基于 EM 算法估计高斯混合模型的参数

摘要

本文主要介绍 EM 算法在高斯混合模型中的应用，以身高数据为例进行了实现和评估。高斯混合模型是一种包含多个高斯分布的概率模型，可用于对数据集中的复杂分布进行建模。EM 算法是一种求解包含隐变量的概率模型的最大似然估计的迭代优化算法，通过交替进行期望和最大化操作，不断逼近对参数的最大似然估计值。在本文中，我们使用 EM 算法对身高数据进行了一维高斯混合模型的参数估计，并通过对模型的评估和性能分析来说明该方法的有效性和实用性。

关键词：EM 算法、高斯混合模型

一、引言

在许多实际问题中，我们往往只能观察到数据的部分特征，而无法直接观察到数据的全部信息，这就需要使用包含隐变量的概率模型来描述数据的生成过程。EM 算法就是一种求解这种包含隐变量的概率模型的最大似然估计的方法。EM 算法包含两个步骤，E 步骤（Expectation step）和 M 步骤（Maximization step）。EM 算法是一种迭代优化算法，通过交替进行期望和最大化操作，不断逼近对参数的最大似然估计值。该算法在统计学、机器学习等领域中得到广泛应用，如高斯混合模型、隐马尔可夫模型、朴素贝叶斯分类器等。

高斯混合模型（GMM）是一种常见的概率模型，被广泛应用于聚类、图像处理、模式识别等领域。在 GMM 中，每个样本被认为是由多个高斯分布的线性组合生成的，因此 GMM 具有更强的表达能力，可以更好地适应真实世界的数据。本文使用 EM 算法估计 GMM 参数。

在本文中，使用 EM 算法来估计身高数据集的一维高斯混合模型的参数。本文的目的是通过对该实验的分析，展示 EM 算法在高斯混合模型参数估计中的应用，并探讨 EM 算法的一些特性和应用中可能遇到的一些问题。

二、EM 算法

EM 算法是一种迭代算法，可用于对含有隐变量的模型进行参数估计。它的基本思路是在每次迭代中，根据当前参数估计隐变量的后验概率重新估计参数，直到收敛为止。EM 算法通常包含两个步骤，分别是 E 步和 M 步。其中，E 步估计隐变量的后验概率，M 步估计参数。

在高斯混合模型中，每个样本被认为是由多个高斯分布的线性组合生成的。假设有k个高斯分布，每个高斯分布的均值为μ_i，方差为σ_i^2，系数为π_i，则每个样本的概率密度函数可以表示为$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x \mid \mu_i, \sigma_i^2)$。其中$\mathcal{N}(x \mid \mu_i, \sigma_i^2)$表示均值为 μ_i，方差为σ_i^2的高斯分布在x处的取值。

假设已经知道每个样本的概率密度函数，可以使用极大似然估计来估计每个高斯分布的参数。设$\theta = \{\pi_i, \mu_i, \sigma_i^2\}$为模型的参数，$\mathcal{N}\{x\} = \{x_1, x_2, \dots, x_N\}$为样本容量为$N$的数据集，则似然函数为$p(x|\theta) = \prod_{i=1}^{N} \sum_{j=1}^{k} \pi_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)$。

我们的目标是找到使得似然函数取最大值的参数θ。但是，由于每个样本都可能来自于不同的高斯分布，因此每个样本的概率密度函数都是未知的。因此，使用 EM 算法来估计参数θ。

EM 算法的步骤如下：

1) 初始化模型参数$\theta = \{\pi_i, \mu_i, \sigma_i^2\}$，通常随机生成。
2) E 步：计算每个样本属于每个高斯分布的概率。对于第i个样本x_i，其属于第j个高斯分布的概率为$w_{ij} = \frac{\pi_j \mathcal{N}(x_i|\mu_j, \sigma_j^2)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i|\mu_l, \sigma_l^2)}$。其中$\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i|\mu_l, \sigma_l^2)$是每个高斯分布在$x_i$处的取值之和，$w_{ij}$表示$x_i$属于第$j$个高斯分布的概率。在 E 步中，通过计算后验概率来估计隐变量z_{ij}。
3) M 步：使用 E 步中计算得到的后验概率重新估计模型参数$\theta = \{\pi_i, \mu_i, \sigma_i^2\}$，即使用最大似然估计来计算每个高斯分布的参数。首先，根据 E 步中计算得到的后验概率计算出每个高斯分布的权重π_i，均值μ_i和方差σ_i^2，计算公式如下。

$$\begin{cases} \pi_i = \dfrac{1}{N} \displaystyle\sum_{j=1}^{n} w_{ij} \\ \mu_i = \dfrac{\sum_{j=1}^{N} w_{ij} x_j}{\sum_{j=1}^{N} w_{ij}} \\ \sigma_i^2 = \dfrac{\sum_{j=1}^{N} w_{ij} (x_j - \mu_i)^2}{\sum_{j=1}^{N} w_{ij}} \end{cases}$$

