随机分布的生成与估计

李一鸣

1160300625

2018 年 9 月 28 日

随机数和随机序列的产生

生成随机序列 $\mathbf{X}=(X_1,X_2,...,X_n)$，其中每个 X_i 服从 $[-\frac{a}{2},\frac{a}{2}]$ 的均匀分布。

生成随机序列 $\mathbf{Y}=(Y_1,Y_2,...,Y_n)$，其中每个 Y_i 服从 $[-\frac{a}{2},\frac{a}{2}]$ 的均匀分布。

蒙特卡罗投点法:

在边长为 a 的正方形内随机投点，设该点落入此正方形的内切圆中的概率为 P_{circle}，则:

$$\begin{aligned} P_{circle} &= \frac{S_{circle}}{S_{square}} \\ &= \frac{\pi(\frac{a}{2})^2}{a^2} \\ &= \frac{\pi}{4} \end{aligned} \tag{1}$$

假定总共生成了 n 个数据，其中有 m 个在圆内，则:

$$f_{circle}=\frac{m}{n} \tag{2}$$

对于任一点 (X_i,Y_i)，如果满足:

$$X_i^2+Y_i^2 \le (\frac{a}{2})^2=\frac{a^2}{4} \tag{3}$$

则其在圆内，计入 m 中。

以频率估计概率，我们有:

$$\pi=\frac{4m}{n} \tag{4}$$

实验结果:

在实验中取 $a=1$（其实 a 取多少都没有关系，精确度只与样本数相关），取 $n=1\,000, 10\,000, 100\,000, 1\,000\,000$ 分别进行实验。

```
n = 1000
m = 778

pi = 3.112
```

```
n = 10000
m = 7853

pi = 3.1412
```

```
n = 100000
m = 78731

pi = 3.14924
```

```
n = 1000000
m = 785167

pi = 3.140668
```

随机分布的计算机模拟

高斯分布的模拟

生成均值为 $\mu = 10$、方差为 $\sigma = 5$ 的正态分布，并画出均值和方差随样本数增加而变化的图。

设样本为 $\mathbf{X}$，总样本数为 N，记前 n 个样本数据的均值、方差分别为 E_n 和 D_n，则得到均值、方差矩阵：

$$
\begin{aligned}
\mathbf{E} = (E_1, E_2, ..., E_N) \\
\mathbf{D} = (D_1, D_2, ..., D_N)
\end{aligned} \tag{5}
$$

样本个数矩阵：

$$
\mathbf{N} = (1, 2, ..., N) \tag{6}
$$

我们只需要作出 $(\mathbf{N}, \mathbf{E})$ 和 $(\mathbf{N}, \mathbf{D})$ 的图像即可。

注意事项：

本来我们可以直接根据 $\mathbf{X}$ 中的前 n 项直接计算 E_n 和 D_n，但在实际问题中数据量往往过大，我们一般不会保存之前的数据，因此我们通常采用递推的计算方式。

1. 均值递推公式

$$\begin{cases} E_{n-1} = \frac{1}{n-1}(X_1 + X_2 + ... + X_{n-1}) \\ \\ E_n = \frac{1}{n}(X_1 + X_2 + ... + X_{n-1} + X_n) \end{cases}$$

解得：

$$\begin{aligned} E_n &= \frac{(n-1)E_{n-1} + X_n}{n} \\ &= \frac{nE_{n-1} - E_{n-1} + X_n}{n} \\ &= E_{n-1} + \frac{X_n - E_{n-1}}{n} \end{aligned} \tag{ps.1}$$

2. 方差递推公式

$$\begin{cases} D_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}(X_i - E_{n-1})^2 \\ \\ D_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - E_n)^2 \end{cases}$$

