2022应用信息论基础项目报告

刘子源 无研223 2022310709

问题描述

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。

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

（b）如果每次只能检测一个位置是否含目标外显子，则应从为目标外显子概率最高的位置开始检测，按照概率递减顺序检测，即首先应该检测第6个位置

（a）如上所述按照6、4、2、1、3、5的顺序检测，注意到当前五个位置都没有检测到目标时，其一定在第六个位置，所以最后一个位置不需要检测

$$l=\frac{8}{23}\times 1+\frac{6}{23}\times 2+\frac{4}{23}\times 3+\frac{2}{23}\times 4+\frac{2+1}{23}\times 5=\frac{55}{23}$$

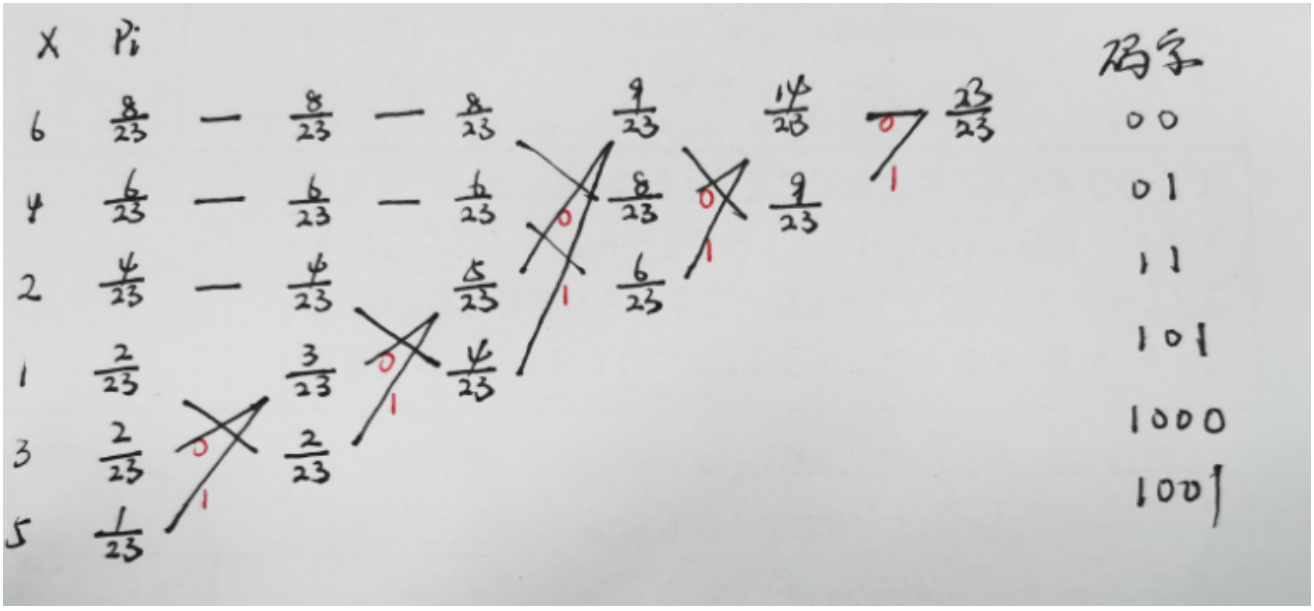
问题二

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。

（c）从开始检测到检测到目标外显子的过程是一个序列，每一个序列能能够唯一确定目标外显子的位置，每一个目标外显子也应该有唯一的检测序列与其对应，说明这其实是一个信源编码的问题

用“0”表示当前检测出了目标外显子，“1”表示当前没有检测出目标外显子，则在检测目标外显子的过程中可以获得一个二元码，该编码的期望长度就是检测策略的期望检测次数；同理，给定一个有效的二元码，我们也可以知道唯一与其对应的检测策略以及目标外显子的位置。综上，使期望检测次数最少等价于使得该信源编码具有最短的期望长度，因此可以使用霍夫曼编码。

