统计信号处理大作业

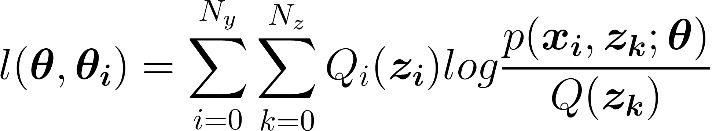
生医81 殷瀚迪 2018013314

**一、查阅资料，学习EM算法和GMM模型的基本知识。并简单阐述EM算法和MLE之间的关系。**

EM算法（Expectation-maximization），又称最大期望算法，是一种迭代算法，用于含有隐变量的概率模型参数的极大似然估计（或极大后验概率估计），分为E步和M步。

Expectation步：





Maximization步：

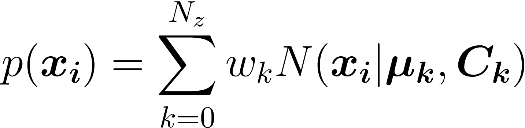


由初值不断迭代以上两个步骤即可求出待估计参数。

EM算法可以解决含有隐变量的MLE估计问题，常用于GMM或HMM的估计。

**二、尝试在理论上推导如何利用EM算法估计GMM模型参数，并给出具体的算法步骤。可以假设最基本的条件，即两个一维高斯分布的混合。并自行设计实验，验证自己算法的正确性。**

GMM模型，高斯混合模型，将观测量的概率模型看作若干个普通高斯概率模型的加权和。换言之，该模型常假设任一观测数据，以不同权重（概率）符合多个不同的高斯分布中的一个：



在使用EM模型对GMM参数进行估计时，隐变量z是每个样本的对应的高斯分布类别，这个类别无法被观察到。

假定该GMM模型由K个高斯分布在概率上加权构成，每个高斯分布的维度是d，每个高斯分布对应一个类别。对于每个属于C的高斯分布，有。在给定的参数下，样本的边际概率为：

表示该样本属于对应的高斯分布。

我们现在来使用EM算法对问题进行估计。承接上文假设，假定该GMM模型由K个高斯分布在概率上加权构成，每个高斯分布的维度是d，共有N个样本，有：

**E步：**

----(1)

样本独立同分布：

----(2)

将(1)式子带入(2)中，对每个样本展开，有：

----(3)

对(3)式中任意i，将P的后验概率展开，有：

----(4)

又有：

----(5)

将(4),(5)带入(3)中，有：

----(6)

由(5)式，项值为1，则有：

---(7)

其中函数是多元高斯分布的概率密度函数。

不妨设为。

**M步：**

现在Q已经确定，我们希望可以通过Q求得下一时刻的参数。

首先求Z的概率分布:

----(8)

这是一个约束优化问题，约束条件为(5)式，使用拉格朗日函数进行求解。

----(9)

求导，令偏导数为0，得到：

----(10)

对所有j求和，得到，从而：

----(11)

得到Z的概率分布后，我们求解组成GMM模型中各高斯分布的均值的更新。

----(12)

将多元高斯分布的概率密度函数展开：

----(13)

无约束，直接求偏导，令偏导数为0，得到：

----(14)

最后，求解协方差的更新。

对Q函数进行整理，有：

----(15)

这里引入矩阵求导法则，必要的求导规则如下：

----(16)

----(17)

----(18)

由(16)(17)(18)式，对(15)式求偏导数，令其为0，解得：

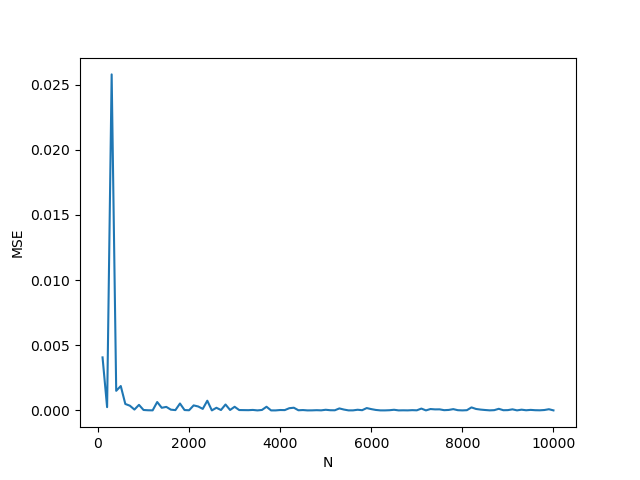
----(19)