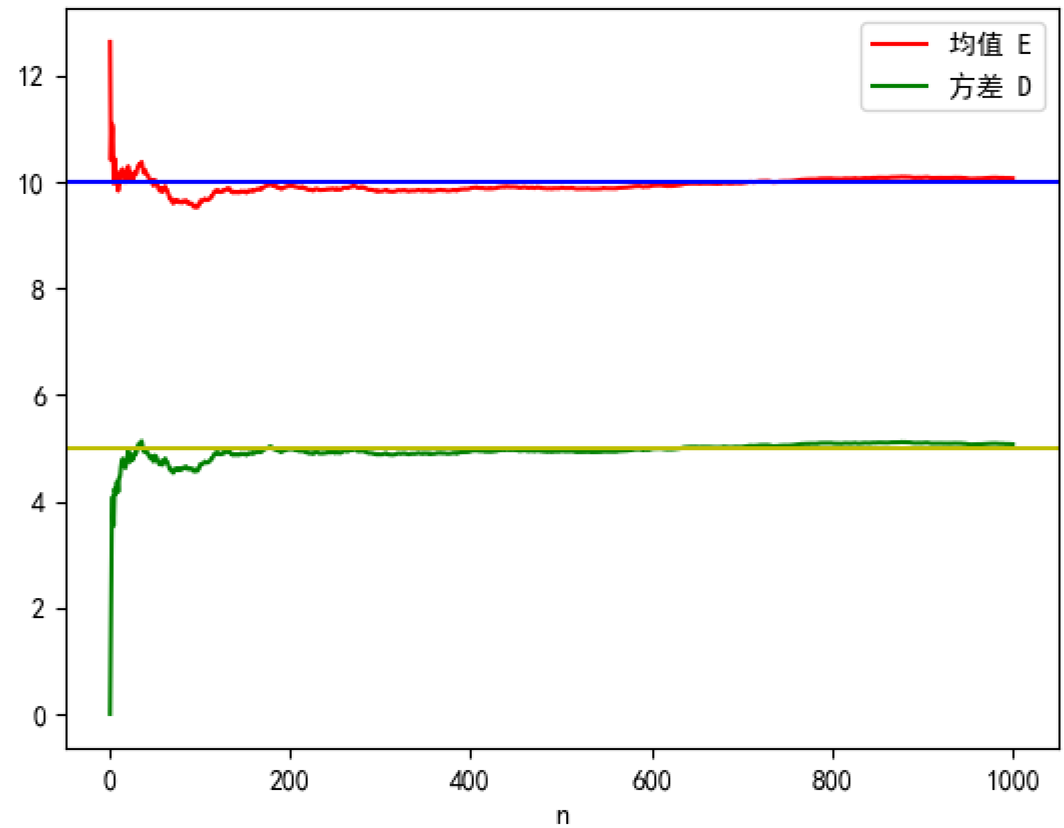
联立式 (ps.1)，得：

$$\begin{aligned} D_n &= \frac{1}{n}\sum_{i=1}^{n}(X_i - E_{n-1} - \frac{X_n - E_{n-1}}{n})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[(X_i - E_{n-1})^2 + (\frac{X_n - E_{n-1}}{n})^2 - 2(X_i - E_{n-1})(\frac{X_n - E_{n-1}}{n})] \\ &= (\frac{X_n - E_{n-1}}{n})^2 + \frac{1}{n}\sum_{i=1}^{n}[(X_i - E_{n-1})^2 - 2(X_i - E_{n-1})(\frac{X_n - E_{n-1}}{n})] \\ &= (\frac{X_n - E_{n-1}}{n})^2 + \frac{1}{n}[(X_n - E_{n-1})^2 - 2(X_n - E_{n-1})(\frac{X_n - E_{n-1}}{n})] \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n-1}[(X_i - E_{n-1})^2 - 2(X_i - E_{n-1})(\frac{X_n - E_{n-1}}{n})] \\ &= \frac{n-1}{n^2}(X_n - E_{n-1})^2 + \frac{1}{n}\sum_{i=1}^{n-1}(X_i - E_{n-1})^2 - 2(\frac{X_n - E_{n-1}}{n})\sum_{i=1}^{n-1}(X_i - E_{n-1}) \\ &= \frac{n-1}{n^2}(X_n - E_{n-1})^2 + \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n-1}(X_i - E_{n-1})^2) \\ &= \frac{n-1}{n^2}(X_n - D_{n-1})^2 + \frac{n-1}{n}D_{n-1} \end{aligned} \tag{ps.2}$$

