信息学竞赛中的概率问题

胡渊鸣

引言
为什么要讲概率论?

- 近年来竞赛中概率问题大量涌现
- 很多算法需要概率理论支撑
- 良好的概率素养能够提高决策水平

概率空间的定义
概率到底是什么?

- 概率空间的三个要素
  - 样本空间 Ω
    - 事件 $A \subset \Omega$
    - 可以暂时认为Ω的每一个子集都是一个事件
  - 所有事件的集合 F
    - 注意, F是集合的集合
  - 概率测度 $P: F \to R$
    - 一个函数

概率公理

不是每个实值函数都是概率测度

- 一个合理的概率测度, 应该满足如下条件:
  - 非负性
    - $\forall A \in F, P(A) \geq 0$
  - 规范性
    - $P(\Omega) = 1$
  - 可加性
    - $\forall A, B \in F, A \cap B = \phi \Rightarrow P(A \cup B) = P(A) + P(B)$

随机变量
名字带来的误解

- 不是随机性来源
- 不是变量
  - 样本空间上的(实值)函数

- 随机变量$X: \Omega \to R$
  - 如Ω为随机快速排序的每次执行情况之集, 可定义X为某次算法的执行时间.

随机变量的期望
一个重要属性

- 期望是随机变量的属性
  - 表示平均情况下随机变量的输出.
  - 继上例, $E[X]$表示快速排序的平均执行时间.

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) = \sum_{x \in X(\Omega)} x\, P(X = x)$$

期望的性质
解题的基础

- 线性性质(可加性)

$$E[\alpha X_1 + \beta X_2] = \alpha E[X_1] + \beta E[X_2]$$

- 独立的随机变量期望可相乘

$$E[X_1 X_2] = E[X_1]E[X_2]$$

独立性

- 我们称随机变量X, Y是独立的, 当且仅当对于任意x, y, 有
$$\Pr[X = x \cap Y = y] = \Pr[X = x] \Pr[Y = y]$$
- 条件概率
  - $\Pr[X = x \cap Y = y] = \Pr[X = x | Y = y] \Pr[Y = y]$

期望的性质

解题的基础

$$
\begin{aligned}
E[X_1X_2] &= \sum_{x\in(X_1X_2)(\Omega)} xP(X_1X_2=x) \\
&= \sum_{x\in(X_1X_2)(\Omega)} xP(X_1=x_1)P(X_2=x_2) \\
&= \sum_{x\in(X_1X_2)(\Omega)} \sum_{x_1\in X_1(\Omega)} xP(X_1=x_1)P(X_2=\frac{x}{x_1}) \\
&= \sum_{x\in(X_1X_2)(\Omega)} \sum_{x_1\in X_1(\Omega)} x_1\frac{x}{x_1}P(X_1=x_1)P(X_2=\frac{x}{x_1}) \\
&= \sum_{x_1\in X_1(\Omega)} x_1P(X_1=x_1) \sum_{x\in(X_1X_2)(\Omega)} \frac{x}{x_1}P(X_2=\frac{x}{x_1}) \\
&= \sum_{x_1\in X_1(\Omega)} x_1P(X_1=x_1) \sum_{x_2\in X_2(\Omega)} x_2P(X_2=x_2) \\
&= E[X_1]E[X_2]
\end{aligned}
$$

一些期望问题

- 尝试一件事情直到成功为止，每次尝试成功率是 p，期望要尝试多少次才能成功？

应用：Maze
Adapted Form CF 123E

- 从树的根节点S出发, 到叶子节点T点停止, 求DFS算法期望要求多少步.
- 每次DFS将从这个点出发未到达过的点random_shuffle以后按这个随机顺序往下试探.
- 注意, DFS时返回(弹栈)的过程也算一步.

- 题目本身不难, 但是需要按部就班说清楚才能让解法使人信服.

