单位代码
学　　号 ZY2103116
分 类 号

北京航空航天大學
BEIHANG UNIVERSITY

EM 算法解决混合硬币的参数估计

深度学习与自然语言处理（NLP）第二次课后作业

院（系）名称　自动化科学与电气工程学院
专 业 名 称　自动化
学 生 姓 名　杨卓昆

2022 年 04 月

EM 算法解决混合硬币的参数估计

深度学习与自然语言处理（NLP）第二次课后作业

杨卓昆 自动化科学与电气工程学院 ZY2103116
yangzhuokun1998@buaa.edu.cn

摘要

本次研究给定三元伯努利混合模型的参数并生成了 200 个投掷硬币结果数据，利用 EM 算法来估计参数并验证。同时，本研究探究参数不同的初始状态对最终结果的影响， 分析了不同参数对实验收敛结果的影响。

目录

1 内容介绍

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。

首先自己给定混合硬币模型的参数，并生成数据，利用 EM 算法来估计参数并验证。最大似然估计算法是一种通过搜索整个概率分布及其参数从而完成对数据集的密度估计的方法。这种方法普遍而有效，是许多机器学习算法的基础，但是这种方法要求训练数据集是完整的，即所有的相关的随机变量都存在。如果存在与数据集中的变量相互作用但被隐藏或未被观察到的变量，即所谓的潜在变量，最大似然法就难以解决问题。此时我们需要引入期望最大化算法，即 EM 算法。EM 算法是一种在存在潜伏变量的情况下进行最大似然估计的方法。它通过首先估计潜在变量的值，然后优化模型，然后重复这两个步骤直到收敛。这是一种有效的通用方法，最常用于有缺失数据的密度估计，例如伯努利混合模型等聚类算法。混合模型是一个由多个概率分布函数组成的模型。对于一个混合模型，我们可以通过统计和学习算法来估计概率分布的参数。

2 实验原理

2.1 伯努利分布

2.1.1 伯努利分布定义

一个非常简单的试验是只有两个可能结果的试验，比如正面或反面，成功或失败，有缺陷或没有缺陷，病人康复或未康复。为方便起见，记这两个可能的结果为 0 和 1，下面的定义就是建立在这类试验基础之上的。

如果随机变量X只取 0 和 1 两个值，并且相应的概率为：

$$Pr(X=1)=p, Pr(X=0)=1-p, 0<p<1$$

则称随机变量X服从参数为p的伯努利分布，若令$q=1-p$，则X的概率函数可写：

$$f(x \mid p)=\begin{cases} p^x q^{1-x}, & \text{x=0,1;} \\ 0, & x \neq 0,1. \end{cases}$$

如果X服从参数为p的伯努利分布，则有：

$$E(X) = 1 \cdot p + 0 \cdot q = p.$$
$$E\left(X^2\right) = 1^2 \cdot p + 0^2 \cdot q = p.$$

$$Var(X) = E\left(X^2\right) - [E(X)]^2 = pq.$$

$$\psi(t) = E\left(e^{tX}\right) = pe^t + q, -\infty < t < +\infty.$$

2.1.2 二维伯努利分布

二维伯努利分布是关于二维布尔向量的概率分布，假设有向量如下：

$$\boldsymbol{x} = [x_1, x_2]$$

其中：

$$x_1, x_2 \in \{0, 1\}$$

且满足：

$$x_1 + x_2 = 1$$

设有参数向量：

$$\boldsymbol{\theta} = [\theta_1, \theta_2]$$

$$\theta_1, \theta_2 \in [0, 1]$$
$$\theta_1 + \theta_2 = 1$$

其概率分布函数为：

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{\theta}) &= \theta_1^{x_1}\theta_2^{x_2} \\ &= \theta_1^{x_1}(1-\theta_1)^{1-x_1} \end{aligned}$$

由此可见，二维伯努利也可考虑成一维伯努利去应用，也就是伯努利实验，这里称其为二维是为了与后面介绍的多维伯努利相一致。二维伯努利最简单的应用就是抛硬币问题，一个硬币共有两个面，每次实验只会出现一个面朝上。

