高斯混合模型的 EM 算法

ZY2203810

吴金旺

Abstract

本文首先介绍了高斯混合模型和 EM 算法，然后推导了高斯混合模型的 EM 算法，并将其用于人工生成的数据集中，最后对结果进行了分析。

Introduction

1.1 高斯分布和高斯混合模型

高斯分布也称正太分布，是统计学中拟合数据最常使用的一类分布，单变量的高斯分布概率密度函数可表示为：

$$\phi(x;\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.1}$$

其中θ是所有参数的集合，μ代表数学期望，σ^2代表则方差。

多维的高斯分布则表示为：

$$\phi(x;\theta)=\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right) \tag{1.2}$$

其中θ是所有参数的集合，x 是多个变量组成的向量，μ为数学期望，Σ为协方差，D 为数据维度。

高斯分布用处极广，日常生活中的很多统计量，如同学们的成绩、工作以后取得的薪资等一般都满足高斯分布。

现实中虽然很多数据满足高斯分布，但很多情况下这些数据会混杂在一起，并且我们也不一定有能力将这些数据区分开，然后单独拟合出它们的具体分布。多个高斯分布叠加在一起，通常不一定是我们期望的高斯分布。可定义高斯混合模型是 K 个高斯分布的组合，用以拟合复杂数据。

假设有 N 个观测数据$x_1,x_2,\dots,x_N$，并认为混合模型中有 K 个子高斯模型，那么高斯混合模型的概率分布可以表示为：

$$p(x|\theta)=\sum_k\alpha_k\phi(x\,|\theta_k)=\sum_k\alpha_k\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x-\mu_k)^2}{\sigma_k^2}\right) \tag{1.3}$$

式中α_k是观测数据属于第 k 个子模型的概率，可以认为是先验概率，这样公式就变为了在参数固定（即认为高斯混合模型参数Θ为常数）的条件下的全概率公式，并且有：

$$\sum_k \alpha_k = 1 \tag{1.4}$$

1.2 EM 算法

EM(Expectation Maximization)算法，也称最大期望算法，或 Dempster-Laird-Rubin 算法，是一类通过迭代进行极大似然估计的优化算法。

1.2.1 最大似然估计

EM 算法是基于最大似然估计法的一种算法。假设一组数据中的各个数据点相互独立，那么所有数据点的总概率可以表示为每个数据点概率的乘积，也就是似然函数，它可表示为：

$$p(x|\Theta) = \prod_i p(x_i, \theta) \tag{1.5}$$

那么根据最大似然估计法，理论上Θ应该使$p(x|\Theta)$取到最大值，即：

$$\Theta = argmax\big(p(x|\Theta)\big) = argmax\left(\prod_i p(x_i, \theta)\right) \tag{1.6}$$

一般我们利用求导来计算使$p(x|\Theta)$取到最大值的Θ，但由于等式右边是多个概率密度的连乘，这样求导起来相当复杂，所以一般对似然函数取对数，由于对数函数是单调的，所以目标变成了求取对数之后的函数的最大值点，即：

$$L(x|\Theta) = \sum_i \ln\big(p(x_i, \theta)\big) = \sum_i \ln\left[\sum_k \alpha_k \phi(x_i\,|\theta_k)\right] \tag{1.7}$$

$$\Theta = argmax\left(\sum_i \ln\big(p(x_i, \theta)\big)\right) = argmax\left(\sum_i \ln\left[\sum_k \alpha_k \phi(x_i\,|\theta_k)\right]\right) \tag{1.8}$$