应用：Maze
Adapted Form CF 123E

- 样本空间Ω是什么?
  - 所有从S到T的可能的DFS路径.
- 概率分布P?
  - $P(\{\omega\})$表示走出路径ω的概率.
  - 对于任意的事件A,$P(A) = \sum_{\omega \in A} P(\{\omega\})$
- 随机变量X?
  - $X(\omega)$表示路径ω的走过的路程.
- 要求什么?
  - $E[X]$, 即走过路程的期望.

应用：Maze
Adapted Form CF 123E

- 路径数目巨大, 无法一一枚举.
- 利用**期望的线性性质**按边分解. 设随机变量$X_e(\omega)$表示路径ω经过边e的次数.

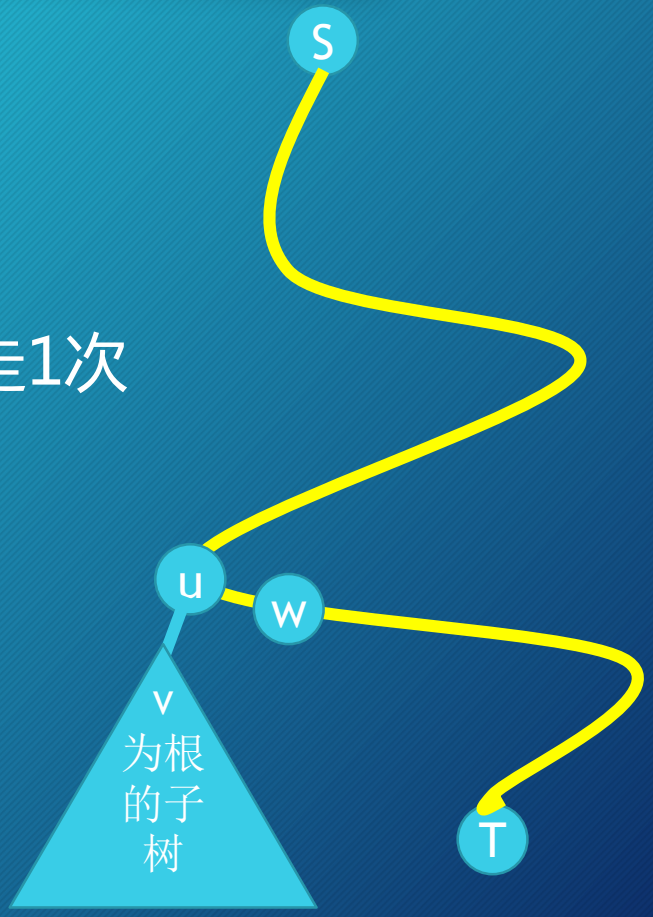
$$X = \sum_{e \in edges} X_e$$

$$E[X] = \sum_{e \in edges} E[X_e]$$

- 只需求出$E[X_e]$即可

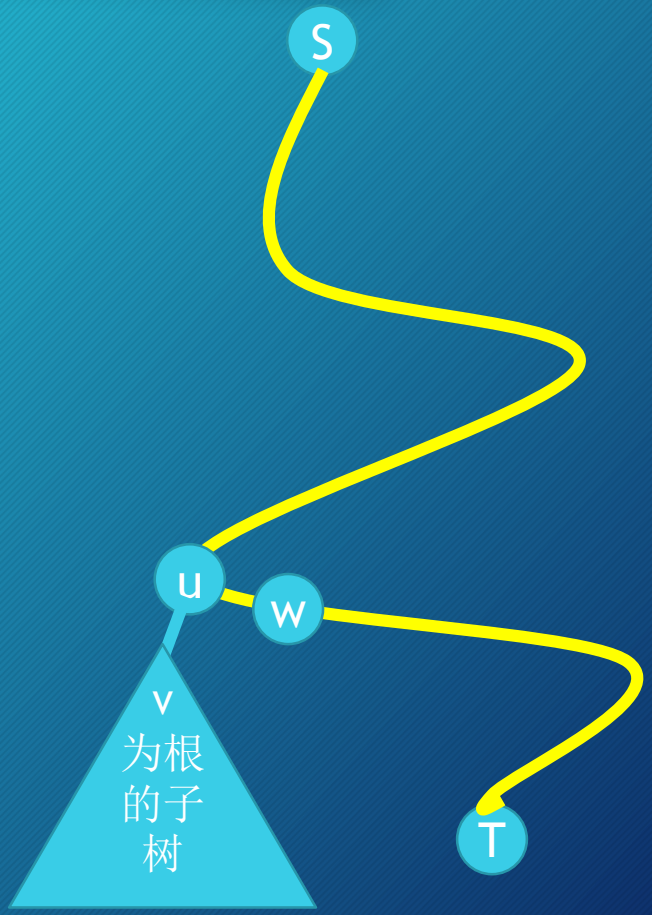
应用：Maze
Adapted Form CF 123E

- 现在问题变成, 如何求$E[X_e]$
- 所有必经路径(黄色)上的边, 必然仅走1次



应用：Maze
Adapted Form CF 123E

- 对于不在必经路径上的边e, 假设:
  - e距离必经路径上最近的点为u
  - 其所在子树以u的儿子v为根
  - T所在子树以u的儿子w为根
- 考虑DFS到达u时
  - 先进入w, 则e走了0次
  - 先进入v, 则e走了2次
  - 两种情况概率分别是$\frac{1}{2}$
  - 期望走的次数为1次.



应用：Maze
Adapted Form CF 123E

- 神奇的结论:
  - 每条边期望走的次数都是一次, 于是期望步数是$n-1$.
  - 这种DFS期望时间仅和树的点数有关, 与形态无关.

BZOJ 列队春游

- 给定n个人的身高(互不相同), 现在将他们随机排成一排, 问期望有多少对人能够互相看到?
- 两个人能够互相看到当且仅当这两个人之间的人的身高比这两个人都矮.