观察形式，与 相当一致，充满美感。

我们已经做完了全部的推导。在算法进行中，只需按照如下的步骤迭代：

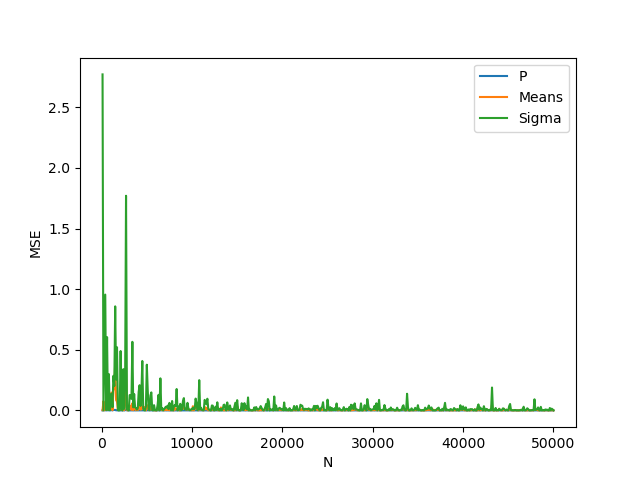
下面我们来做个检验。使用python生成了两个一维高斯的混合，即k=2,d=1,两个高斯过程的样本量之比为2:1，对样本进行shuffle，改变总点数，观察EM算法对两个高斯过程各估计量与预设值的均方误差：

下图为对概率分布的估计：



因为每次生成的数据完全随机，因此推测N=100-200处出现的误差来源是样本的随机性，但可以看到，随着样本数量N的增加，MSE在逐渐较小，对于高斯分布的概率估计越来越准确。

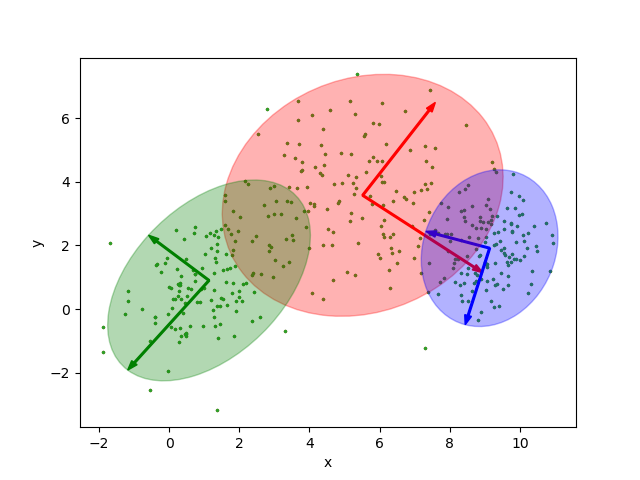
下图为概率，均值和标准差的均方误差随着样本量增加的变化趋势：



可以看到，对概率分布和均值的估计误差很小，并且随着点数增多误差越来越小；标准差的估计上出现了明显的震荡，推测原因与每次生成的样本的随机性有关，但可以观察到，震荡越来越小，误差趋于稳定

另外，选取10000作为固定样本量，可以发现当两个高斯过程的可分性（重叠范围）越高，EM算法的效果越好。

**三、提交需要在实验报告中写明最后求得的模型参数，并以可视化的形式给出数据分类结果。**



求得的参数如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| k | 0.33652494 | | 0.34510405 | | 0.31837101 | |
| MEANS | 5.51747305 | | 1.13615956 | | 9.13775339 | |
| 3.57423868 | | 0.90138959 | | 1.91416631 | |
| COVS | 2.6761569 | -0.38187242 | 1.39419546 | 0.6808004 | 0.63427932 | 0.11931604 |
| -0.38187242 | 2.40937331 | 0.6808004 | 1.66042151 | 0.11931604 | 1.01097545 |