

**计算建模实验报告**

题 目 随机模拟、泊松分布

学 院 计算机科学与技术

专 业 计算机科学与技术

学 号 2021112932

学 生 肖瑶

任 课 教 师 刘绍辉

哈尔滨工业大学计算机科学与技术学院

2023.9

**实验一:** **随机模拟、泊松分布**

1. **实验内容或者文献情况介绍**
   1. **随机数和随机序列的生成**

随机序列仿真是在计算机科学、统计学、数学等领域中非常重要的工具之一。如何产生一个所需分布的随机变量成为关键性问题，我们必须考虑到，实际系统试验消耗人力物力太多，风险代价太大；各种环境的干扰因素，电子设备的热噪声、通信信道的畸变、图像中的灰度失真、飞行器高度表接收的地面杂波，机械系统的振动噪声。

产生随机变量的方法基本有以下几种：

* 加同余法
* 乘同余法
* 加法与乘法相结合
* 混合同余法
* 反馈位移寄存器发生器
* 组合发生器

以上是几种主流的随机数生成法，在求取不同分布的随机数时，若已知一种特定分布，我们还可以做累计函数逆变换，将一个均匀分布的随机数转化为一个连续分布的随机数。

实验1-1不限方法，要求生成随机序列，具体要求如下所示：

* + - * 1. 生成随机序列X，其中每个Xi服从[-1,1]的均匀分布
        2. 生成随机序列Y，每个Yi服从[-1,1]的均匀分布。
        3. 利用随机序列{(Xi,Yi)}计算圆周率（蒙特卡罗投点法）
        4. 请给一幅BMP图像（灰度和彩色都可以）添加均值为m,方差为σ^2的高斯噪声，并计算添加噪声后与原图像的PSNR和SSIM值（PSNR和SSIM请查阅相关资料，用于评价图像的质量），尝试估计噪声图像的噪声均值和方差。
  1. **随机分布的计算机模拟**

1. 利用normrnd生成均值为10，方差为5的正态分布
2. 利用1生成的样本估计分布的均值和方差，并画出均值和方差随样本数增加而变化的图。
3. 敌坦克分队对我方阵地实施突袭，其到达规律服从泊松分布，平均每分钟到达４辆．（1）模拟敌坦克在３分钟内到达目标区的数量，以及在第１、２、３分钟内各到达几辆坦克．（2）模拟在3分钟内每辆敌坦克的到达时刻．（1.用poissrnd(4)进行模拟。2.用exprnd（1/4）模拟）
4. 模拟两状态马尔科夫链，其转移矩阵为,a和b可随意取值，例如取a=1/2,b=1/3,运行1000步，然后统计状态1和状态2出现的次数，与理论上的静态分布进行比较。
   1. **高斯混合模型及EM算法（水果聚类）**

编程完成水果聚类问题，以水果数据的密度以及含糖量为分类依据，为30个不同数据进行分类。

1. **算法简介及其实现细节**

在实验1.1中，实验要求我们实现随机数和随机序列的生成，在这里，采用梅森旋转Mersenne Twister 伪随机数生成器 （PRNG）[1]。该算法的具体实现细节如下所示：

1. 初始化：该类用于使用种子值初始化生成器。种子用于设置生成器的初始状态。在本例中，它使用‘MT19937’，seed=666。
2. 状态初始化：该方法初始化生成器的状态。它设置一个包含 624 个元素的列表，其中第一个元素设置为种子，其余元素使用按位运算和加法基于种子计算。
3. 提取随机数：该方法用于生成随机数。它通过首先调用方法来实现此目的，该方法更新生成器的状态。然后，它将一系列按位运算应用于当前状态以生成随机数。
4. 扭曲操作：该方法更新生成器的状态。它遍历状态数组，并对元素执行一系列按位运算，包括 XOR 和位移。
5. 随机序列的生成：在代码的主要部分，它生成一个随机数序列（在本例中为数字）并将它们存储在列表中。它还使用这些随机数填充 NumPy 数组。
6. 频率分析：然后对生成的随机数执行频率分析。它创建每个数字频率的直方图。
7. 最后，可视化结果'vis.jpg'并存储。

在实验2中，要求生成均值为10，方差为5的服从高斯分布的随机分布，这里使用了numpy库中的函数，得到一系列数据之后，对其进行参数估计，公式即为求均值以及求方差的方程，最后，记录下size变大过程中均值以及方差的变化，结果保存在results文件夹中。

在实验2的第三问中，以坦克到达为背景，求取坦克到达的数量以及每辆到达的时间。因为已知其服从泊松分布，所以设置好平均值，可以用泊松分布估计出1,2,3分钟内到达的坦克数量，显然，这是一个随机数。然后，定义坦克的到达率为坦克到达数量的倒数，生成一组敌坦克在一段时间内的到达时刻，这些时刻符合指数分布。

最后，在实验3中，使用GMM算法对30个水果数据进行聚类。

GMM（Gaussian Mixture Model）[2]是一种基于高斯分布的混合模型，用于对数据进行聚类或密度估计。它假设数据是由多个高斯分布组合而成的，每个高斯分布称为一个“成分”。每个成分代表了数据中的一个聚类。GMM的参数估计通常使用EM算法进行，EM算法是一种迭代优化算法，用于在含有隐含变量的模型中进行参数估计。首先，根据当前的参数估计计算每个样本属于每个成分的概率（即责任）。使用责任重新估计模型的参数，包括均值、协方差和权重。重复进行以上两步，直到收敛。

1. **实验设置及结果分析（包括实验数据集）**

在实验1中，随机数生成中种子使用‘MT19937’，seed=666。

在实验2中，设置生成的高斯分布均值为10，方差为5，size设置为100到10000，步长为100。

实验2中，采用data.txt为数据集，数据集大小为30，特征分量为含糖量与密度。

1. **结论**

在本次实验中，尝试了多种方法，成功生成了服从均匀分布、泊松分布、高斯分布的随机数序列，同时应用生成的随机数序列，较为合理地解决了坦克到达的预测问题。最后，采用GMM算法，解决了水果聚类问题，将给定的30个水果数据分成了3类。

1. **参考文献**

[1] Guillaro F, Cozzolino D, Sud A, et al. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization[J]. 2022.

[2] [2023-10-18]. https://las.inf.ethz.ch/courses/introml-s20/tutorials/tutorial\_em.pdf.