BZOJ 列队春游

- 给定n个人的身高(互不相同), 现在将他们随机排成一排, 问期望有多少对人能够互相看到?
- 期望的线性性质的典型应用.

一些期望问题

- n个在[0,1]上均匀分布的i.i.d.(independent identically distributed, 独立同分布)的随机变量的最大值是多少?
  - 猜一下?
  - 答案是$\frac{n}{n+1}$
  - 需要用到一些微积分的知识
- 最小值是多少?

概率论感觉测试

- 下面的几个小题大部分是从一年前网上很流行的<概率论感觉测试>中选出的, 非常具有代表意义.
- 通过对这些问题的思考, 能够提高大家对概率论的掌握程度
- 我会对每一个问题给出大家能够接受的解答

概率论感觉测试

- 假设考试周为1个礼拜（周一到周日），且考试时间为均匀分布，假使你有3门考试，则最后一门考试大约在
- A 周五
- B 周六
- C 周日

概率论感觉测试

- 有以下几个国家，每个国家有自己的习俗。问哪个国家长期以后男人的比例最大

- A. 每个家庭不断的生孩子直到得到第一个男孩为止
- B. 每个家庭不断的生孩子直到得到第一个女孩为止
- C. 每个家庭不断的生孩子直到得到一男一女为止
- D. 以上几个国家最后男女比例基本一样

概率论感觉测试

- 给一个1-n的排列，与原来位置相同的数字的个数的期望大约是（如 n=5 则51324 与原来位置只有3是相同的）
-
- A. 1
- B. log n
- C. ln n

概率论感觉测试

- 美国的25分硬币共有50种，上面有50个州的图案，如果我们每次得到的硬币是随机的，则大约收集多少可以收集全

- A. 200
- B. 300
- C. 400
- D. 500

概率论感觉测试

- 假设在一根长为1米的绳子上随机的分布5只蚂蚁，他们的位置和初始的方向都是均匀随机的。从时刻0开始，他们朝着他们初始的方向以每分钟1米的速度开始爬，直到离开绳子或者碰到另外一只蚂蚁。当他们碰到另外一只蚂蚁时，两只蚂蚁会分别转向然后继续前进。问期望大约多少时间之后所有蚂蚁都将离开绳子?
- A．50秒
- B． 1分钟
- C． 2分钟
- D．5分钟

