北京航空航天大學
BEIHANG UNIVERSITY

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

1 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。

2 实验原理

2.1 EM 算法

对于 n 个样本观察数据$x = (x_1, x_2, \ldots, x_n)$，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：

$$\hat{\theta} = argmax \sum_{i=1}^{n} logp(x_i; \theta)$$

如果我们得到的观察数据有未观察到的隐含数据$z = (z_1, z_2, \ldots, z_n)$，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：

$$\hat{\theta} = argmax \sum_{i=1}^{n} logp(x_i; \theta) = argmax \sum_{i=1}^{n} log \sum_{z_i} p(x_i, z_i; \theta)$$

2.2 算法步骤

（1）随机初始化模型参数θ的初值θ_0

（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：

$$Q_i(z_i) = p(z_i | x_i, \theta)$$

$$l(\theta, \theta_j) = \sum_{i=1}^{n} \sum_{z_i} Q_i(z_i) log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}$$

M 步：极大化$l(\theta, \theta_j)$,得到θ_{j+1}:

$$\theta_{j+1} = argmax\ l(\theta, \theta_j)$$

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。

3 实验步骤

3.1 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。

3.2 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 μ_1，μ_2 和 μ_3 表示

$$\mu_1(x_i) = \frac{p^{x_i}(1-p)^{1-x_i}}{p^{x_i}(1-p)^{1-x_i} + q^{x_i}(1-q)^{1-x_i} + r^{x_i}(1-r)^{1-x_i}}$$

$$\mu_2(x_i) = \frac{q^{x_i}(1-r)^{1-x_i}}{p^{x_i}(1-p)^{1-x_i} + q^{x_i}(1-q)^{1-x_i} + r^{x_i}(1-r)^{1-x_i}}$$

$$\mu_3(x_i) = \frac{r^{x_i}(1-r)^{1-x_i}}{p^{x_i}(1-p)^{1-x_i} + q^{x_i}(1-q)^{1-x_i} + r^{x_i}(1-r)^{1-x_i}}$$

3.3 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$\theta = [s_1，s_2，p，q，r]$。通过一下公式计算极大似然估计：

$$s_1 = \frac{\sum_{i=1}^{N} \mu_1(x_i)}{N}$$

$$s_2 = \frac{\sum_{i=1}^{N} \mu_2(x_i)}{N}$$

$$p = \frac{\sum_{i=1}^{N} (\mu_1(x_i) x_i)}{S \sum_{i=1}^{N} \mu_1(x_i)}$$

$$q=\frac{\sum_{i=1}^{N}(\mu_2(x_i)x_i)}{S\sum_{i=1}^{N}\mu_2(x_i)}$$

$$r=\frac{\sum_{i=1}^{N}(\mu_3(x_i)x_i)}{S\sum_{i=1}^{N}\mu_3(x_i)}$$

4 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。