EM 算法估计高斯混合模型参数

彭海琅 ZY2203322

plunck@buaa.edu.cn

摘要

对于所含变量可观测的样本数据集，我们一般使用极大似然估计法或贝叶斯估计法估计模型参数，但如果样本数据集存在不可观测的隐变量，单纯的极大似然估计就不再可行。EM 算法（Expectation-Maximization algorithm, EM）是一种迭代算法，E 步主要用初始值或者上一个 M 步估计的参数值计算似然函数的期望，M 步寻找似然函数最大化时对应的参数，通过 E 步和 M 步两个迭代步骤，每次迭代都使极大似然函数增加。但是，由于初始值的不同，可能会使似然函数陷入局部最优。本文利用 EM 算法求解了由两个高斯分布混合产生的身高数据。

导论

一、高斯混合模型

高斯混合模型是指具有如下形式的概率分布参数模型

$$P(y|\theta) = \sum_{k=1}^{K} \alpha_k \emptyset(y|\theta_k)$$

其中，α_k是系数，$\alpha_k \geq 0$，$\sum_{k=1}^{K} \alpha_k = 1$；$\emptyset(y|\theta_k)$是高斯分布密度，$\theta_k = (\mu_k, \sigma_k^2)$，$\emptyset(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$称为第$k$个分模型。

二、EM 算法

最大期望算法（Expectation-Maximization algorithm, EM）是一类通过迭代进行极大似然估计的优化算法，通常作为牛顿迭代法的替代用于对包含隐变量或缺失数据的概率模型进行参数估计。EM 算法的标准计算框架由 E 步（Expectation-step）和 M 步（Maximization step）交替组成，算法的收敛性可以确保迭代至少逼近局部极大值。

EM 算法是基于极大似然估计（Maximum Likelihood Estimation, MLE）理论的优化算法。给定相互独立的观测数据$X = \{X_1, X_2, \dots, X_N\}$，和包含隐变量$Z$、参数$\theta$的概率模型$f(X, Z, \theta)$，根据 MLE 理论，$\theta$的最优单点估计在模型的似然取极大值时给出：$\theta = \underset{\theta}{argmax}\, p(x|\theta)$。考虑隐变量，模型的似然有如下展开

$$p(x|\theta) = \sum_{c=1}^{k} p(x, Z_c|\theta),\ Z = \{Z_1, \dots, Z_k\}$$

隐变量可以表示缺失数据，或概率模型中任何无法直接观测的随机变量,求和的部分也被称为X，Z的联合似然（joint likelihood）。

由 MLE 的一般方法，对上式取自然对数后可得：

$$logP(x|\theta) = log \prod_{i=1}^{N} P(X_i|\theta) = \sum_{i=1}^{N} logP\,(X_i|\theta) = \sum_{i=1}^{N} log\left[\sum_{c=1}^{k} P(X_i, Z_c|\theta)\right]$$

上述展开考虑了观测数据的相互独立性。引入与隐变量有关的概率分布$q(Z)$，即隐分布（可认为隐分布是隐变量对观测数据的后验），由 Jensen 不等式，观测到的对数似然有如下不等关系：

$$logP(x|\theta) = \sum_{i=1}^{N} log\left[\sum_{c=1}^{k} \frac{q(Z_c)}{q(Z_c)} P(X_i, Z_c|\theta)\right]$$

$$\geq \sum_{i=1}^{N}\sum_{c=1}^{k}\left[q(Z_c) log \frac{P(X_i, Z_c|\theta)}{q(Z_c)}\right] \equiv L(\theta, q)$$

当θ，q使不等式右侧取全局极大值时，所得到的θ至少使不等式左侧取局部极大值。因此，将不等式右侧表示为$L(\theta,q)$后，EM 算法有如下求解目标：

$$\hat{\theta} = \underset{\theta}{argmax} L(\theta,q)$$

对于 E 步，当$q(Z) = P(Z|X,\theta)$时，$L(\theta,q)$取极大值。

对于 M 步，求解模型参数使得$L(\theta,q)$取极大值，该极值问题的必要条件是$\frac{\partial L(\theta,q)}{\partial \theta} = 0$。