概率论感觉测试

- 假设有1000次100m短跑大赛，每次比赛的冠军成绩都在9.7-10之间均匀分布，问期望有多少次比赛比赛能够破纪录
-
- A. 7
- B. 10
- C. 15
- D. 32

概率论感觉测试

- 几乎所有的自然数都含有9

- A. 对
- B. 错

概率论感觉测试

- 几乎所有的自然数都含有9

- A. 对
- B. 错

概率论感觉测试(生日悖论)

- 将23个数填入大小为365的Hash表, 可以认为每个数会被独立均匀地放入表中的一个随机的位置, 发生Hash冲突的概率是
- A. 小于1%
- B. 1% 到10 %
- C. 10% 到 50%
- D. 50% 以上

概率论感觉测试(随机游走问题)

- 如果一个物体在3维随机游动，也即每一刻他可以向左，右，上，下，前，后等概率的走，长久来看，则会发生什么情况

- A. 此物体无穷多次回到原点
- B. 此物体无穷多次回到任何一条坐标轴上，但不会无穷多次回到原点
- C. 此物体不会无穷多次回到任何一条坐标轴上

概率论感觉测试(中心极限定理)

- 打10000副拱猪，总共持有9500-10500个A的概率大约在
-
- A. 80%-90%
- B. 90%-95%
- C. 95%-99%
- D. 99%以上

概率论感觉测试(采样偏见)

- 实验室测试灯泡的寿命。在灯泡坏的时候立刻换新灯泡。灯泡寿命约为1小时。考察10000小时时亮着的那个灯泡

- A. 那个灯泡的寿命期望也约为1小时
- B. 那个灯泡的寿命期望约为其他灯泡的2倍
- C. 那个灯泡的期望寿命约为其他灯泡的1/2
- D. 以上说法都不对

常见认知偏见

- 为什么父母从来不挑食, 而子女经常被指责挑食?
- 比尔盖茨, 乔布斯等人都没有好好上大学, 最后也很成功→大学无用?
- 我之前这么倒霉, 这次一定能够有好运的?
- “网上秀恩爱, 分手分得快”

常见认知偏见

- 为什么父母从来不挑食, 而子女经常被指责挑食?
  - 因为菜是父母买的, 他们只买自己喜欢吃的菜, 自然不会挑食.
- 比尔盖茨, 乔布斯等人都没有好好上大学, 最后也很成功→大学无用?
  - 大部分不好好上大学的人都不会像他们那样成功, 他们只是少数.
- 我之前这么倒霉, 这次一定能够有好运的?
  - 即使你之前再倒霉, 这也不会让你这次变得走运.
- “网上秀恩爱, 分手分得快”
  - 不秀恩爱的人, 即使分手也不会被发现.

概率转移网络上的相关问题

概率转移网络是OI中常见的一大类问题的模型, 现在来看看这个模型中的各种问题如何解决.

首先要明白什么是概率转移网络. 概率转移网络(以下简称网络)是一个**有向网络**, 由**点集(状态集)**V, **转移概率矩阵(一个二元函数)**$G : V \times V \to [0, 1]$, 以及**起点**$v_0$组成. 其中, 对于每个$u \in V$, 有$\sum_v G[u, v] \leq 1$.

有了数学上的定义, 再来看看这个模型的实际意义. 一个移动的质点, 一开始(时刻0)位于v_0. 对于每一个时间段, 如果质点位于顶点u, 那么对于任意$v \in V$这个点有$G[u, v]$的概率转移到v, 还有$1 - \sum_v G[u, v]$的概率会消失(或者理解为移动到了一个虚空的点).