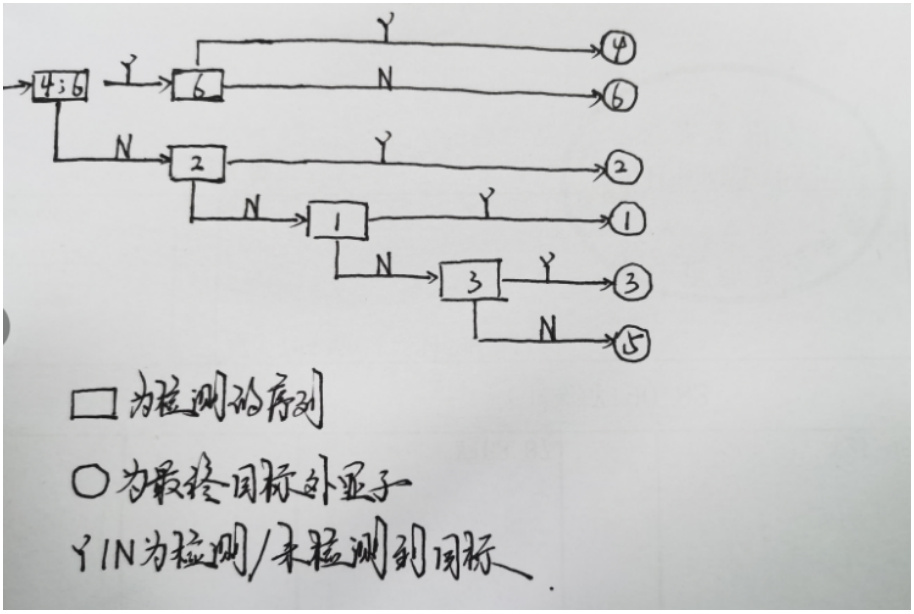
（a）



最少期望检测次数为

$$\sum_{i=1}^{6} p_i l_i = \frac{8}{23} \times 2 + \frac{6}{23} \times 2 + \frac{4}{23} \times 2 + \frac{2}{23} \times 3 + \frac{2}{23} \times 4 + \frac{1}{23} \times 4 = \frac{54}{23}$$

（b）检测策略如图所示



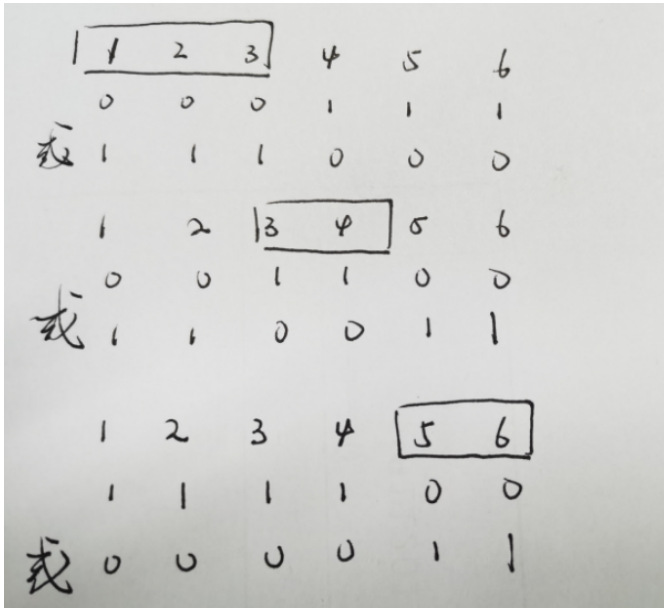
问题三

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$\{i, i+1, i+2, \ldots, i+k\}$，并检测其中是否含有目标外显子

（a）考虑码本的第一位码字，若每次检测只截取一段连续的区域，记每个目标外显子对应编码的第一个码字集和为$\{c(i)\}$，则其只可能为

$$c(i) = 0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0 \ or \ 1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1$$

不存在这样的霍夫曼编码策略，即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，在本题约束下的平均码长会比霍夫曼编码平均长度长



