机器学习与自然语言处理实验二

院（系）名称	自动化科学与电气工程学院
学生姓名	赵怡然
学生学号	ZY2103208
指导老师	秦曾昌

2022 年 04 月 21 日

一、 实验背景介绍

EM 算法即期望最大化算法(expection maximization algorithm)是一种迭代算法，作为一种数据添加算法，在目前的 DL 算法中被广泛运用。

二、 实验目标

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。

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。

三、 EM 算法与抛硬币模型

3.1 EM 算法的数学原理

1. 琴生不等式：

若 ϕ(x)为严格凹函数，x1、x2、…、xn 为定义域内的 n 个点，a1、a2、…、a3 是 n 个正实数，且满足 $\sum_{i=1}^{n} a_i = 1$则下述不等式成立：

$$\phi(\sum_{i=1}^{n} a_i x_i) >= \sum_{i=1}^{n} a_i \phi(x_i)$$

当且仅当 x1=x2=…=xn 时，不等式取等号。

2. EM 算法推导

对于一个似然函数：

$$L(\theta|x^{(1)},x^{(2)},...,x^{(m)}) = \sum_{i=1}^{m} log(P(x^{(i)};\theta)) \\ = \sum_{i=1}^{m} log(\sum_{z^{(i)}} P(x^{(i)},z^{(i)};\theta))$$

直接通过求导极大化比较困难，因此借助琴生不等式，可以得到：

$$log(\sum_{z^{(i)}} P(x^{(i)},z^{(i)};\theta)) = log(\sum_{z^{(i)}} Q^i(z^{(i)}) \frac{P(x^{(i)},z^{(i)};\theta)}{Q^i(z^{(i)})}) \\ >= \sum_{z^{(i)}} Q^i(z^{(i)}) log(\frac{P(x^{(i)},z^{(i)};\theta)}{Q^i(z^{(i)})})$$

其中$Q^i(z^{(i)})$满足$\sum_{z^{(i)}} Q^i(z^{(i)}) = 1$且$Q^i(z^{(i)}) >= 0$，由取等条件可以得到：

$$\frac{P(x^{(i)},z^{(i)};\theta)}{Q^i(z^{(i)})} = k \\ \Rightarrow \quad P(x^{(i)},z^{(i)};\theta) = kQ^i(z^{(i)}) \\ \Rightarrow \quad \sum_{z^{(i)}} P(x^{(i)},z^{(i)};\theta) = k\sum_{z^{(i)}} Q^i(z^{(i)}) \\ \Rightarrow \quad P(x^{(i)};\theta) = k \\ \Rightarrow \quad Q^i(z^{(i)}) = \frac{P(x^{(i)},z^{(i)};\theta)}{P(x^{(i)};\theta)} = P(z^{(i)}|x^{(i)};\theta)$$

其中 k 为常数，由 Q 函数表达式可知，这是一个关于参数的后验分布函数。原有的似然函数可以写为：

$$L(\theta|x^{(1)},x^{(2)},...,x^{(m)}) = \sum_{i=1}^{m} log(\sum_{z^{(i)}} P(x^{(i)},z^{(i)};\theta)) \\ = \sum_{i=1}^{m} log(\sum_{z^{(i)}} \frac{P(x^{(i)},z^{(i)};\theta)}{P(z^{(i)}|x^{(i)};\theta)} P(z^{(i)}|x^{(i)};\theta)) \\ = \sum_{i=1}^{m} \sum_{z^{(i)}} P(z^{(i)}|x^{(i)};\theta) log \frac{P(x^{(i)},z^{(i)};\theta)}{P(z^{(i)}|x^{(i)};\theta)} \\ \geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q^i(z^{(i)}) log \frac{P(x^{(i)},z^{(i)};\theta)}{Q^i(z^{(i)})}$$

通过上述分析，提供了一种不直接极大化似然函数本身而是极大化似然函数的下底由$Q^i(z^{(i)})$组成的函数 $\sum_{i=1}^{m}\mathbb{E}_{Q^i(z^{(i)})}\left[log\frac{P(x^{(i)},z^{(i)};\theta)}{Q^i(z^{(i)})}\right]$的方法。