CF 113 D Museum

Petya和Vasya在进行一次旅行, 他们决定去参观一座博物馆. 这座博物馆由m条走廊连接的n间房间, 并且满足可以从任何一间房间到任何一间别的房间, 两个人决定分头行动, 去看各自感兴趣的艺术品. 他们约定在下午六点到一间房间会合. 然而他们忘记了一件重要的事:他们并没有选好在哪儿碰面. 等时间到六点, 他们开始在博物馆里到处乱跑来找到对方. 不过, 尽管他们到处乱跑, 但他们还没有看完足够的艺术品, 因此他们每个人采取如下的行动方法:每一分钟做决定往哪里走, 有p_i 的概率在这分钟内不去其他地方(即呆在房间不动), 有$1 - p_i$的概率他会在相邻的房间中等可能的选择一间并沿着走廊过去. 这里的i指的是当前所在房间的序号. 每条走廊会连接两个不同的房间, 并且任意两个房间至多被一条走廊连接.

CF 113 D Museum

两个男孩同时行动. 由于走廊很暗, 两人不可能在走廊碰面, 不过他们可以从走廊的两个方向通行. 此外, 两个男孩可以同时地穿过同一条走廊却不会相遇. 两个男孩按照上述方法行动直到他们碰面为止. 更进一步地说, 当两个人在某个时刻选择前往同一间房间, 那么他们就会在那个房间相遇.

两个男孩现在分别处在a, b两个房间, 求两人在每间房间相遇的概率. $(n \leq 22)$

CF 113 D Museum

- 建模

我们尝试将这个问题转化为上面提到的概率转移网络. 不难发现, 由于两个人物的存在, 我们需要将状态集V定义为$V_0 \times V_0$, 其中V_0为题目中涉及的房间的集合. 通过题目中给出的条件, 我们不难求出每个状态(两个任务的位置)转移到另一个状态的概率, 即矩阵G. 同时, 初始状态$v_0 = (a, b)$. 由于问题的特殊性, 还要再定义停止状态集合$S = \{(a, a)|a \in V_0\}$. 接下来有两种方法来解决这个问题.

CF 113 D Museum

- 迭代法

如果将S中的状态的转移特殊处理, 将其转移概率除了转移到自己为1外其余全部为0, 不难发现, 我们要求的其实是经过足够多步骤的移动以后, 质点所在的位置(经过足够长时间, 质点一定停留在S中). 记网络中时刻t时, 质点处于每个点的概率为x^t, 其中x_u^t 为质点时刻t 在状态u 的概率, 不难发现, 有$x^{t+1} = Gx^t$. 初始状态x^0中, 只有$x^0_{(a,b)}$ 为1, 其余全部为0.

这时, 需要求的解是x^∞. 这个值没法在有限的时间之内算出来的, 不过可以发现, 当t足够大时, x^t和x^∞ 其实是非常接近的. 我们可以利用快速幂尽可能高地算出G^{2^K}, 然后利用$x^t = G^t x^0$得出一个相当好的近似解.

对于大部分比较弱的数据, 这种方法还是可以接受的, 但是一旦数据比较强, x^t收敛的速度不理想的时候, 该方法就无能为力了. 因此, 必须找一个更好的解法.

CF 113 D Museum

- 解法二:解线性方程组

区别于解法一, 我们采用另一种方法来处理S中的状态, 将其转移概率全部设为0, 即到达了S中的状态以后下一步必然转移到“虚空”. 这样做的目的是便于列方程.

下面考虑, 落入虚空之前, 质点停留在每个点的次数的期望E_u, 即每个时刻质点位于这个点的概率之和$\sum_t x_u^t$(这里使用了**期望的线性性质**).

对于异于v_0的点u, 我们有:

$$E_u = x_u^0 + \sum_{t=1}^{\infty} x_u^t = x_u^0 + \sum_{t=1}^{\infty} \sum_v x_v^{t-1} G[v,u] = x_u^0 + \sum_v G[v,u] \sum_{t=0}^{\infty} x_v^t = x_u^0 + \sum_v G[v,u] E_v$$

显然, $x_{v_0}^0 = 1, x_u^0 = 0, u \neq v_0$.

这样我们就建立了一个方程组, 解之即可. 由于S中的状态只能停留一次, 所以质点停留在这些点的期望次数就等于质点最后一步停留在这个点的概率.

CF 113 D Museum

- 通过高斯消元, 可以得到一个$O(n^6)$的解法.