实验内容

一、问题描述

给定一组数据集，有 2000 位同学的身高数据，已知男生，女生的身高都服从高斯分布，这两个高斯分布的均值和方差我们都不知道，男生和女生的人数我们也不知道，利用 EM 算法求解出这两个不同的分布。

二、EM 算法实现步骤

1、取参数的初始值开始迭代。

2、E 步(expectation)：依据当前模型参数，计算分模型k对观测数据y_j的响应度，男生女生的身高满足两个不同的正态分布，因此只有两个分模型（$k = 1，2$）

$$\widehat{\gamma_{jk}} = \frac{\alpha_k \emptyset(y_j|\theta_k)}{\sum_{k=1}^{2} \alpha_k \emptyset(y_j|\theta_k)}, j = 1,2,3,\ldots,2000; k = 1,2$$

3、M 步(maximization)：计算新一轮迭代的模型参数

$$\widehat{\mu_k} = \frac{\sum_{j=1}^{2000} \widehat{\gamma_{jk}} y_j}{\sum_{j=1}^{2000} \widehat{\gamma_{jk}}}, k = 1,2$$

$$\widehat{\sigma_k}^2 = \frac{\sum_{j=1}^{2000} \widehat{\gamma_{jk}} (y_j - \mu_k)^2}{\sum_{j=1}^{2000} \widehat{\gamma_{jk}}}, k = 1,2$$

$$\widehat{\alpha_k} = \frac{\sum_{j=1}^{2000} \widehat{\gamma_{jk}}}{2000}, k = 1,2$$

4、重复第 2 步和第 3 步，直到收敛。

实验结果

初始值的选取对结果有较大影响，定理只能保证参数估计序列收敛到对数似然函数序列的稳定点，不能保证收敛到极大值点．所以在应用中，初值的选择变得非常重要，常用的办法是选取几个不同的初值进行迭代，然后对得到的各个估计值加以比较，从中选择最好的。

第一组	初始值		收敛值
μ_1	175	μ_1	173.064648
μ_2	175	μ_2	173.064648
σ_1	1	σ_1	6.93355098
σ_2	1	σ_2	6.93355098
α_1	0.5	α_1	0.5
α_2	0.5	α_2	0.5

第二组	初始值		收敛值
μ_1	175	μ_1	164.204960
μ_2	180	μ_2	176.225170
σ_1	1	σ_1	3.096558
σ_2	1	σ_2	4.879851
α_1	0.5	α_1	0.262934
α_2	0.5	α_2	0.737066

第三组	初始值		收敛值
μ_1	175	μ_1	176.22517
μ_2	175	μ_2	164.204961
σ_1	1	σ_1	4.87985
σ_2	10	σ_2	3.096558
α_1	0.5	α_1	0.737066
α_2	0.5	α_2	0.262934

第四组	初始值		收敛值
μ_1	180	μ_1	176.22515
μ_2	150	μ_2	164.204935
σ_1	10	σ_1	4.879863
σ_2	10	σ_2	3.096544
α_1	0.5	α_1	0.737068
α_2	0.5	α_2	0.262932

第五组	初始值		收敛值
μ_1	180	μ_1	176.22515
μ_2	150	μ_2	164.204935
σ_1	10	σ_1	4.879863
σ_2	10	σ_2	3.096544
α_1	0.9	α_1	0.737068
α_2	0.1	α_2	0.262932

第六组	初始值		收敛值
μ_1	180	μ_1	164.20496
μ_2	170	μ_2	176.225169
σ_1	100	σ_1	3.096558
σ_2	100	σ_2	4.879851
α_1	0.9	α_1	0.262934
α_2	0.1	α_2	0.737066

结论

正态分布参数分析：

	μ_1	μ_2	σ_1	σ_2	α_1	α_2
真实值	176	164	5	3	0.75	0.25
第 4、5 组	176.22515	164.204935	4.879863	3.096544	0.737068	0.262932
第 3 组	176.22517	164.204961	4.87985	3.096558	0.737066	0.262934
第 6 组	176.225169	164.20496	4.879851	3.096558	0.737066	0.262934

可知选取不同的初始值，参数可能会收敛到不同的局部极大值。选择方差之和比较小的组别，可能会比其他组更贴近真实的模型。