2.2 EM 算法

2.2.1 EM 算法介绍

EM 算法是一种在存在潜伏变量的情况下进行最大似然估计的方法。它通过

首先估计潜在变量的值，然后优化模型，然后重复这两个步骤直到收敛。这是一种有效的通用方法，最常用于有缺失数据的密度估计，例如高斯混合模型、伯努利混合模型等聚类算法。对于服从混合分布的数据 Y，数据集由许多点组成，这些点恰好由两个不同的分布产生。每个点都属于一个概率分布，但数据是组合在一起的，分布足够相似，以至于某个点可能属于哪个分布并不明显。可以使用两个分布$\phi_{\theta_1}(x)$和$\phi_{\theta_2}(x)$表示 Y 的概率密度，其中参数为$\theta_i = (\mu_i, {\sigma_i}^2)$，则可以将 Y 的概率密度表示如下：

$$g_Y(y) = (1-\pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y)$$

其中参数为$\theta = (\pi, \mu_1, {\sigma_1}^2, \mu_2, {\sigma_2}^2)$，π 为混合比例。则参数估计的对数似然函数可以表示如下：

$$l(\theta; Z) = \sum_{i=1}^{N} \log[(1-\pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)]$$

其中 N 为训练数据的个数。通过对上述对数似然函数进行变形，可得到如下表示：

$$l(\theta; Z, \alpha) = \sum_{i=1}^{N}[(1-\alpha_i)\log\phi_{\theta_1}(y_i) + \alpha_i\log\phi_{\theta_2}(y_i)] + \sum_{i=1}^{N}[(1-\alpha_i)\log(1-\pi) + \alpha_i\log\pi]$$

其中 α 为隐变量，当α_i取值为 0 时表示Y_i来自于模型一，当α_i取值为 1 时表示Y_i来自于模型二。在这种情况里，$(\mu_1, {\sigma_1}^2)$的极大似然估计是$\alpha_i = 0$的数据样本均值和方差，$(\mu_2, {\sigma_2}^2)$的极大似然估计是$\alpha_i = 1$的数据样本均值和方差。由于α_i的实际值未知，我们使用迭代的方式进行处理，用期望代替α_i，即：

$$\gamma_i(\theta) = E(\alpha_i|\theta, Z) = \Pr(\alpha_i = 1|\theta, Z)$$

下面介绍二分量参数估计的 EM 算法的具体步骤。首先初始化模型参数$\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\pi}$。之后设置 E 步算法如下：

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, i = 1, 2, \ldots N$$

该期望表示数据y_i属于ϕ_{θ_1}的概率，可以看到$\widehat{\gamma}_i$越接近 0 表示y_i越可能属于模型一，反之，可以看到$\widehat{\gamma}_i$越接近 1 表示y_i越可能属于模型二。最后设置 M 步

算法如下所示：

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)y_i}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}, \hat{\mu}_2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i y_i}{\sum_{i=1}^{N}\hat{\gamma}_i}$$

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)(y_i-\hat{\mu}_1)^2}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}, \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i(y_i-\hat{\mu}_2)^2}{\sum_{i=1}^{N}\hat{\gamma}_i}$$

$$\hat{\pi} = \sum_{i=1}^{N}\hat{\gamma}_i / N$$

在 EM 算法中，估计步骤将为每个数据点的过程潜变量估计一个值，而最大化步骤将优化概率分布的参数，以试图最好地捕捉数据的密度。这个过程不断重复，直到有一组好的潜变量值和最大似然值符合数据。E 步骤。估计每个潜变量的期望值。M 步骤。使用最大似然法优化分布的参数。