（b）证明

充分性：

若A、B可合并，则存在问题“is X in S”可将合并后的节点分开，其中S是$\{1,\ldots,6\}$的连续子集

若A或B是连续的，由于对称性，不妨设A是连续的，只需令$S=A$，即可将合并后的结果分开；

若A和B是不连续的，可知$\#A\geq 2,\#B\geq 2$，因为有一个元素的子集可看作是连续的

现在需证明$minA>maxB$或$maxA<minB$，出于对称性，不妨设$maxA<maxB$，则只需证$maxA<minB$

用反证法，假设$maxA>minB$，即$minA<minB<maxA<maxB$

记$P=A\cup B$，则S可将P分为子集$Q_1=S\cap P$和$Q_2=S^c\cap P$

若$maxS<maxA$，则$maxA,maxB\in Q_2$，合并后结果无法分开；

若$maxA\leq maxS<maxB$，若$minS\leq minA$，则$minA,minB\in Q_1$；若$minS>minA$，则$minA,maxB\in Q_2$，均无法将P分为集和A、B；

若$maxB\leq maxS$，若$minS\leq maxA$，则$maxA,maxB\in Q_1$；若$minS>maxA$，则$minA,minB\in Q_2$，均无法将P分为集和A、B；

得出矛盾，所以若A、B可合并，且A和B是不连续的，设$maxA<maxB$，必须有$maxA<minB$；同理，若设$maxA>maxB$，必须有$maxB<minA$

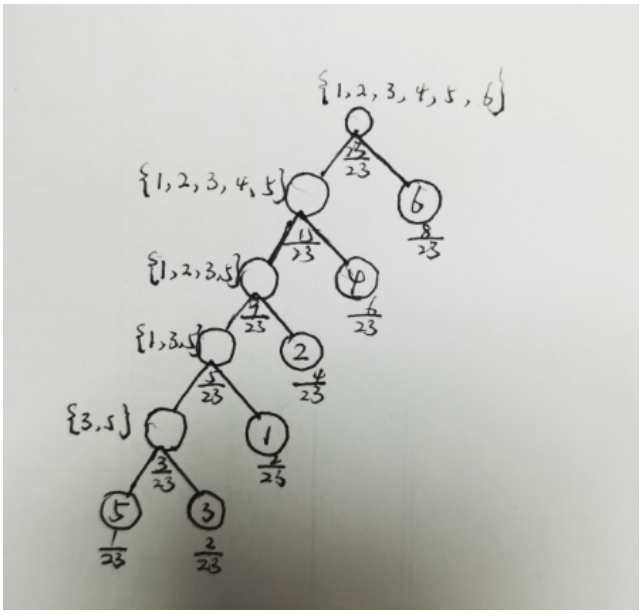
必要性：

若A或B连续，假设A连续，只需令$S=A$，即可将合并后结果分开；

$minA>maxB$或$maxA<minB$，假设$minA>maxB$，只需令$S=\{1,2,\ldots,maxB\}$，即可将合并后结果分开

综上所述，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A或B是连续的或$minA > maxB$或$maxA < minB$

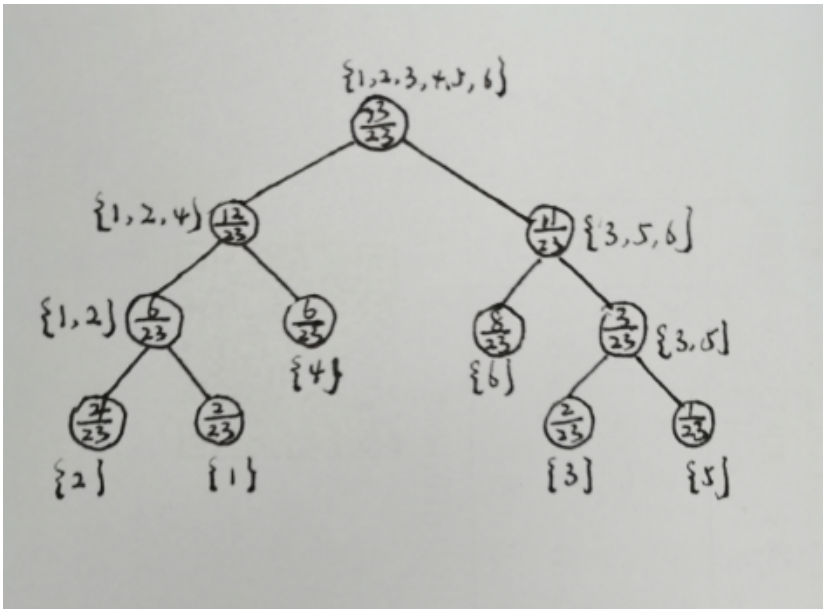
利用该贪婪算法构建的决策树如图所示



最少期望检测次数为

$$\sum_{i=1}^{6} p_i l_i = \frac{1}{23} \times 5 + \frac{2}{23} \times 5 + \frac{2}{23} \times 4 + \frac{4}{23} \times 3 + \frac{6}{23} \times 2 + \frac{8}{23} \times 1 = \frac{55}{23}$$

