**Report of Deep Learning for Natural Langauge Processing**

高志磊

gaozhilei@buaa.edu.cn

**Introduction**

极大似然算法可以估计模型参数，但是如果有多个分布，比如已知男生、女生身高分别服从正态分布，但是其身高数据混杂在一起，如何确定男生、女生的分布情况？针对这一问题，需要引入隐变量，隐变量表示男生、女生的比例，随后进行参数估计，这就是EM算法的核心思想。

**Methodology**

**M1: 高斯混合模型**

高斯混合模型（Gaussian Mixed Model，GMM）指的是多个高斯分布函数的线性组合，理论上GMM可以拟合出任意类型的分布，通常用于解决同一集合下的数据包含多个不同分布的情况（或者是同一类分布但参数不一样，或者是不同类型的分布，比如正态分布和伯努利分布）。GMM的概率密度函数如下：



其中，是第个高斯模型的概率密度函数，可以看成选定第个模型后，该模型产生的概率。是第个高斯模型的权重，称作选择第个高斯模型的先验概率，且满足。

**M2: EM算法**

EM是一种用于含有隐变量的模型参数的最大似然估计，是一种迭代算法，通过E步和M步交替迭代直至收敛估计出模型参数。

E步：求第个数据属于第个模型的后验概率：



其中，是每个高斯模型的权重，是第个高斯模型的概率密度函数。

M步：利用所有数据的后验概率重新估计高斯混合模型的参数：



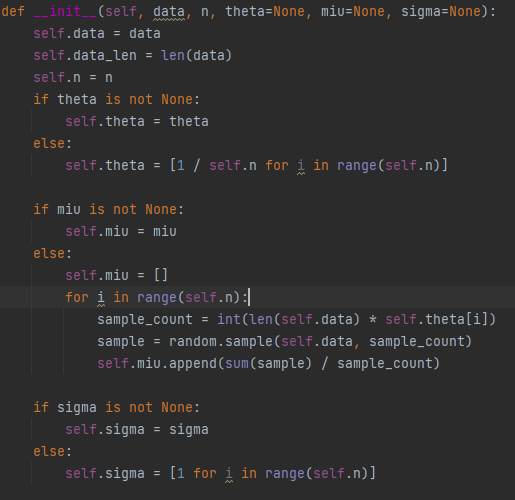




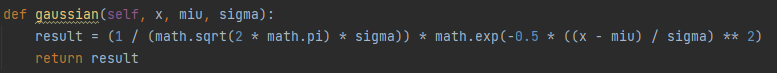
重复计算E步和M步直至收敛。

**M3: 程序实现**

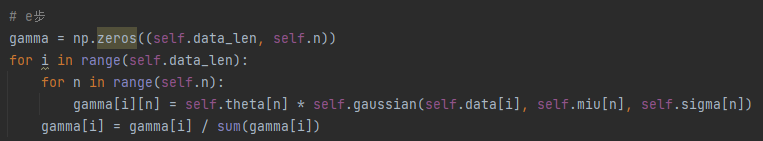
（一）参数初始化，如果指定初始参数，则使用给定的初始参数，如果没有，则假设模型概率相等，所有模型均值为随机抽取一定样本的均值，所有模型方差为1。



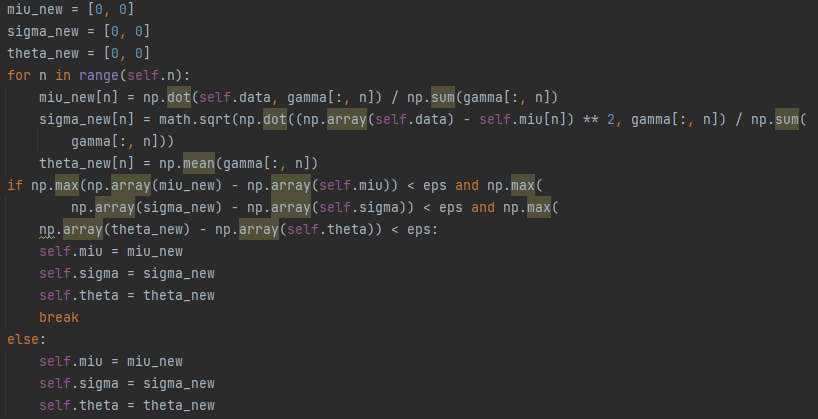
（二）高斯密度函数：



（三）E步：



（四）M步：



**Experimental Studies**

给定2000个身高数据，已知男生身高和女生身高都服从正态分布，使用EM算法估计男女生身高分布的参数。实验结果如表1所示：

Table 1：实验结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| 第一组 | 真实值 | 0.75 | 0.25 | 176 | 164 | 5 | 3 |
| EM算法初始值 | 0.5 | 0.5 | 170 | 160 | 1 | 1 |
| EM算法估计值 | 0.7674 | 0.2326 | 175.8614 | 163.5913 | 5.2711 | 2.9442 |
| 第二组 | 真实值 | 0.75 | 0.25 | 176 | 164 | 5 | 3 |
| EM算法初始值 | 0.5 | 0.5 | 170 | 170 | 1 | 1 |
| EM算法估计值 | 0.5 | 0.5 | 173.0078 | 173.0078 | 7.0859 | 7.0859 |

**Conclusions**

从上述结果可以看出，EM算法并不能保证收敛到全局最优值，算法执行结果与初始值的选取有关。