Report of Deep Learning for Natural Langauge Processing

刘新宇
LXYbhu@buaa.edu.cn

Abstract

高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，这 K 个子模型是混合模型的隐变量（Hidden variable），对于单高斯模型，我们可以用最大似然法（Maximum likelihood）估算参数θ的值，然而对于混合高斯分布的参数估计，通常需要使用 EM 算法进行求解。本次作业首先简要介绍了高斯模型和EM算法，随后利用 EM 算法，求解了由两个高斯分布混合产生的身高数据。

Methodology

1. 高斯混合模型

高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，这 K 个子模型是混合模型的隐变量（Hidden variable）。一般来说，一个混合模型可以使用任何概率分布，这里使用高斯混合模型是因为高斯分布具备很好的数学性质以及良好的计算性能。首先定义如下信息：

- x_j表示第j个观测数据，$j = 1,2, \ldots\ldots,N$
- K是混合模型中子高斯模型的数量，$k = 1,2, \ldots\ldots,K$
- α_k是观测数据属于第k个子模型的概率，$\alpha_k \geq 0$，$\sum_{k=1}^{K} \alpha_k = 1$
- $\varphi(x|\theta_k)$是第k个子模型的高斯分布密度函数
- γ_{jk}表示第j个观测数据属于第k个子模型的概率

高斯混合模型的概率分布为：$P(x|\theta_k) = \sum_{k=1}^{K} \alpha_k \varphi(x|\theta_k)$，对于这个模型而言，参数

$\theta = (\widetilde{\mu_k}, \widetilde{\theta_k}, \widetilde{\alpha_k})$，也就是每个子模型的期望、方差（或协方差）、在混合模型中发生的概率。

2. EM算法

EM 算法是一种迭代算法，1977 年由 Dempster 等人总结提出，用于含有隐变量（Hidden variable）的概率模型参数的最大似然估计。

每次迭代包含两个步骤：

（1）E-step：求期望 $E(\gamma_{jk}|X,\theta)$ $for\ all\ j = 1,2,\ldots\ldots,N$

（2）M-step：求极大，计算新一轮迭代的模型参数

通过 EM 迭代更新高斯混合模型参数的方法，我们有样本数据$x_1, x_2, \ldots\ldots, x_n$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，首先初始化参数，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：

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

M-step：计算新一轮迭代的模型参数：

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

$$\Sigma_k = \frac{\sum_j^N \gamma_{jk}(x_j - \mu_k)(x_j - \mu_k)^T}{\sum_j^N \gamma_{jk}}, k = 1,2,\ldots,K \quad \text{（用这一轮更新后的 } \mu_k \text{ ）}$$

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

重复计算 E-step 和 M-step 直至收敛。至此，我们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。解决方法是初始化几次不同的参数进行迭代，取结果最好的那次。

Experimental Studies

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，使用EM算法来估计高斯混合模型的参数，并使用这些参数来进行预测，进行模型评估，并解释模型的性能。

根据Methodology一节中的方法，首先取数据集中参数的初始值开始迭代。这些值是EM 算法的初始化参数，初始值的选取对结果有较大影响。本次实验中，我们设置 $\alpha_1 = \alpha_2 = 0.5$，表示男女比例各占 50%；$\mu_1 = 180, \mu_2 = 150$表示男女身高的初始值；$\theta_1 = \theta_2 = 10$表示男女身高的方差。随后循环进行 E 步和 M 步的迭代，实验结果如表1：

表1 实验结果

参数	初始值	收敛值
α_1	180	176.22515
α_2	150	164.204935
θ_1	10	4.879863
θ_2	10	3.096544
μ_1	0.5	0.737068
μ_2	0.5	0.262932

根据给出的初始数据和EM算法迭代值，可以发现男女身高的平均值和方差都有所偏差，男生身高数据偏差不大，但是女生偏差大。尤其是平均值只有小数点后一位的差别。同时，男女的人数比例和方差也有所偏差。在实验中，采用了笔者的想法对参数进行初始化，结果EM算法表现非常不鲁棒，即使误差收敛到很小的值，但明显参数估计是错误的。这说明EM算法对于参数的初值是很敏感的。

Conclusion

本次作业使用EM算法对一组男女身高数据进行了分类，根据对 EM 算法的计算结果，我们可以看到该算法可以用于混合高斯模型的拟合，从而对观测数据进行聚类和分类。

References

[1]https://zhuanlan.zhihu.com/p/304