（c）构建的决策树如图所示

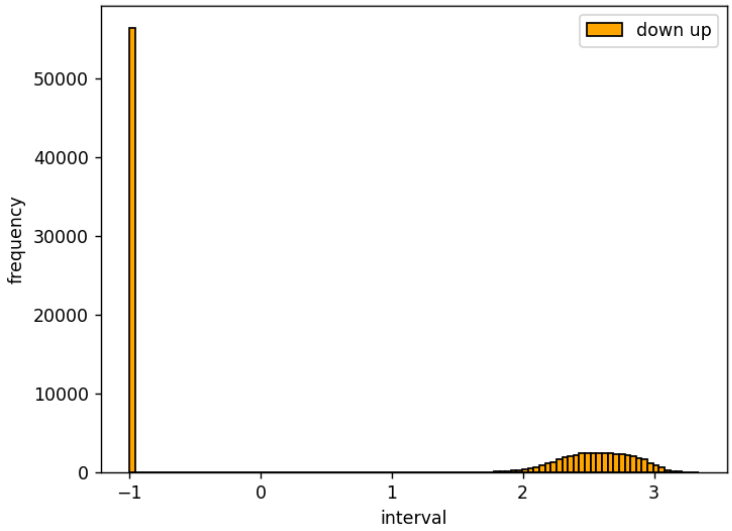


最少期望检测次数为

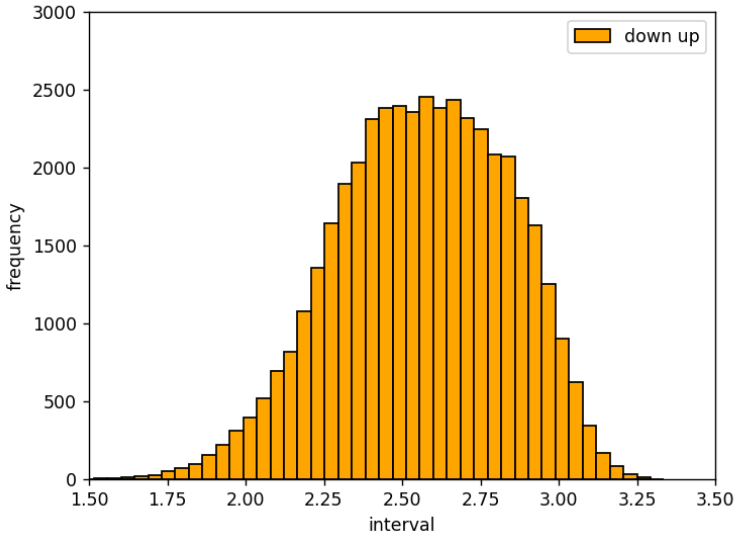
$$\sum_{i=1}^{6} p_i l_i = \frac{1}{23} \times 3 + \frac{2}{23} \times 3 + \frac{2}{23} \times 3 + \frac{4}{23} \times 3 + \frac{6}{23} \times 2 + \frac{8}{23} \times 2 = \frac{55}{23}$$

