2022 应用信息论基础课程 Project

说明：

1.本次 Project 意在鼓励大家通过理论推导和仿真设计，加深对信源压缩编码理论的理解。

2.大家可以针对问题中自己感兴趣的角度进行探讨，题目中没有明确定义的参数或约束条件也可以自行给出，在报告中予以说明即可。

3.引用他人的学术成果或开源代码应在参考文献中说明

4.请大家**在截至日期前于网络学堂提交报告和代码文件**。其中，报告要求提交 pdf 文档，应包含问题描述、基本原理、求解过程、仿真思路及结果、结论及分析等部分，**无需在报告中粘贴代码**。仿真代码建议采用 Python，Matlab 等实现。需注明运行环境并对代码文件与问题的对应进行说明。

5.本次 Project 不限制大家讨论交流，但报告和代码要求独立完成，若发现抄袭现象，则记 0 分。

问题介绍：

（DNA 检测）外显子是 DNA 序列中的一个区间，且不和其他外显子相重叠。为了简化问题，我们假设待检测 DNA 片段是一串外显子序列，且其中只含有一个目标外显子。而我们需要探测其位置，记为随机变量$X \in \{1,2,3,\dots,n\}$，位置i处为目标的概率记为p_i。取$n=6$，$(p_1,\dots,p_6)=\left(\frac{2}{23},\frac{4}{23},\frac{2}{23},\frac{6}{23},\frac{1}{23},\frac{8}{23}\right)$

问题一、如果每次只能检测一个位置是否含目标外显子，则求（a）最少期望检测次数是多少（b）首先应该检测哪个位置

问题二、如果每次可以任意截取多个外显子，并一起检测其中是否含有目标，则求（a）最少期望检测次数是多少（b）应该采取何种检测策略（c）说明该问题和信源编码的等价性。

问题三、考虑到截取的成本问题，改为每次检测只截取一段连续的区域$\{i,i+1,i+2,\dots,i+k\}$，并检测其中是否含有目标外显子。

（a）证明：在这种情况下哈夫曼编码得到的最小期望检测次数是错误的。

（b）每次检测，等价于提问“is X in set S ?”，其中S为$\{1,2,3,\dots,n\}$的子集，相比问题二，S的选择范围受限于连续区域，使得我们在构造哈夫曼树时不能任意的合并节点。尽管如此，我们还是可以沿用其思想，每次**尽可能**的将概率和最小的节点进行合并。这里，尽可能指存在一个问题能够将合并后节点分开，证明合并的充要条件为：

Proposition 1：假设 A,B 分别为决策树中两个节点对应的 X 的可能取值集合，则两个节点可以合并当且仅当以下任意一个条件满足

（i）A 或 B 集合是连续的

（ii）min A>max B or max A<min B

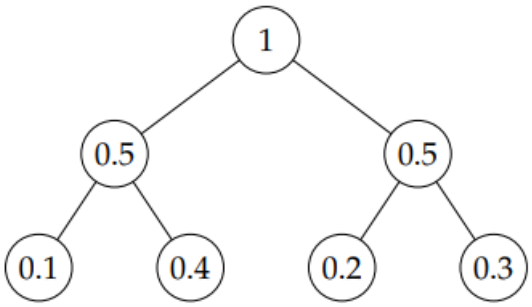
并利用该贪婪算法构建对应的决策树，并计算期望探测次数。

（c）相比贪婪哈夫曼算法从底向上构建决策树，我们也可以尝试从上到下来构建决策树。我们给出贪婪二分算法

Definition 1. 最优划分指，将某个节点对应的可能取值集合 S 划分为两个不相交且互补的集合 A,B，且对于任意的不相交且互补集合 C,D，有$|p(A)-p(B)| \leq |p(C)-p(D)|$

初始时，根节点集合$S=\{1,2,3,\dots,n\}$，然后反复的通过最优划分将决策树进行分裂。可以用以下例子说明其算法。

Example 1：对于分布$(p_1,p_2,p_3,p_4)=(0.1,0.2,0.3,0.4)$，其贪婪二分决策树为



利用该贪婪算法构建对应的决策树，并计算期望探测次数。

（d）固定 n=6，改变概率分布，观察两种贪婪算法得到的期望检测次数与最小期望检测次数（可由暴力搜索得到）的差值和分布。

（e）分别从理论和仿真两个角度，比较两种贪婪算法在鲁棒性，运行速度，期望检测次数等性能上的不同。

问题 4、在问题 2 的设置下，如果在两个位置都有目标基因（相同且都需要检测），试用上述两种方法给出最小期望检测次数和检测方法。取 n=4，$(p_{12},p_{13},p_{14},p_{23},p_{24},p_{34})=(0.1,0.1,0.15,0.15,0.3,0.2)$