NLP 大作业——EM 算法

学院： 自动化科学与电气工程学院 **姓名：**王明贤 **学号：**ZY2103526

一、EM 算法简介

EM（Expectation-Maximum）算法即期望最大化算法，曾入选“数据挖掘十大算法”。EM 算法是最常见的隐变量估计方法，常被用来学习高斯混合模型（Gaussian mixture model）的参数；隐式马尔科夫算法（HMM）、LDA 主题模型的变分推断等等。

二、问题描述与分析

1.问题描述：

一个袋子中三种硬币的混合比例为：s1, s2 与 1-s1-s2 (0<=si<=1), 三种硬币掷出正面的概率分别为：p, q, r。 （1）自己指定系数 s1, s2, p, q, r，生成 N 个投掷硬币的结果（由 01 构成的序列，其中 1 为正面，0 为反面），利用 EM 算法来对参数进行估计并与预先假定的参数进行比较。 截至日期：4 月 22 日晚 12 点前

2.问题分析

每个硬币抛掷结果均服从伯努利分布，抛掷为正面记为 1，抛掷为反面记为 0。对于 n 个观测样本 $x=(x_1,x_2,\cdots,x_n)$，我们不知道隐含变量性别 $z=(z_1,z_2,\cdots,z_n)$，此时模型参数的对数似然函数为

$$\hat{\theta}=\arg\max\sum_{i=1}^{n}\log\sum_{z_i}p(x_i,z_i\mid\theta)$$

$$p(x_i,z_i\mid\theta)=p(x_i\mid z_i,\theta)p(z_i\mid\theta)$$

$$\hat{\theta}=\arg\max\sum_{i=1}^{n}\log(s_1p^x(1-p)^{1-x}+s_2q^x(1-q)^{1-x}+(1-s_1-s_2)r^x(1-r)^{1-x})$$

我们可以给定估计参数初值，迭代求解逼近真实值。

硬币为第一类硬币：

$$\alpha=\sum_{i=1}^{n}\frac{s_1 p^{x_i}(1-p)^{1-x_i}}{s_1 p^{x_i}(1-p)^{1-x_i}+s_2 q^{x_i}(1-q)^{1-x_i}+(1-s_1-s_2)r^{x_i}(1-r)^{1-x_i}}$$

硬币为第二类硬币：

$$\beta=\sum_{i=0}^{n}\frac{s_2 q^{x_i}(1-q)^{1-x_i}}{s_1 p^{x_i}(1-p)^{1-x_i}+s_2 q^{x_i}(1-q)^{1-x_i}+(1-s_1-s_2)r^{x_i}(1-r)^{1-x_i}}$$

硬币为第三类硬币：

$$\gamma=\sum_{i=0}^{n}\frac{(1-s_1-s_2)r^{x_i}(1-r)^{1-x_i}}{s_1 p^{x_i}(1-p)^{1-x_i}+s_2 q^{x_i}(1-q)^{1-x_i}+(1-s_1-s_2)r^{x_i}(1-r)^{1-x_i}}$$

三、算法设计

该问题是一个混合高斯模型求解问题，可以使用 EM 算法估计隐含量。

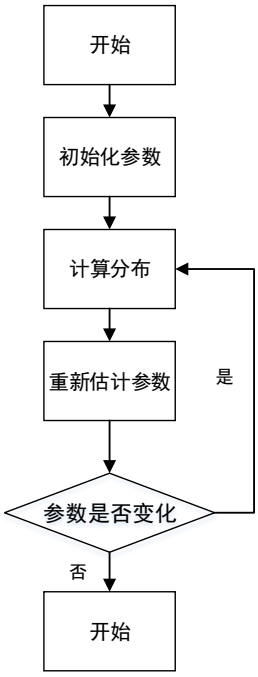

图 1

如图 1 所示，该算法分为以下 4 步：

（1）初始化参数：先初始化三类硬币抛掷正面的概率如：p=0.6,q=0.4,r=0.5，s1=0.3,s2=0.3,s3=0.4；

（2）计算每类硬币所占比例期望；

（3）通过极大似然估计更新硬币抛掷概率的估计。

（4）这时候三类硬币的概率分布进行了更新，然后重复步骤（1）至（3），直到参数不发生变化为止。

四、运行结果

1.运行结果

表 1 生成数据的真实参数

s1	s2	p	q	r
0.3	0.3	0.6	0.4	0.5

表 2 数据初值设置

s1	s2	p	q	r
0.2	0.3	0.55	0.35	0.6

设定迭代 1000 次终止，三种硬币抛掷 pqr 变化曲线如下图所示。

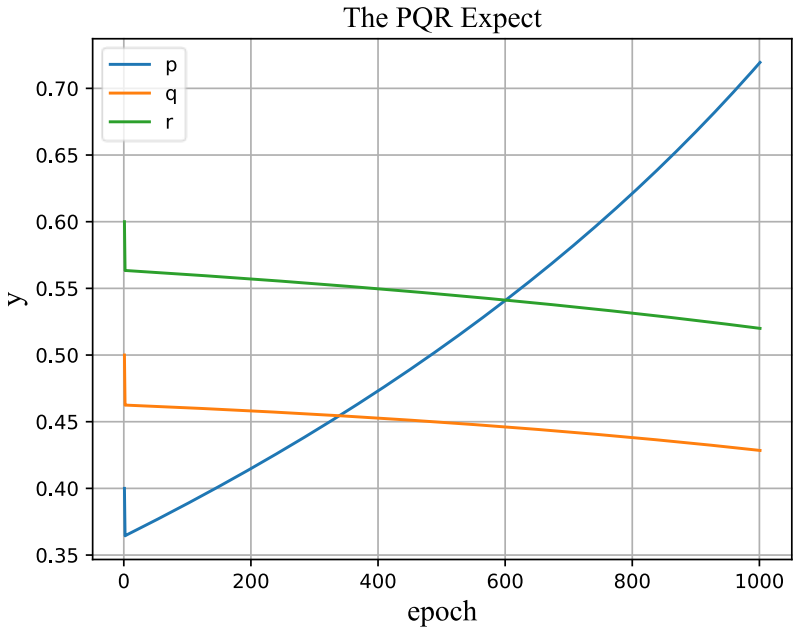

图 2PQR 变化曲线

最终得到各个参数的估计值如表 2。

表 3 估计数据参数

s1	s2	p	q	r
0.101	0.411	0.720	0.428	0.520

2. 结果分析

数据规模和迭代次数对 EM 算法估值有较大影响，更多的数据有利于迭代调整。同时 EM 算法对初始值较为敏感，虽然可以保证收敛，但是可能收敛的不是全局最优点。但是在本次实验中，最终算法似乎没有收敛，算法程序问题经排查暂未发现。

五、总结体会

该问题是一个混合伯努利模型，通过 EM 算法对其隐含变量进行了估计。通过自己生成数据进行实验对比。该算法是一个很经典的隐变量估计方法，通过这次大作业，巩固了课程的算法，同时通过查阅资料，也了解了该算法的局限性。