可以看出，尽管已经取了对数，但公式中仍然包含两重求和，其中一重求和在对数函数内部，直接求导难度依旧很大，因此要进行进一步的化简。

1.2.2 隐函数

对于 1.2.1 节中提出的求导困难的问题，EM 算法提出用迭代的方法解决，通过对最优的高斯混合模型进行逼近而进行优化。为了帮助迭代算法的过程，EM

算法提出了隐参数 z, 每次迭代，先使用上一次的参数计算隐参数 z 的分布，然后使用 z 更新似然函数，对目标参数进行估计。

在高斯混合模型（GMM）估计中，EM 算法所设的隐变量 z 一般属于 1，2，⋯，K。那么，计算出 GMM 中 K 组高斯模型的参数之后，某个数据点x_i属于第 z 个高斯模型的概率可表示为：

$$p(z|x_i,\theta_k)\text{或}p(z|x_i,\mu_k,\sigma_k) \tag{1.9}$$

另根据全概率公式，有：

$$p(x_i|\theta)=\sum_k p(x_i|z=k,\mu_k,\sigma_k)p(z=k) \tag{1.10}$$

对比高斯混合分布，即式（1.3）中的$p(x|\theta)=\sum_k \alpha_k\phi(x\,|\theta_k)$，将$x=x_i$带入式（1.3），对比即可发现$\alpha_k$就是式（1.10）中$z$的先验分布，即：

$$\alpha_k=p(z=k) \tag{1.11}$$

而$z=k$情况下x_i的条件概率也就是第k个高斯模型的概率密度函数，即：

$$\phi(x_i|\theta_k)=\phi(x_i|\mu_k,\sigma_k)=p(x_i|z=k,\mu_k,\sigma_k) \tag{1.12}$$

因此可以把原有的部分参量替换为隐变量，将（1.11）和（1.12）带入到式（1.10）中，并引入隐变量 z，有：

$$\begin{aligned}L(x|\Theta)&=\sum_i \ln[p(x_i,z|\mu_k,\sigma_k)]\\&=\sum_i \ln\sum_k p(x_i|z=k,\mu_k,\sigma_k)p(z=k)\\&=\sum_i \ln\sum_k p(z=k|,x_i,\mu_k,\sigma_k)\frac{p(x_i|z=k,\mu_k,\sigma_k)p(z=k)}{p(z=k|x_i,\mu_k,\theta_k)}\end{aligned} \tag{1.13}$$

1.2.3 似然函数简化

由式（1.13）推导出的似然函数依然相对复杂，因此需要进一步简化。

引入对凸函数成立的 Jensen 不等式：

$$f[E(x)]\ge E[f(x)] \tag{1.14}$$

对比式（1.13）和式（1.14），令:

$$\begin{aligned}u&=\frac{p(x_i|z=k,\mu_k,\sigma_k)p(z=k)}{p(z=k|x_i,\mu_k,\theta_k)}\\f(u)&=lnu\\E(u)&=\sum_k p(z=k|,x_i,\mu_k,\sigma_k)u\end{aligned} \tag{1.15}$$

那么：

$$L(x|\Theta)=\sum_i f\big(E(u)\big)\geq\sum_i E[f(u)]$$
$$=\sum_i\sum_k p(z=k|,x_i,\mu_k,\sigma_k)\ln\frac{p(x_i|z=k,\mu_k,\sigma_k)p(z=k)}{p(z=k|x_i,\mu_k,\theta_k)} \tag{1.16}$$

于是似然函数简化成对数函数的两重求和。等式右侧给似然函数提供了一个下界，我们可以根据贝叶斯准则进行推导其中的后验概率，并令其等于参数$\omega_{i,k}$:

$$\begin{aligned}p(z=k|,x_i,\mu_k,\sigma_k)&=\frac{p(x_i|z=k,\mu_k,\sigma_k)p(z=k)}{\sum_k p(x_i|z=k,\mu_k,\sigma_k)p(z=k)}\\&=\frac{\alpha_k\phi(x|\mu_k,\sigma_k)}{\sum_k\alpha_k\phi(x|\mu_k,\sigma_k)}\\&=\omega_{i,k}\end{aligned} \tag{1.17}$$

则：

$$L(x|\Theta)=\sum_i ln\sum_k\omega_{i,k}\frac{\alpha_k\phi(x_i|\mu_k,\sigma_k)}{\omega_{i,k}}$$
$$\geq\sum_i\sum_k\omega_{i,k}ln\frac{\alpha_k\phi(x_i|\mu_k,\sigma_k)}{\omega_{i,k}} \tag{1.18}$$