由初始参数的到函数$Q^i(z^{(i)})$，然后通过极大化由$Q^i(z^{(i)})$组成的下底函数得到新的参数，再由得到的新参数重复上述步骤及 E-STEP 和 M-STEP。通过不断调整$Q^i(z^{(i)})$函数使得下底上升，最后估计参数收敛与真实值。

3.2 抛硬币模型与 EM 算法

题目：

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。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。

变量：

参数有 s1, s2, p, q, r 共 5 种参数。实验已知每次硬币正反的实验结果，未知所抛硬币的种类 Z（A、B、C 三种）。

E-STEP：

$$Q(Z=A)=\frac{s_1*p^x*(1-p)^{n-x}}{s_1*p^x*(1-p)^{n-x}+s_2*q^x*(1-q)^{n-x}+(1-s_1-s_2)*r^x*(1-r)^{n-x}}$$

$$Q(Z=B)=\frac{s_2*q^x*(1-q)^{n-x}}{s_1*p^x*(1-p)^{n-x}+s_2*q^x*(1-q)^{n-x}+(1-s_1-s_2)*r^x*(1-r)^{n-x}}$$

$$Q(Z=C)=1-Q(Z=A)-Q(Z=b)$$

其中 n 为每枚硬币抛出的次数。

M-STEP：

通过下底函数求导可得：

$$s1 = \frac{\sum_{i=1}^{m} Q^i(Z^{(i)} = A)}{m}$$

$$s2 = \frac{\sum_{i=1}^{m} Q^i(Z^{(i)} = B)}{m}$$

$$p = \frac{\sum_{i=1}^{m} Q^i(Z^{(i)} = A)x^{(i)}}{n\sum_{i=1}^{m} Q^i(Z^{(i)} = A)}$$

$$q = \frac{\sum_{i=1}^{m} Q^i(Z^{(i)} = B)x^{(i)}}{n\sum_{i=1}^{m} Q^i(Z^{(i)} = B)}$$

$$p = \frac{\sum_{i=1}^{m} Q^i(Z^{(i)} = C)x^{(i)}}{n\sum_{i=1}^{m} Q^i(Z^{(i)} = C)}$$

四、 实验结果与分析

根据上述 EM 算法原理和相应的抛硬币模型。采用 Python 完成实验数据的生成以及 EM 算法迭代估计，对比了估计参数与真实参数之间的差异。同时分析了不同实验次数以及每枚硬币抛出次数情况下精准度的差异。

为控制变量，初始参数与目标参数均设为相同。

初始参数：s1=0.2, s2=0.5, p=0.1, q=0.3, r=0.5。

目标参数：s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75。

4.1 总实验次数为 1000 次

4.1.1 每枚硬币抛出次数为 1 次

实验次数: 1000. 每个硬币抛出次数为: 1. 总迭代次数为: 5。

估计参数为:

s1:0.1403492647058835,

s2:0.4864430147058773,

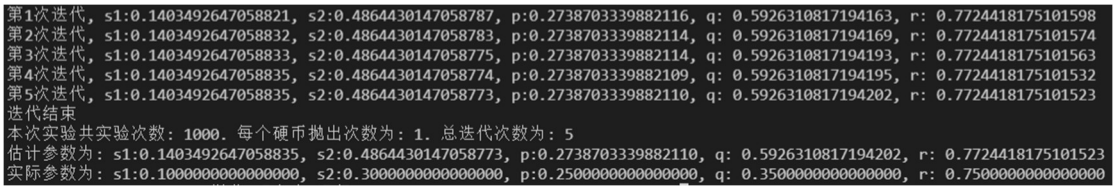
p:0.2738703339882110,

q: 0.5926310817194202,

r: 0.7724418175101523

实际参数为：

s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75



4.1.2 每枚硬币抛出次数为 10 次

实验次数：1000. 每个硬币抛出次数为：10. 总迭代次数为：5514。

估计参数为:

s1:0.2105960005738045,

s2:0.2164117061808579,

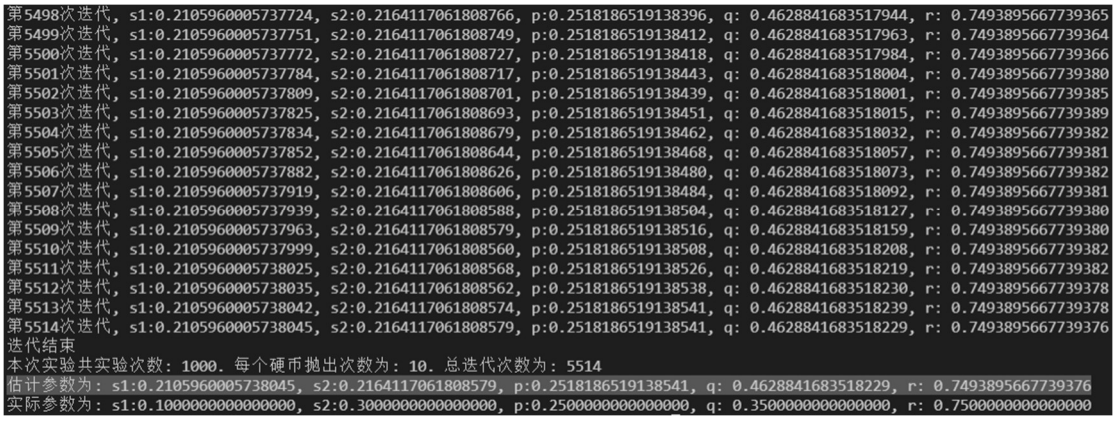
p:0.2518186519138541,

q: 0.4628841683518229,

r: 0.7493895667739376

实际参数为：

s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75



4.1.3 每枚硬币抛出次数为 100 次

实验结果：

实验次数：1000. 每个硬币抛出次数为：100. 总迭代次数为：377。

估计参数为:

s1:0.1248450826529275,

s2:0.2801549344349364,

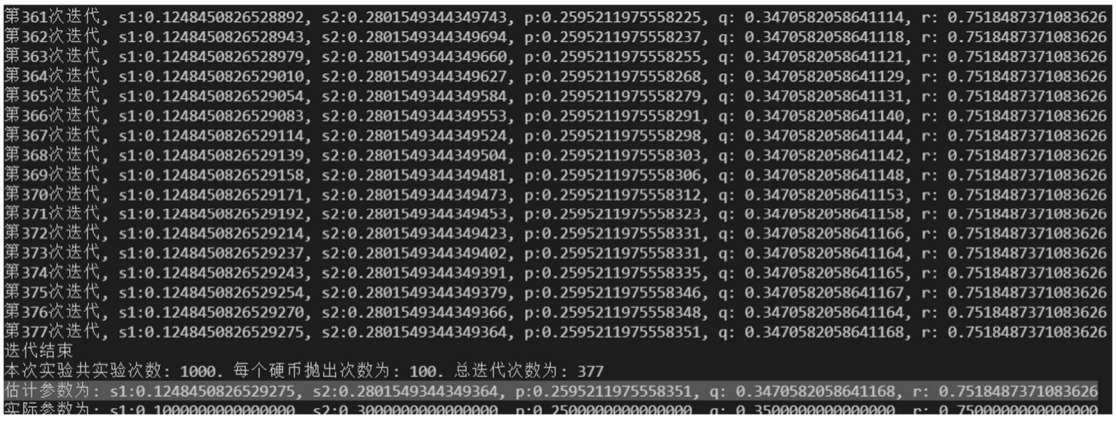
p:0.2595211975558351,

q: 0.3470582058641168,

r: 0.7518487371083626

实际参数为：

s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75



4.2 总实验次数为 100 次

4.2.1 每枚硬币抛出次数为 1 次

实验次数：100. 每个硬币抛出次数为：1. 总迭代次数为：2。

估计参数为:

s1:0.1373161764705884

s2:0.4857536764705885

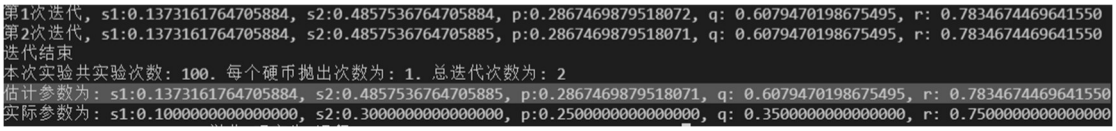
p:0.2867469879518071

q: 0.6079470198675495

r: 0.7834674469641550

实际参数为：

s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75



4.2.2 每枚硬币抛出次数为 10 次

实验次数：100. 每个硬币抛出次数为：10. 总迭代次数为：4382。

估计参数为:

s1:0.2903668728569482

s2:0.2100830819739999

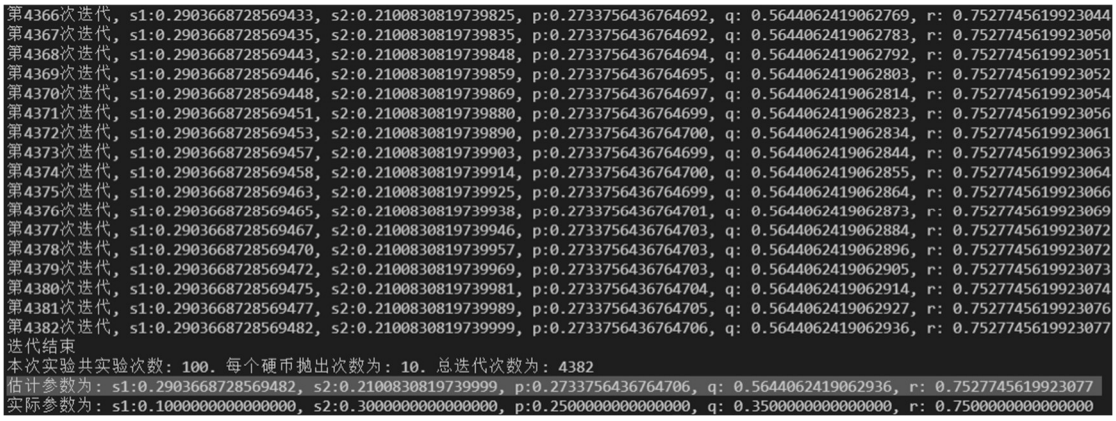
p:0.2733756436764706

q: 0.5644062419062936

r: 0.7527745619923077

实际参数为：

s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75



4.2.3 每枚硬币抛出次数为 100 次

实验结果：

实验次数：100. 每个硬币抛出次数为：100. 总迭代次数为：898。

估计参数为:

s1:0.0944890923710993

s2:0.3955114127933256

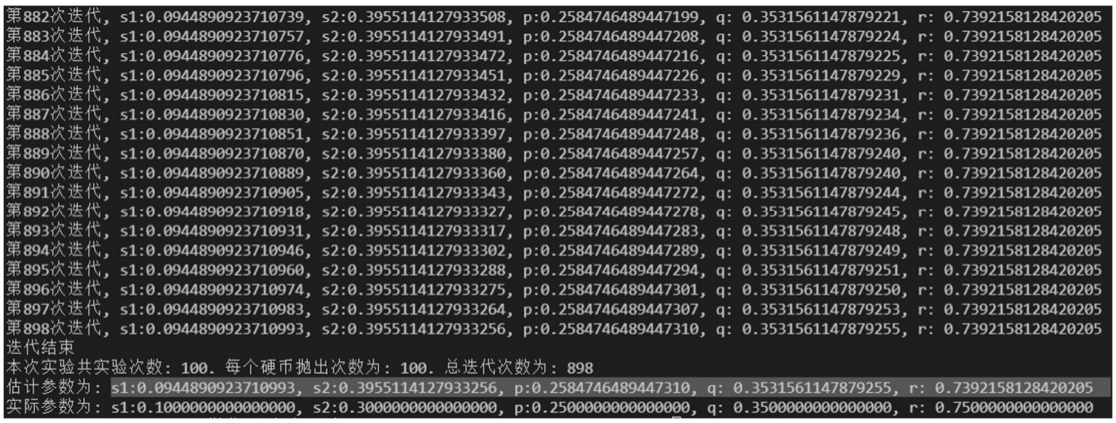
p:0.2584746489447310

q: 0.3531561147879255

r: 0.7392158128420205

实际参数为：

s1=0.1, s2=0.3, p=0.25, q=0.35, r=0.75



4.3 结果分析

通过上述结果可知，随着每枚硬币抛出次数的上升，估计参数与实际参数吻合度逐渐提高。实验次数及样本数提高对吻合度有略微提升。

结果原因分析：

随着样本提供的信息增加及试验次数和每枚硬币抛出次数，EM 算法能更好的估计硬币模型的参数。每枚硬币抛出一次时，估计参数与实际参数有较大的差异，是由于实验数据提供的过少导致；及多种参数下均能较好的拟合所提供的数据样本。