EM 算法估计高斯混合模型参数

——自然语言处理的深度学习报告(第二次)

崔潇雅

ZY2203502

1 摘要

计算单个高斯分布的参数时可以使用最大似然估计法，但计算混合高斯模型时，由于隐变量的存在无法使用最大似然估计法求导，因此使用迭代的思想使用 EM 算法进行求解。本文以给定参数的两个标准高斯模型生成的混合高斯模型分布数据为样本，按照 3：1 划分数据集和测试集，求解每个高斯分布参数后在测试集上预测并评估模型性能。

2 论文理论基础

2.1 问题描述

运行 students_height.py，自己给定一组由两个分模型生成的 2000 个混合高斯分布身高数据，已知男、女生身高都服从标准高斯分布，利用 EM 算法求解两个分模型的参数并且使用这些参数进行预测，评价模型性能。

2.2 高斯混合模型

高斯混合模型就是用高斯概率密度函数精确的量化事物，将一个事物分解为若干基于高斯概率密度函数的模型。任何一个曲线，无论多么复杂，都可以用若干个标准高斯曲线来无限逼近它，这就是高斯混合模型的基本思想。

假设高斯混合模型是从多个高斯分布中生成的，每个单独的分模型都是标准的高斯模型，定义高斯混合分布公式为：

$$P(x \mid \theta)=\sum_{k=1}^{K} \alpha_k \phi(x \mid \theta_k)$$

其中，α为各个高斯模型的在混合模型中发生的概率，$\sum\alpha$=1；$\phi(x \mid \theta_k)$是标准高斯分布密度函数：$\emptyset(y \mid \theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$，$\mu_k$、$\sigma_k$为第 k 个分模型的期望、方差。

2.3 EM 求解高斯混合模型

本文不再赘述一般的EM算法，只介绍通过 EM 迭代更新高斯混合模型参数的方法（有样本数据X_1，X_2，$X_3 \ldots X_N$和一个有K个子模型的高斯混合模型，想要推算出这个高斯混合模型的最佳参数）：

(1) 首先初始化参数。

(2)E-step：依据当前参数，计算每个数据 j来自子模型 k 的可能性：

$$\gamma_{jk}=\frac{\alpha_k \phi(x_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(x_j|\theta_k)}, j=1,2,\ldots,N; k=1,2,\ldots,K$$

(3)M-step：计算新一轮迭代的模型参数:

$$\mu_k = \frac{\sum_j^N (\gamma_{jk} x_j)}{\sum_j^N \gamma_{jk}}, k = 1, 2, \ldots, K$$

$$\Sigma_k = \frac{\sum_j^N \gamma_{jk}(x_j - \mu_k)(x_j - \mu_k)^T}{\sum_j^N \gamma_{jk}}, k = 1, 2, \ldots, K$$

$$\alpha_k = \frac{\sum_{j=1}^N \gamma_{jk}}{N}, k = 1, 2, \ldots, K$$

(4)重复计算 E-step 和 M-step 直至收敛（可以极小正数收敛法和迭代次数收敛法，本文使用足够大的迭代次数）

至此，就找到了高斯混合模型的参数。需要注意的是，EM 算法具备收敛性，但并不保证找到全局最大值，有可能找到局部最大值。本文解决方法是初始化几次不同的参数进行迭代，取方差之和更小的那次。

2.3 Log-Likelihood 函数

混合高斯模型的对数似然估计定义为：

$$logL(\theta) = \sum_{j=1}^N logP(x_j|\theta) = \sum_{j=1}^N log(\sum_{k=1}^K \alpha_k \phi(x|\theta_k))$$

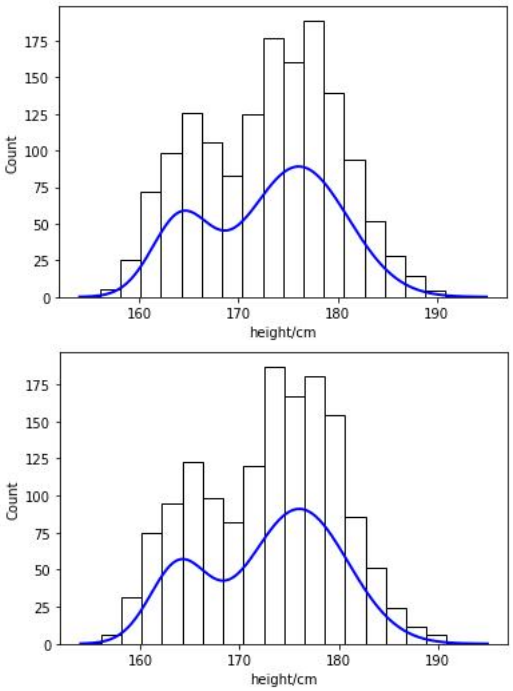
3 实验结果

4.1 参数计算及绘图

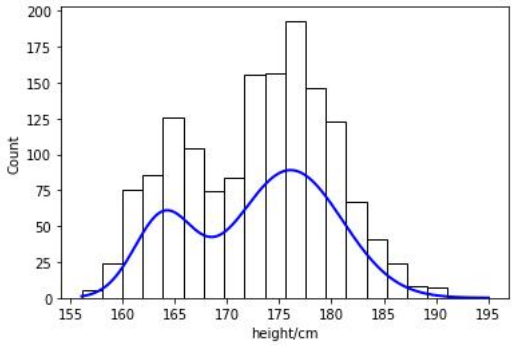
首先将2000个身高数据按照3：1划分成训练集和测试集。选取初始值，依据生活经验设置男女生比例各占50%：α_1=α_2=0.5，男女生初始平均身高值：μ_1=175，μ_2=160，男女生初始身高方差：σ_1=σ_2=10，之后循环E步和M步运行程序迭代500次得到结果。为避免陷入局部最优，选取不同的初值进行迭代，下面展示三组不同初始参数时的收敛值及对应的混合高斯分布直方图：

第一组	初始值	收敛值
μ_1	175	176.083796
μ_2	160	164.169820
α_1	0.5	0.742030
α_2	0.5	0.257970
σ_1	10	4.982411
σ_2	10	2.895594

第二组	初始值	收敛值
μ_1	180	176.034192
μ_2	160	163.875168
α_1	0.5	0.747093
α_2	0.5	0.252907
σ_1	10	4.917021
σ_2	10	2.687023



第三组	初始值	收敛值
μ_1	175	176.083411
μ_2	150	163.930068
α_1	0.5	0.738727
α_2	0.5	0.261273
σ_1	10	4.969938
σ_2	10	2.786199



选取不同的初始值，参数收敛值也不同，男女生平均身高值、比例值三组数据偏差不大，相对而言方差值有所差距，第二组数据更小，比其他模型更贴近真实模型。

4.2 测试集

在测试集上计算三组模型Log-Likelihood值，发现第二组模型预测效果更好。

	第一组	第二组	第三组
Log-Likelihood	-1649.7476	-1651.8370	-1647.8863

4.3 模型评价

本文使用EM算法对两组男女身高数据进行了分类，结果表明，EM算法在有限次内收敛，并且计算出的参数值与真实性非常接近。但是对初始值敏感，给定不同的初始参数会收敛到不同的结果，容易陷入局部最优导致不能得到全局最优解。

6 参考文献

[1](13条消息) 机器学习最易懂之EM算法详解与python实现_Elenstone的博客-CSDN博客