EM 算法提出迭代逼近的方法，不断提高等式右边的下界的值，从而逼近似然函数。

1.2.4 迭代求解过程

每次迭代的目标函数为：

$$Q(\Theta,\Theta^{\mathrm{t}})=\sum_i\sum_k\omega_{i,k}^t ln\frac{\alpha_k\phi(x_i|\mu_k,\sigma_k)}{\omega_{i,k}^t} \tag{1.19}$$

迭代开始前，首先选定一组初始参数值Θ，然后间隔的更新ω和Θ。具体的来说，假设已经进行了 t 次迭代，此时有更新后的参数Θ^{t}，那么可以根据高斯混合模型计算$\omega_{i,k}^t$，即利用公式（1.17）。这一步称为 expectation step，或 E-step，有：

$$\omega_{i,k}^t=\frac{\alpha_k\phi(x|\mu_k,\sigma_k)}{\sum_k\alpha_k\phi(x|\mu_k,\sigma_k)} \tag{1.20}$$

然后，利用式（1.19），求出参数ω更新后，使目标函数Q最大的参数$\Theta^{\mathrm{t+1}}$，这一步称为 maximization step，或 M-step，即：

$$\Theta^{\mathrm{t+1}}=argmax\sum_i\sum_k\omega_{i,k}^t ln\frac{\alpha_k\phi(x_i|\mu_k,\sigma_k)}{\omega_{i,k}^t} \tag{1.21}$$

这样循环迭代，就完成了似然函数下界的不断提高和逼近。

Methodology

将式（1.1）带入式（1.19）上，就对单变量高斯混合模型使用 EM 算法，完整的目标函数：

$$
\begin{aligned}
Q(\Theta,\Theta^{\mathrm{t}}) &= \sum_{i}\sum_{k}\omega_{i,k}^{t} ln\frac{\alpha_k}{\omega_{i,k}^{t}\sqrt{2\pi}\sigma_k}\ \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right) \\
&= \sum_{i}\sum_{k}\omega_{i,k}^{t}\left(ln\alpha_k - ln\omega_{i,k}^{t} - ln\sqrt{2\pi}\sigma_k - \frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right)
\end{aligned} \tag{2.1}
$$

2.1 E-step

E-step 的目标就是计算隐参数的值，也就是对每一个数据点，分别计算其属于每一种高斯模型的概率，所以隐参量ω是一个 N×K 矩阵。每一轮新的迭代开始时，利用式（1.20）更新ω的值，由于公式上述已有，因此不过多介绍。

2.2 M-step

M-step 的任务就是最大化目标函数，从而求出高斯参数的估计。

α_k是观测数据属于第 k 个子模型的概率，由于它仅有$\sum\alpha_k = 1$以及非负的约束，因此这是一个受限优化的问题。

$$
\begin{gathered}
\alpha_k^{t+1} \coloneqq argmax\sum_{i}\sum_{k}\omega_{i,k}^{t} ln\alpha_k \\
subject\ to\ \sum_{k}\alpha_k = 1, \alpha_k > 0
\end{gathered} \tag{2.2}
$$

这种问题类似于数学中的多元函数求最值，通常用拉格朗日乘子法计算,下面构造拉格朗日乘子：

$$
L(\alpha_k,\lambda) = \sum_{i}\sum_{k}\omega_{i,k}^{t} ln\alpha_k + \lambda\left[\sum_{k}\alpha_k - 1\right] \tag{2.3}
$$

只需对式（2.3）求偏导数并令其等于 0，就可解得我们所需的是目标函数最大的α_k值，也就是我们所需的α_k^{t+1}，有：

$$
\begin{gathered}
\frac{\partial\mathcal{L}(\alpha_k,\lambda)}{\partial\alpha_k} = \sum_{i}\omega_{i,k}^{t}\frac{1}{\alpha_k} + \lambda = 0 \\
\alpha_k = -\frac{\sum_i \omega_{i,k}^{t}}{\lambda}
\end{gathered} \tag{2.4}
$$

