag{4.2}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{n} z_{ik} (x - \mu_k)^2}{\sum_{i=1}^{n} z_{ik}} \tag{4.3}$$

$$a_k = \frac{\sum_{i=1}^{n} z_{ik}}{n} \tag{4.4}$$

4)重复步骤 2，3，直至连续两轮迭代，数据变化量小于 0.0001，设为达到收敛精度。

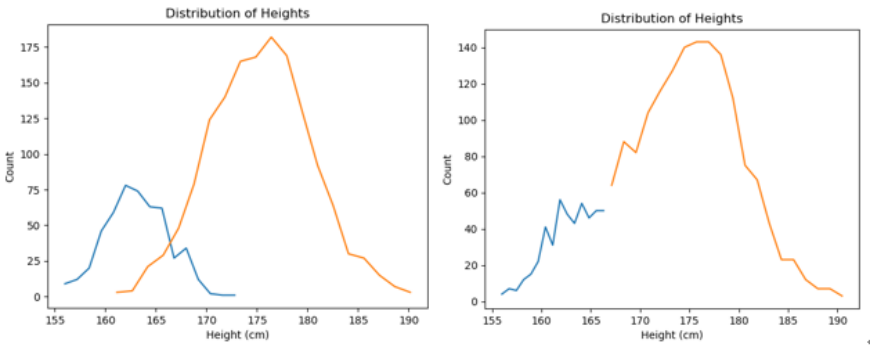
五 实验结果及评估

经过上述实验步骤，进行共 53 次迭代，获取到的最终高斯混合模型为:$a_1 = 0.2291, \mu_1 = 163.4399, \sigma_1 = 2.8735$，$a_2 = 0.7709, \mu_2 = 175.8243, \sigma_2 = 5.1834$。

与真实数据$a_1 = 0.25, \mu_1 = 164, \sigma_1 = 3$，$a_2 = 0.75, \mu_2 = 176, \sigma_2 = 5$相比较，相对误差低于 8%，有较好的收敛效果。

为了进一步分析 EM 模型的准确度，需要观察 EM 模型对于观测点的模型划分。根据隐变量z_{ik}的定义可知，当隐变量在第 i 个高斯模型中的期望最大时，可以认为该观测点属于该高斯模型，原始数据中，属于模型 1 的观测点有 500 个，模型 2 的观测点由 1500 个。而在 EM 模型中，属于模型 1 的观测点由 485 个，属于模型 2 的观测点由 1515 个。

为了进一步查看 EM 的划分准确度，本文分别绘制原始数据与 EM 模型数据的分布直方图，如图 1 所示。



图（a）原始数据直方图

图（b）EM 模型划分数据直方图

1 隶属于两种模型的身高分布直方图

可见在数据交叠较为严重的部分，存在着一定的划分错误情况，本文经计算可得，误判数据 174 个，误判率为 8.7%。

六 结论

高斯混合模型在诸多领域都有着较为广泛的应用，基于 EM 算法的高斯混合模型参数估计是一种有效的方法，可以用来求解高斯混合模型的最大似然估计，从而对数据进行聚类、分类、降维等任务。

本文基于 EM 模型，对求解高斯混合模型参数进行推到域代码实现，并对结果进行分析，获得了较好的实验效果。EM 算法作为一种迭代算法，对于含有隐函数的模型有较好的解决能力，针对不同的概率分布模型可以通过不同的估计形式进行推导已经被广泛应用于统计学、机器学习、计算机视觉等领域，在未来仍有广泛的应用前景。