（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

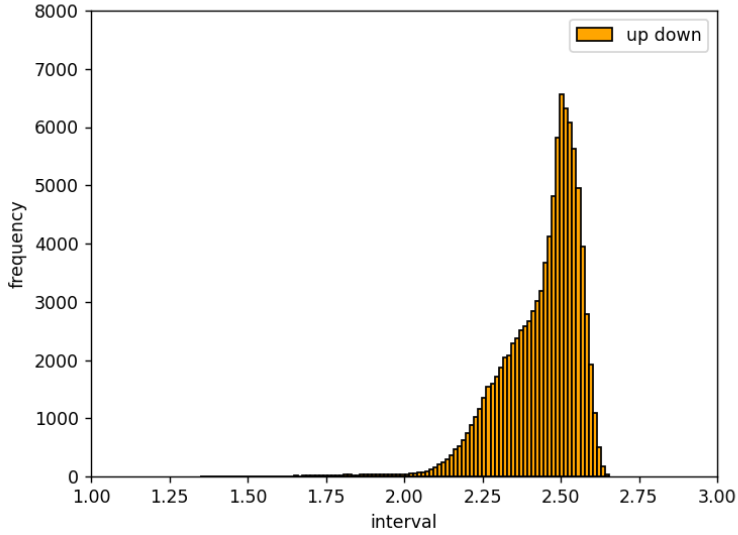
对于自下而上的贪婪决策树算法，并不是每一种分布都可以构造出决策树，每次实验固定$n=6$并随机生成不同的概率分布，结果如下图所示，其中-1为无法构建决策树的情况。统计发现，对于$\mathrm{np.random.random()}$函数随机生成的概率而言，可成功自下而上构建决策树的概率分布仅占43.74%



去除无法成功构建生成树的情况后重新绘图，其期望检测次数分布如下图所示，对于随机生成的概率分布，期望检测次数的平均值为2.5672，方差为0.0758，最小期望检测次数为1.4104，最大期望检测次数为3.3532



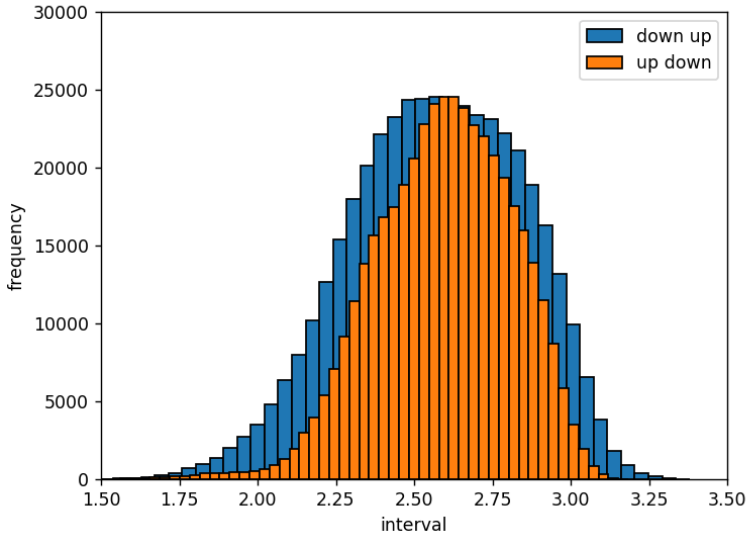
对于自上而下的贪婪决策树算法，任何有效的概率分布均可以构造出决策树，每次实验固定$n=6$并随机生成不同的概率分布，结果如下图所示



对于随机生成的概率分布，期望检测次数的平均值为2.4388，方差为0.0154，最小期望检测次数为1.2544，最大期望检测次数为 2.6538

由上仿真可知，自上而下的贪婪算法的期望检测次数的平均值、方差、最小值、最大值以及算法的稳定性均优于自下而上的贪婪算法

进行10^6次仿真实验，每次实验固定$n=6$并随机生成不同的概率分布，将成功自下而上构建决策树的情况和对应的自上而下的算法构造决策树的期望检测次数的分布在一起绘制，结果如下图所示。仍然可以看出自上而下算法得到的“钟型”更扁，方差更小



（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

鲁棒性

自上而下构建贪婪决策树的鲁棒性更强

从仿真角度分析，从仿真角度分析，给定任意有效的概率分布，自上而下的算法都可以构建决策树并计算期望检测次数，但是只有43.74%的数据可通过自下而上的算法成功构建决策树

从理论角度分析，自上而下的算法对于任意集合$\{1,\ldots,n\}$，根据Definition 1，必然存在一个最优划分，即必然可以构造一个贪婪决策树；但对自下而上算法而言，采用Proposition 1提供的策略进行若干次合并后，会存在无法满足a、b两条从而无法继续进行合并的情况

运行速度

自上而下构建贪婪决策树的算法速度更快

从仿真角度分析，进行10^5次仿真实验，自上而下算法运行时间为$42.3257\mathrm{s}$，自下而上算法运行时间为$43.1871\mathrm{s}$