2.2.2 算法处理伯努利模型原理

对于服从伯努利混合分布的数据 Y，数据集由许多点组成，这些点恰好由三个不同的分布产生。每个点都属于一个伯努利概率分布，但数据是组合在一起的，分布足够相似，以至于某个点可能属于哪个分布并不明显。可以使用三个伯努利分布s_{p_1}和s_{p_2}、s_{p_3}表示 Y 是三重分布的概率，其中参数为p_i，则可以将 Y 的概率密度表示如下：

$$g_Y(y) = \sum s_{p_1}(x) \times p_i$$

其中参数为$\theta = (p1, p2, p3)$，s_{p_1}和s_{p_2}、s_{p_3}为混合比例。则参数估计的似然函数可以表示如下：

$$p(x|\theta) = \prod_i \left[\sum_k s_k * N(x_i|\theta_k)\right]$$

对数极大似然：

$$L(x|\theta) = \sum_i ln\left[p(x_i|\theta_k)\right] = \sum_i ln\left[\sum_k s_k * N(x_i|\theta_k)\right]$$

其中 N 为训练数据的个数。通过对上述对数似然函数进行变形。

在 EM 算法中，估计步骤将为每个数据点的过程潜变量估计一个值，而最大化步骤将优化概率分布的参数，以试图最好地捕捉数据的密度。这个过程不断重复，直到有一组好的潜变量值和最大似然值符合数据。E 步骤。估计每个潜变量的期望值。M 步骤。使用最大似然法优化分布的参数。

EM 算法所设定的隐参量 z 一般属于 1，2，k，…，K。用于描述计算出 K 组伯努利分布模型的参数后某个数据点 xi 属于第 k 个伯努利模型的概率：

将隐函数做为后验概率，可以得到：

$$\mu_k{}^i = p(z_i = k|x_i, \theta_k) = \frac{s_k * N(x_i|\theta_k)}{\sum_{k=1}^{3} s_k * N(x_i|\theta_k)} = \frac{s_k * \theta_k{}^{x_i} * (1-\theta_k)^{1-x_i}}{\sum_{k=1}^{3} s_k * \theta_k{}^{x_i} * (1-\theta_k)^{1-x_i}}$$

引入 Jensen 不等式来简化似然函数可得：

$$L(x|\theta) = \sum_i ln \sum_k \mu_k{}^i \frac{s_k * N(x_i|\theta_k)}{\mu_k{}^i} \geq \sum_i \sum_k \mu_k{}^i * ln \frac{s_k * N(x_i|\theta_k)}{\mu_k{}^i}$$

不等式的右侧给似然函数提供了一个下界。EM 算法提出迭代逼近的方法，不断提高下界，从而逼近似然函数。每次迭代都以下面这个目标函数作为优化目标：

$$Q(\theta, \theta^t) = \sum_i \sum_k \mu_k{}^i{}_t * ln \frac{s_k * N(x_i|\theta_k)}{\mu_k{}^i{}_t}$$

这个式子表示，在第t次迭代后，获得参数θt，然后就可以计算隐参数概率wti,k。将隐参数代回$Q(\theta,\theta t)$,进行最大似然优化，即可求出更优的参数$\theta t+1$。

3 算法实现

输入：观察到的数据样本，最大迭代次数iters_num，判断收敛阈值D。

算法步骤:

（1）随机初始化模型参数 θ 的初值为 θ_0，随机生成距离真实参数距离不过大的初始参数。

（2）开始 EM 算法迭代：

E 步：计算联合分布的条件概率期望：

$$\mu_k{}^i = p(z_i = k|x_i, \theta_k) = \frac{s_k * N(x_i|\theta_k)}{\sum_{k=1}^{3} s_k * N(x_i|\theta_k)} = \frac{s_k * {\theta_k}^{x_i} * (1-\theta_k)^{1-x_i}}{\sum_{k=1}^{3} s_k * {\theta_k}^{x_i} * (1-\theta_k)^{1-x_i}}$$

M 步：极大化参数最大似然估计

$$\Theta := \arg\max_{\Theta} Q(\Theta, \Theta^t)$$

$$s_k^{t+1} = \frac{\sum_i {\mu_k}^i{}_t}{N}$$

$$\theta_k^{t+1} = \frac{\sum_i {\mu_k}^i{}_t * x_i}{\sum_i {\mu_k}^i{}_t}$$

（3）判断是否收敛，计算两次参数似然估计的距离，计算方法如下所示：

$$d(\theta_j, \theta_{j+1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{|\theta_{j,i} - \theta_{j+1,i}|}{|\theta_{j,i} + \theta_{j+1,i}|}$$