其中，N为数据集中样本的数量，w_{ij}表示样本x_i属于第j个高斯分布的概率。

4) 重复步骤 2 和 3，直到收敛为止。可以定义一个收敛条件，例如当两次迭代的似然函数值之差小于一个预设的阈值时停止迭代。也可以定义最大迭代次数。

三、实验结果

本文使用一个包含了 2000 个身高样本身高数据集。利用直方图绘制身高数据集，如图 1 所示。

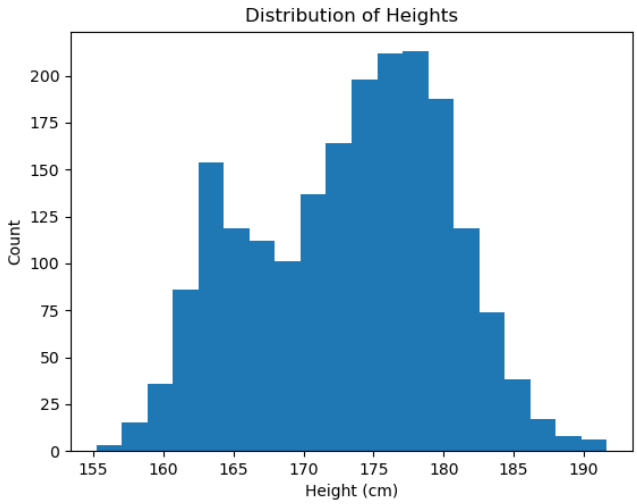

图 1 身高数据直方图

可以看出这些身高数据呈现出包含了两个近似高斯分布的形态。那么假设身高数据来自服从两个高斯分布的数据集，不妨设数据集X_i服从参数为μ_i, σ_i^2的高斯分布，即$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), i = 1,2$。而对于某个给定的样本，设其来自数据集$X_i$的概率是$\pi_i$，且$\sum_{i=1}^{2} \pi_i = 1$。

然后使用 EM 算法来估计身高数据集的一维高斯混合模型的参数μ_i, σ_i, π_i $(i = 1,2)$。EM 算法估计得到了两个高斯分布的参数：一个均值为 163.97539521，方差为 2.73077224；另一个均值为 176.22763695，方差为 4.92303133；权重π为[0.24682851, 0.75317149]。将估计得到的高斯混合模型和原始数据进行可视化比较，绘制身高分布直方图和估计的高斯混合模型的概率密度函数，如图 2 所示。

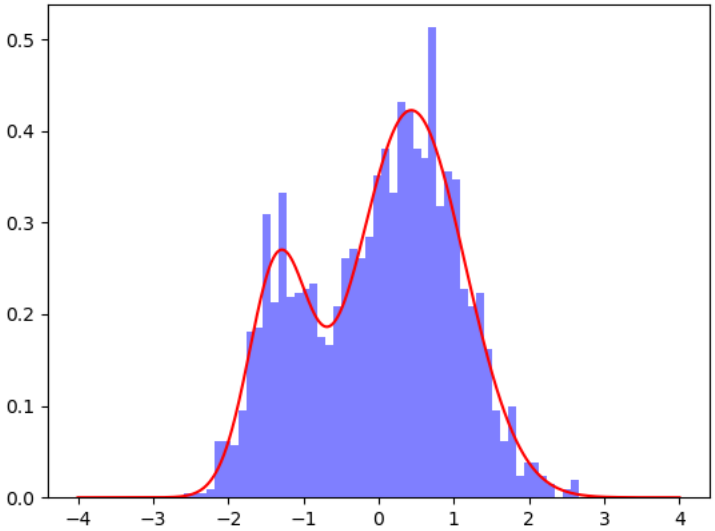

图 2 EM 算法估计的高斯混合模型

可以看到，估计得到的高斯混合模型能够较好地拟合身高数据的分布，因为它包含了两个高斯分布，其中一个分布对应于身高比较矮的人群，另一个分布对应于身高比较高的人群。这表明身高数据是一个双峰分布，这一结论可以帮助我们更好地理解身高数据。

四、结论

本文使用 EM 算法对身高数据集估计一维高斯混合模型的每个高斯分布的均值、方差和权重，能够较准确地估计数据的分布，具有较好的性能和可解释性，可以应用于数据聚类、异常检测、密度估计等多个领域。

但是，使用 EM 算法估计高斯混合模型参数可能会受到数据的影响。如在数据不足或者噪声较大的情况下，估计结果可能会不稳定或者偏差较大。因此，在应用该模型时，我们需要对数据进行适当的预处理，并结合实际应用场景来做出综合判断。