从理论角度分析，设待划分集合长度为N，对自上而下算法而言，由Definition 1定义可知，单次最优划分的复杂度为$O(N^2)$；

设构造的决策树有h层，则对于第n层节点，一共需要进行2^n次划分，且每次最优划分集合的平均大小为$\frac{N}{2^n}$，则单层的计算复杂度为$\frac{O(N^2)}{2^n}$；

总复杂度为

$$\sum_{n=1}^{h}\frac{O(N^2)}{2^n}=O(N^2)(1-2^{-h})=O(N^2)$$

对自下而上算法而言，由Proposition 1定义可知，单次最优合并的复杂度为$O(N^2)$，一共需要进行N次合并，所以总复杂度为$O(N^3)$

期望检测次数

自上而下构建贪婪决策树的算法的期望检测次数更优

从仿真角度分析，由（d）中分析可知，考虑期望检测次数的平均值、方差、最小值、最大值，均是自上而下算法更小

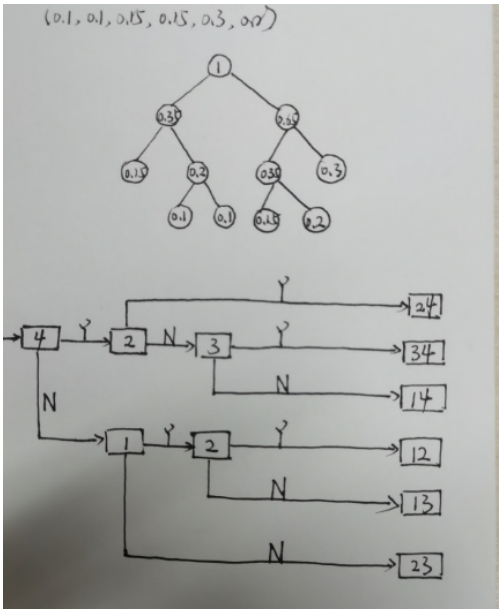
从理论角度分析，自上而下的构建的决策树平衡性更好，即在相同概率分布下，通常它的层数更浅，叶子节点深度更相似，这也解释了其方差更小、期望检测次数的最小值、最大值更小的原因；对于极端概率分布情况自下而上倾向于构建很深的决策树，导致其检测均值更大

问题四

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

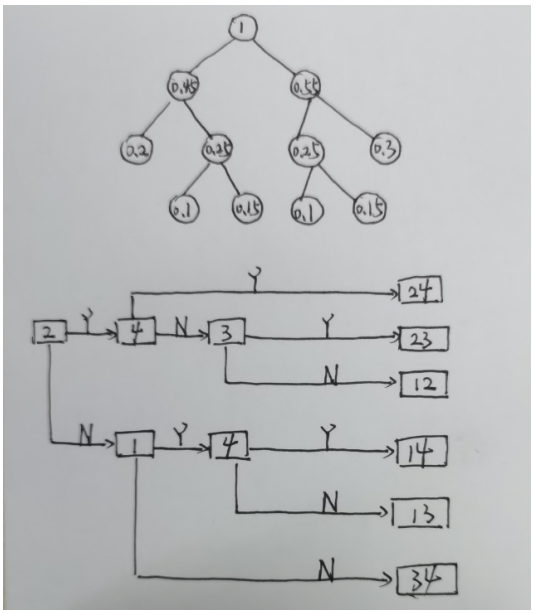
自下而上构造的决策树即对应检测策略如图所示，最小期望检测次数为

$$\sum_{i=1}^{6}p_il_i=0.1\times3+0.1\times3+0.15\times3+0.15\times2+0.3\times2+0.2\times3=2.55$$



自上而下构造的决策树即对应检测策略如图所示，最小期望检测次数为

$$\sum_{i=1}^{6} p_i l_i = 0.1 \times 3 + 0.1 \times 3 + 0.15 \times 3 + 0.15 \times 3 + 0.3 \times 2 + 0.2 \times 2 = 2.5$$



附录

代码运行环境：

python3.9、numpy、matplotlib

使用说明：

分别为两种算法对比、自下而上、自上而下、测速四部分，将想运行的部分注释取消即可