

深度学习与自然语言处理

EM算法估计参数

|  |  |
| --- | --- |
| 学院名称 | 自动化科学与电气工程学院 |
| 学生学号 | ZY2103809 |
| 学生姓名 | 王海腾 |
| 指导老师 | 秦曾昌 |

2022年4月

# 任务描述

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）

（2）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。

# 实验原理

## EM算法

对于n个样本观察数据，找出样本的模型参数θ, 极大化模型分布的对数似然函数如下：

如果我们得到的观察数据有未观察到的隐含数据 ，即上文中每个样本属于哪个分布是未知的，此时我们极大化模型分布的对数似然函数如下：

## 算法步骤

（1）随机初始化模型参数的初值

（2）j=1,2,...,J 开始EM算法迭代：

E步：计算联合分布的条件概率期望：

M步：极大化 ,得到:

如果参数收敛，则算法结束。否则继续进行E步和M步进行迭代。

# 实验步骤

## 生成硬币序列

采用𝑁×L的二维数组储存生成的数据。其中包括𝑁个硬币，每个硬币投掷L次.使用seed保证实验可复现性，数组中1表示为正面，0表示为反面。

## E步

E步，计算每个样本是硬币1、硬币2、硬币3所掷出的后验概率，分别用μ1，μ2和μ3表示

## M步

在M步中需要对参数进行一次极大似然估计，实现参数迭代。在本次作业中，需要迭代的参数为。通过一下公式计算极大似然估计：

# 实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 硬币个数N | 投掷次数L | 设置(s1,s2,p,q,r) | 初始(p,q,r) | 预测(s1,s2,p,q,r) |
| 1000 | 100 | (0.2, 0.3, 0.2, 0.8, 0.6 ) | (0.1, 0.9, 0.5) | (0.185,0.294, 0.201,0.80100,0.597) |
| 1000 | 100 | (0.1, 0.4, 0.6, 0.3, 0.9) | (0.5, 0.2, 0.8) | (0.10,0.407, 0.594,0.30000,0.898) |
| 10 | 100 | (0.2, 0.3, 0.2, 0.8, 0.6 ) | (0.1, 0.9, 0.5) | (0.2, 0.43 0.15, 0.788, 0.6015) |
| 1000 | 10 | (0.2, 0.3, 0.2, 0.8, 0.6 ) | (0.1, 0.9, 0.5) | (0.271,0.364, 0.273,0.775,0.642) |
| 1000 | 100 | (0.2, 0.3, 0.2, 0.8, 0.6 ) | (0.5, 0.7, 0.7) | (0.293,0.353, 0.286,0.717,0.717) |

所做实验如上表所示，分别修改了硬币个数，硬币投掷次数，所设置的参数分布，初始化参数进行预测。所得结论如下

1. 硬币个数和投掷次数需要较大才可以得到较好的预测结果，当硬币个数较少时，会对硬币分布的预测产生较大影响。
2. 参数的初始化会对预测结果产生较大的影响。初始的概率值需要与所设置的概率值在较近的范围内，否则不能得到正确预测结果。