将所有 k 项累加，就可以求得λ，即：

$$\lambda = -N$$
$$\alpha_k^{t+1} = \frac{\sum_i \omega_{i,k}^t}{N} \tag{2.5}$$

类似的，有：

$$\mu_k^{t+1} = \frac{\sum_i \omega_{i,k}^t x_i}{\sum_i \omega_{i,k}^t} \tag{2.6}$$

$$(\sigma_k^2)^{t+1} = \frac{\sum_i \omega_{i,k}^t (x_i - \mu_k^{t+1})^2}{\sum_i \omega_{i,k}^t} \tag{2.7}$$

这样就完成了参数的迭代。

Experimental Studies

3.1 数据集

一般来说人类的身高近似服从正太分布，但男女身高分布的期望和方差都有所不同。

我们的数据集是一个人工随机生成的单变量，K=2 的混合高斯模型，其中包括了 500 个女性样本，1500 个男性样本，其分布如表 1 所示。

表 1 样本分布

	男	女
均值	176	164
方差	25	9
人数	1500	500

图 3.1 表明样本的整体分布，它具有两个峰。

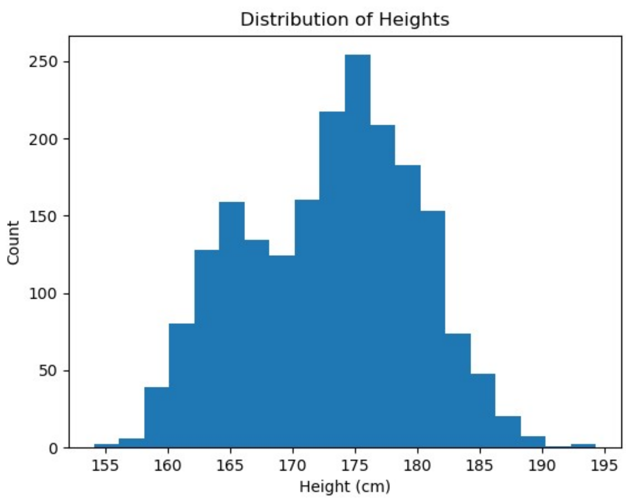

图 3.1 样本整体分布

3.2 超参数

超参数主要包括初值与迭代次数。如表 2 所示。

表 2 模型超参数

超参数	数值
迭代次数	100
$[\alpha_1^0, \alpha_2^0]$	[0.5, 0.5]
$[\mu_1^0, \mu_2^0]$	[160, 170]
$[(\sigma_1^2)^0, (\sigma_2^2)^0]$	[10, 10]
$\omega_{i,k}^0$	1

3.3 实验结果

可以使用第 2 章中建立的模型和算法进行求解，迭代得到的μ_k和σ_k^2直接代表第k个高斯模型的均值和方差；$\omega_{i,k}$衡量了第i个样本属于第k个高斯模型的概率，我们取其中概率最高的高斯模型作为样本预测的结果，这样可以得到预测的男女人数。

迭代 100 次后，得到实验结果如表 2 所示。

表 3 实验结果分布

	男	女
均值	175.88	164.12
方差	25.50	8.80
人数	1481	519

可以看出，表 3 与表 1 中样本的实际分布极其接近，基本可以认为，EM 算法成功得到了两个高斯模型的参数。

此外，如果我们将每个样本的预测结果与其实际标签作比较，会发现预测的正确率是 93.9%，不过这一参数意义不大，在两个高斯分布交界处的样本，是无法分辨出其具体属于哪一个高斯分布的。

Conclusion

本次作业分析了 EM 算法和其在高斯混合模型下的使用方法，这是一种十分有效的算法，可以将两个高斯模型分开，并分别求出它们的分布。EM 算法分为 E-step 和 M-step，两者共同将目标函数的下限不断提高，已到达求解最优值的目的。实验结果也表明了 EM 算法的有效性。