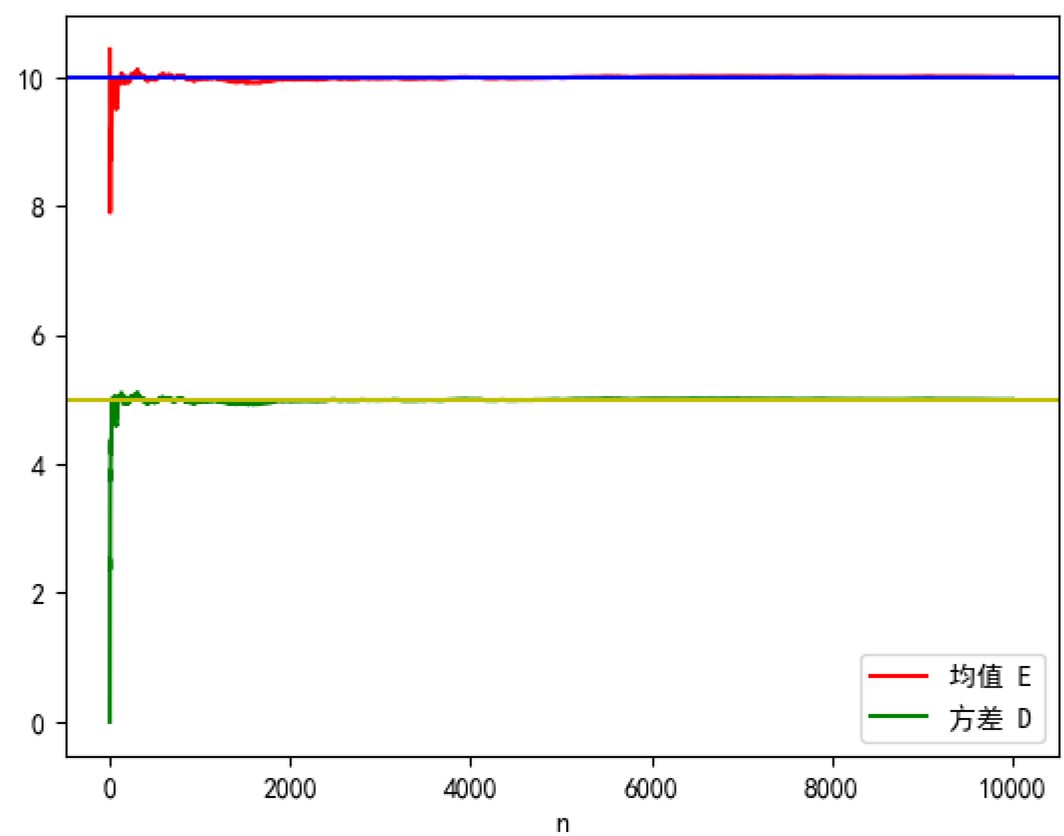
实验结果：

在实验中取 $a = 1$，取 $N = 1\,000, 10\,000$ 分别进行实验。

正态分布样本的均值、方差随样本数增加而变化的图像
$\mu = 10.0, \sigma^2 = 5.0, N = 1000$



正态分布样本的均值、方差随样本数增加而变化的图像
$\mu = 10.0, \sigma^2 = 5.0, N = 10000$



敌军坦克到达情况的模拟

敌军坦克分队到达我方阵地规律服从泊松分布，平均每分钟到达 λ 辆。

泊松分布的期望值是 λ，也就是说在一分钟之内，到达的坦克数量 T 的分布列为：

$$P(T=k)=\frac{\lambda^k e^{-\lambda}}{k!} \tag{7}$$

我们可以生成 N 组数据 $\mathbf{T_1},\mathbf{T_2},...,\mathbf{T_N}$，分别用它们的均值 $E(\mathbf{T_1}),E(\mathbf{T_2}),...,E(\mathbf{T_N})$ 表示第 $1,2,...,N$ 分钟内到达的坦克数量存入 $\mathbf{A}=(A_1,A_2,...,A_N)$中，则在 N 分钟内坦克到达总数量 A_{total} 满足：

$$A_{total}=\sum_{n=1}^{N}A_n \tag{8}$$

实验结果：

取 $\lambda=4,N=3$，对样本数进行改变，得到如下实验结果：

```
size = 1000
lambda = 4.0
N = 3

A = [3.947 4.07  3.993]
A_total = 12.01
```

```
size = 10000
lambda = 4.0
N = 3

A = [4.014  3.9791 4.0143]
A_total = 12.0074
```

```
size = 100000
lambda = 4.0
N = 3

A = [4.01122 3.99846 3.99746]
A_total = 12.00714
```

```
size = 1000000
lambda = 4.0
N = 3

A = [3.997912 3.99892  3.999104]
A_total = 11.995936
```

每辆敌军坦克到达的时刻服从期望为 $\frac{1}{\lambda}$ 的指数分布，也就是说坦克到达的时间 S 的分布函数为：

$$F_S(x) = e^{-\lambda x} \tag{9}$$

我们可以生成 M 组数据 $\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_M}$，分别用它们的均值 $E(\mathbf{S_1}), E(\mathbf{S_2}), ..., E(\mathbf{S_M})$ 表示第 $1, 2, ..., M$ 辆坦克到达所需的时间，将其存入 $\mathbf{B} = (B_1, B_2, ..., B_M)$中，则在 N 分钟内每辆敌军坦克到达时间为：

$$\mathbf{B'} = (B'_i = \sum_{1}^{j} B_j | \quad j \in [1, M] \quad where \quad B'_i < N) \tag{10}$$