若$d(\theta_j, \theta_{j+1}) < d$，则输出 θ_{j+1}为最终估计结果。否则重复步骤（2）。

输出：模型参数 θ，并评估收敛结果，评估指标计算方法如下：

$$d(\theta_0, \theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{|\theta_{0,i} - \theta_i|}{|\theta_{0,i} + \theta_i|}$$

4 实验过程

4.1 生成数据

首先，我们需要生成一个数据集，其中点由三个伯努利过程之一生成。第一个伯努利分布的概率为 0.2，第二个伯努利分布的概率为 0.5，第三个伯努利分布的概率为 0.8。我们选择混合系数为 0.2、0.3，则第三个模型结果出现的概率为 0.5，总共生成 200 个数据，即在第一个伯努利分布中采样 40 个数据，在第二个分布中采样 60 个数据，其余数据由第三个分布产生。具体参数如下所示：

表 4.1 生成数据真实参数

名称	符号	数据		
分布（硬币）	C	分布 1	分布 2	分布 3
概率	P	p = 0.2	q = 0.5	r = 0.8
混合系数	S_i	0.2	0.3	0.5
数量	N_i	40	60	100
总数	N	200		

5 实验结果与分析

5.1 实验结果

通过 EM 算法求出实验预测结果如下表所示：

表 4.2 生成数据结果

名称	符号	数据		
分布（硬币）	C	分布 1	分布 2	分布 3
概率	P	0.205	0.501	0.807
混合系数	S_i	0.217	0.316	0.469
总数	N	200		

表 4.3 生成数据参数预测评价

	$\boldsymbol{\theta}$	$\boldsymbol{\theta_{True}}$	$\boldsymbol{d(\theta_{0,i}, \theta_i)}$
p	0.2	0.205	0.005
q	0.5	0.501	0.001
r	0.8	0.807	0.007
$\boldsymbol{S_1}$	0.2	0.217	0.017
$\boldsymbol{S_2}$	0.3	0.316	0.016
$\boldsymbol{S_2}$	0.5	0.469	0.031
$\boldsymbol{d(\theta_0, \theta)}$	-	-	0.014

可以看到，EM 算法对数据参数进行了较好的预测，距离真实参数误差仅为 0.014 左右。

具体的收敛过程如下所示:

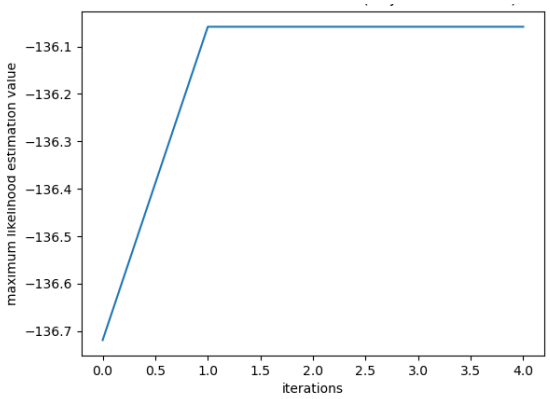

图 4.1 最大似然函数变化曲线

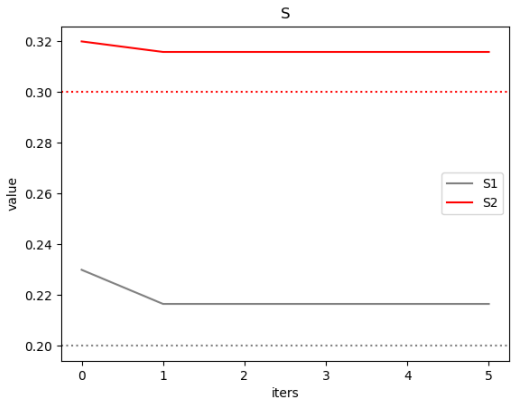

图 4.2 混合系数（S1，S2）变化曲线

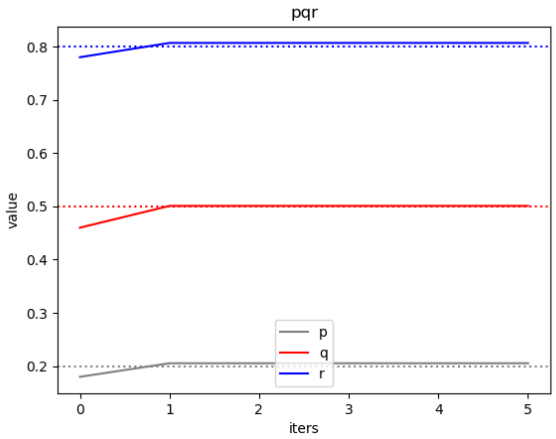

图 4.3 p，q，r 变化曲线

5.2 分析

在实验之中可以发现，程序并不是每一次运行都能得到最佳的收敛结果，这

是由于 EM 算法并不能找到全局最优的结果。虽然 EM 算法在原理上是总能够收敛的，但是其收敛的结果是受初始状态影响的。为了探究参数不同的初始状态对最终结果的影响，本文将进行初始均值、标准差、混合系数对最终收敛结果的影响进行分析。同时可以看到在第一次更新时，参数和目标函数有一个大的变化，之后几次变化就很小几乎接近收敛。并且第一次更新时候很难收敛到真实的值。因此可以得出 EM 算法对三硬币模型参数的求解能力很差，很快陷入局部最优收敛，按照 EM 算法的推导公式，参数在第一次更新后，之后参数变化很小几乎为零。根据公式推导和数学原理，EM 算法本身能保证最终一定收敛，但不能保证收敛到最优。

5.3 探索

为了分析混合系数对实验收敛结果的影响设计如下实验以对 EM 算法性能进行测试。这里采用二类混合模型生成数据以探索初始化参数对于结果的影响。

这里改变初始参数的初始化方法，使得混合系数有着更大的初始化随机范围，而其余参数限制于更加接近真实参数的范围对参数初始化方法如下：

当$\theta_i = \pi$时：

$$\theta_{0,i} = \mathrm{Random}(0,1)$$

当$\theta_i \neq \pi$时：

$$\theta_{0,i} = \theta_{\mathrm{True},i} \times (1 + \mathrm{Random}(-0.1,0.1))$$