实验结果：

取 $\lambda = 4, N = 3$，对样本组数 M 进行改变，得到如下实验结果：

```
size = 1000000
lambda = 4.0
M = 1000
N = 3

B' = [0.         0.13342413 0.31777741 1.03763729 1.21159522 1.29430663
 1.40031872 1.84761008 1.87534276 1.88153913 2.10736653 2.11029332
 2.11131691]
size of B' = 13
```

```
size = 1000000
lambda = 4.0
M = 10000
N = 3

B' = [0.         0.17706403 0.41156082 0.4709648  0.84327938 0.90410084
 1.2769691  1.34718539 1.34834828 1.49136868 1.79251071 2.13449103
 2.14307735 2.33244413 2.63948843 2.67663826]
size of B' = 16
```

```
size = 1000000
lambda = 4.0
M = 100000
N = 3

B' = [0.         0.25616824 0.34234413 0.56016894 0.93668404 0.97535618
 1.01152144 1.10112408 1.11153903 1.42089177 1.67121369 1.98186432
 2.36122109 2.49658838 2.86648867 2.97963822]
size of B' = 16
```

```
size = 1000000
lambda = 4.0
M = 1000000
N = 3

B' = [0.         0.179777   0.39791508 0.53242681 0.616768   0.66081224
 0.68747736 0.78975181 1.18480444 1.27434074 1.36375675 1.55436283
 1.59032769 1.72499968 1.8993587  2.15947932 2.3233332  2.88118565]
size of B' = 18
```

基于高斯分布混合模型的模式分类方法

考虑水果聚类问题，水果的属性 $\mathbf{X}$ 满足高斯分布，其均值向量、协方差矩阵分别为 $\mu, \mathbf{\Sigma}$，将其概率密度记为 $p(\mathbf{X}|\mu, \mathbf{\Sigma})$。

定义高斯混合分布：

$$\begin{aligned} p_M(\mathbf{X}) &= \sum_{i=1}^{k} \alpha_i p(\mathbf{X}|\mu_\mathbf{i}, \mathbf{\Sigma_i}) \\ p(\mathbf{X}|\mu_\mathbf{i}, \mathbf{\Sigma_i}) &= \frac{exp\{-\frac{1}{2}(\mathbf{X}-\mu_\mathbf{i})^T \mathbf{\Sigma_i^{-1}} (\mathbf{X}-\mu_\mathbf{i})\}}{(2\pi)^{D/2}|\Sigma_i|^{\frac{1}{2}}} \end{aligned} \tag{11}$$

该分布由 k 个混合分布组成，每个混合成分对应一个高斯分布，其中 μ_i, Σ_i 是第 i 个高斯混合成分的参数，$\alpha_\mathbf{i}$ 为选择第 i 个混合成分的概率，满足：

$$\alpha_i > 0, \sum_{i=1}^{k} \alpha_i = 1 \tag{12}$$

记样本 X_j 属于第 i 个高斯成分的后验概率为 y_{ji}，有：

$$\begin{aligned} y_{ji} &= \frac{\alpha_i p(x_j|\mu_i, \Sigma_i)}{p_M(x_j)} \\ &= \frac{\alpha_i p(x_j|\mu_i, \Sigma_i)}{\sum_{l=1}^{k} \alpha_l p(x_j|\mu_l, \Sigma_l)} \end{aligned} \tag{13}$$

为了得到混合分布的各个组成部分的分布参数，我们需要利用 EM 算法 (Expectation–maximization algorithm) 不断迭代来获取 k 个类的均值和方差参数。

E 步：

根据当前参数计算样本后验概率 $\mathbf{Y} = (y_{ji})_{ji}$

M 步：

根据后验概率更新模型参数 $\{\alpha_i, \mu_i, \Sigma_i | 1 \le i \le k\}$，新的参数与后验概率应该满足下面的关系：

$$
\alpha_i' = \frac{\sum_{j=1}^{N} y_{ji}}{N}
$$

$$
\mu_{\mathbf{i}}' = \frac{\sum_{j=1}^{N} y_{ji} x_j}{\sum_{j=1}^{N} y_{ji}} \tag{14}
$$

$$
\mathbf{\Sigma_i'} = \frac{\Sigma_{j=1}^{N} y_{ji}(x_j - \mu_i')(x_j - \mu_i')^T}{\sum_{j=1}^{N} y_{ji}}
$$