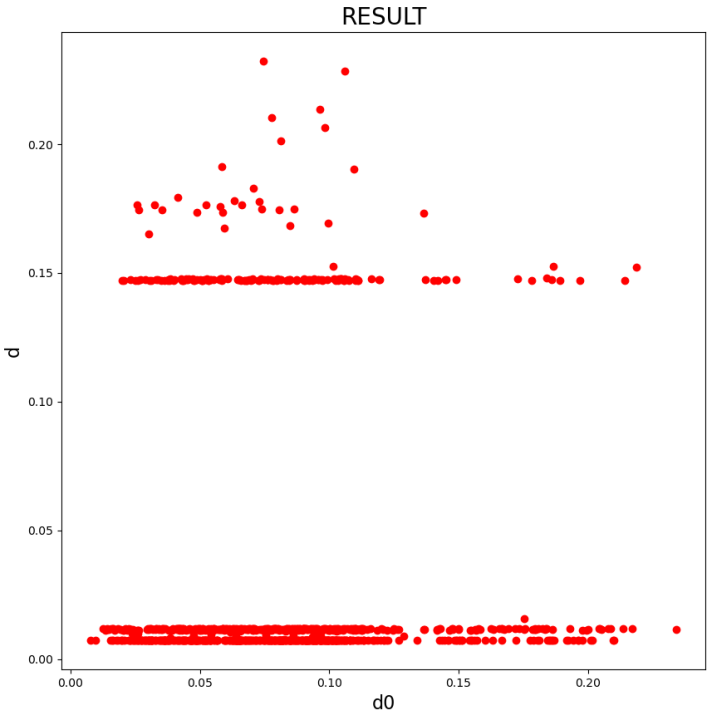

图 5.1 不同初始化混合系数下实验的收敛情况

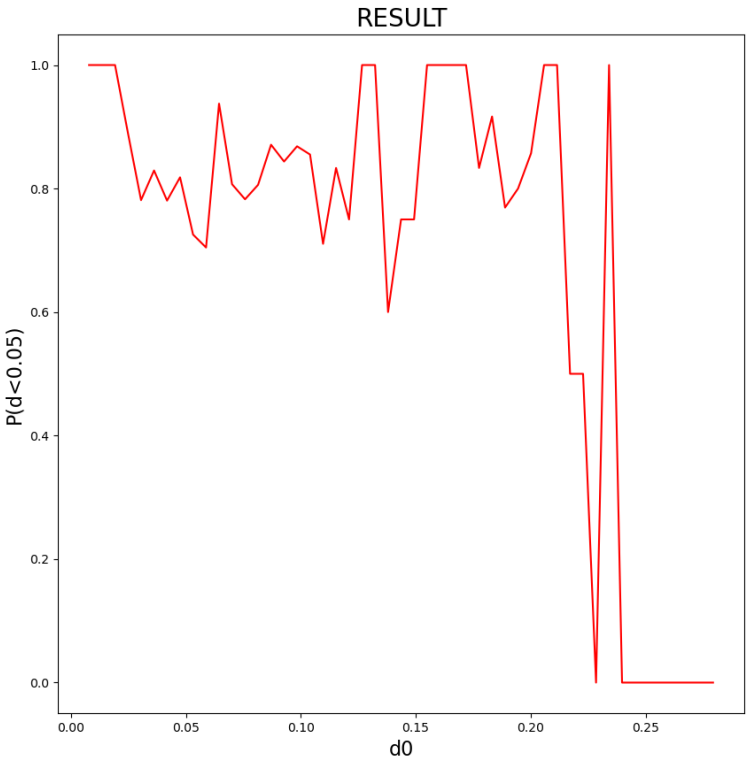

图 5.2 不同初始化混合系数下实验的成功收敛比例

可以看到，收敛结果对于混合系数初始化并不敏感，算法始终具有较高的成

功收敛比例。

传统 EM 算法对初始值敏感，聚类结果随不同的初始参数变化而发生较大波动。总的来说，EM 算法收敛的优劣很大程度上取决于其初始参数。而当参数初始化在距离真实参数距离小于 0.05 时，算法具有较高的成功收敛比例。进一步分析了不同参数对实验收敛结果的影响，这里改变初始参数的初始化方法，使得被考察参数有着更大的初始化随机范围，而未被考察参数限制于更加接近真实参数的随机范围，从而统计分析对于某一个具体参数初始化偏差对真实收敛结果的影响。

6 总结

本次研究给定三元伯努利混合模型的参数并生成数据，利用 EM 算法来估计参数并验证。同时，本研究充分讨论了各种影响因素，探究参数不同的初始状态对最终结果的影响，分析了不同参数对实验收敛结果的影响。

由于 EM 算法容易陷入局部最优，本文探究了参数不同的初始状态对最终结果的影响，进行了对最终收敛结果的影响进行分析。同时研究发现，收敛结果对于均值初始化非常敏感，当参数初始化在距离真实参数距离小于 0.05 时，算法具有较高的成功收敛比例，反之收敛结果与真实结果差异很大。收敛结果对于标准差初始化较为敏感，当参数初始化在距离真实参数距离小于 0.15 时，算法具有较高的成功收敛比例。收敛结果对于混合系数初始化并不敏感，算法始终具有较高的成功收敛比例。

传统 EM 算法对初始值敏感，聚类结果随不同的初始值而波动较大。总的来说，EM 算法收敛的优劣很大程度上取决于其初始参数，其中均值对于收敛结果影响较大。本次实验增强了我对于 EM 算法的掌握，加深了对于聚类算法的理解，提升了对于问题的分析能力与实现程序过程中的设计能力与完成能力。

7 参考文献

[1]https://docs.qq.com/pdf/DZnBvaHVSa3psdXlL
[2]https://zhuanlan.zhihu.com/p/93513123
[3]https://baike.baidu.com/item/%E4%BC%AF%E5%8A%AA%E5%88%A9%E5%88%86%E5%B8%83/7167021?fr=aladdin