不断重复 E、M 两步直到收敛。

实验结果:

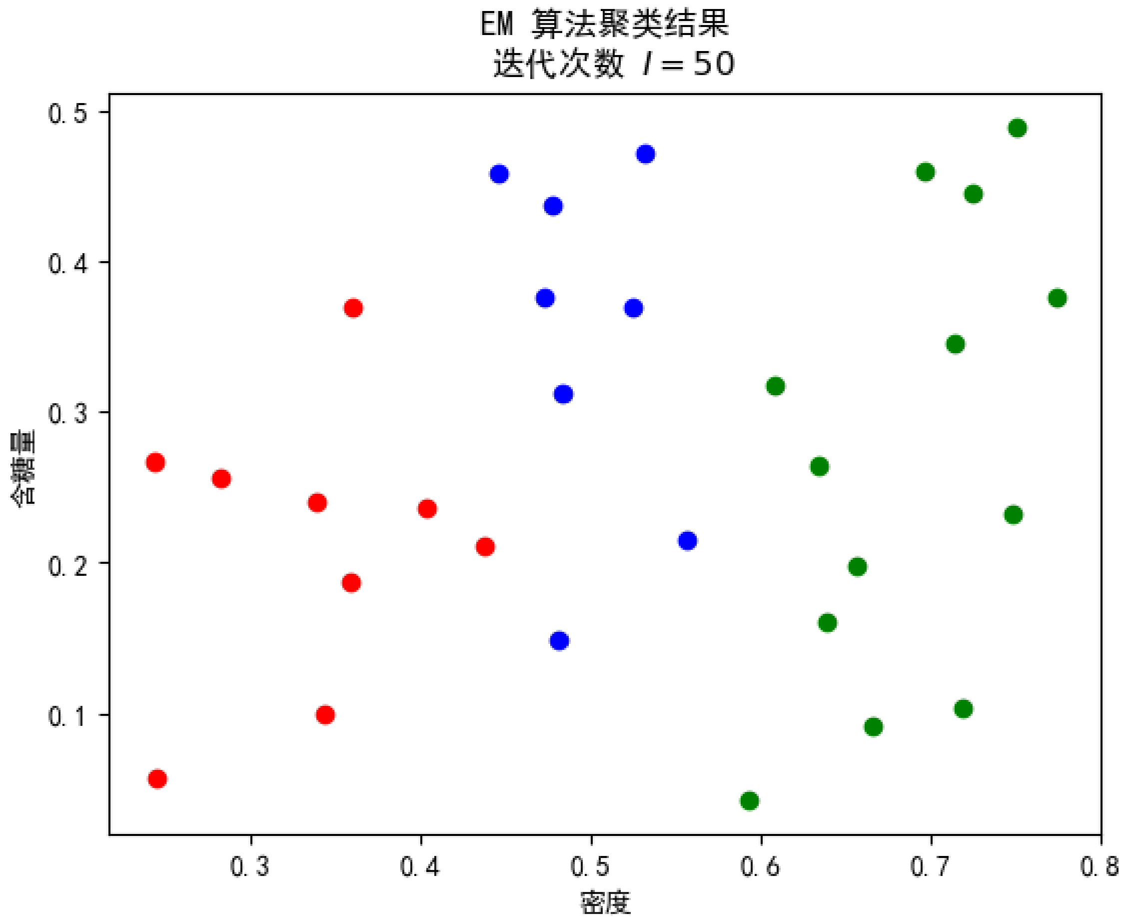
现有水果数据 $\mathbf{S}$:

$$
\mathbf{S} = (\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_N}) \tag{15}
$$

其中 N = 30，S_i 为二维列向量，包含密度、含糖率两个属性，我们随机初始化一组参数：

$$
\begin{aligned}
&\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3} \\
&\mu_{\mathbf{1}} = \mathbf{S_6}, \mu_{\mathbf{2}} = \mathbf{S_{22}}, \mu_{\mathbf{3}} = \mathbf{S_{27}} \\
&\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma_3} = \begin{pmatrix} 0.1 & 0.0 \\ 0.0 & 0.1 \end{pmatrix}
\end{aligned} \tag{16}
$$

令迭代次数 $I = 50$，将得到的结果以散点图绘制如下：



详细计算过程参见 em-50.txt。

参考文献

1. [Monte Carlo method | Wikipedia]
2. [Expectation–maximization algorithm | Wikipedia]