SDOI 2012 走迷宫

Morenan被困在了一个迷宫里. 迷宫可以视为N个点M条边的有向图, 其中Morenan处于起点S, 迷宫的终点设为T. 可惜的是, Morenan 只会从一个点出发随机沿着一条从该点出发的有向边, 到达另一个点. 这样, Morenan走的步数可能很长, 也可能是无限, 更可能到不了终点, 若到不了终点, 则步数视为无穷大. 但你必须想方设法求出Morenan所走步数的期望值.

- $N \leq 2{,}000$.
- 最大强连通块大小不超过200

CF 229 E Gifts

- 有m个罐子, 每个罐子里面有一些钞票. 一个罐子里面不会有相同面值的钞票. 主持人告诉你每个罐子里面的钞票面额, 现在你需要去抽出n张钞票. 只有当你抽取出全部的n张钞票以后, 你才会知道抽出的钞票的面额.
- 很明显, 你希望能够抽出面值之和最大的n张钞票. 有一些抽取的方案是有可能达成这个目标的, 而有一些方案不能. 你将会在所有可能的方案中随机选择一种, 按照选出的方案去抽取.
- 求你抽出的钞票面值之和最大的概率.
- $n, m \leq 1{,}000$

CF 229 E Gifts

- 求你抽出的钞票面值之和最大的概率.
- 答案为每种方案的成功概率除以方案数
- 如何计算每种方案的成功概率?

随机化方法

- 无向图的割
  - 定义
    - 割C是指集合对(S,T), s.t. $S \cap T = \Phi$, $S \cup T = V$
    - 割C的大小$|C| = \sum_{(u,v)\in E,[u\in S]\neq[v\in S]} 1$
- 对于一个随机的割，其大小的期望是多少？
  - 随机是指, 每个点等概率地被分配到两个集合中

随机化方法

- 对于一个随机的割，其大小的期望是多少？
  - 随机是指, 每个点等概率地被分配到两个集合中
- $|C| = \sum_{(u,v)\in E,[u\in S]\neq[v\in S]} 1$
- $|C| = \sum_{(u,v)\in E} [[u \in S] \neq [v \in S]]$
- $E[|C|] = E[\sum_{(u,v)\in E} [[u \in S] \neq [v \in S]]] = \sum_{(u,v)\in E} E[[u \in S] \neq [v \in S]] = \sum_{(u,v)\in E} \frac{1}{2} = |E|/2$

随机化方法

- $E[|C|] = |E|/2$
- 推论:
  - 对于一个无向图, 一定存在一个割, 其大小不少于边数的一半
  - 为什么?
- 如何求出这个割?
  - 多次随机?
    - 随机次数没有保证
  - 去随机化(Derandomization)!

随机化方法

- 去随机化(Derandomization)
- 目标: 确定性地求出无向图的一个大小不少于边数的一半的割
- 方法:
  - 按顺序枚举每一个点, 考虑这个点应该在S中还是T中.
  - 注意到, 我们可以计算, 在确定了一些点的归属以后, 剩下的点随机分配时, 割的期望大小.
  - 所以, 我们只需要去分别计算当前枚举的点分别属于两个集合时, 剩下的点随机分配形成的割的期望大小, 取期望不少于边数一半的那个方案即可.
  - 会不会两个方案的期望割大小都小于边数一半?

矩阵乘法

- 给定3个矩阵A, B, C, 规模均为3000×3000, 判断是否有$AB = C$
- 要求$O(n^2)$的复杂度.
- 允许一定的错误率

占领数

- 给定n个数, 有Q个询问.
- 每次询问为:
  - 第l个数到第r个数之间的所有数中, 是否存在一个数出现次数大于$\frac{r-l}{2}$?

- 随机算法?

两个不等式
对极端情况的估计

- Markov不等式
  - 对于非负随机变量X, 任意正实数a, 有

$$P(X \geq a) \leq \frac{E[X]}{a}$$

- Chebyshev不等式
  - 对于任意随机变量有

$$P(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}$$

$$P(|X - \mathrm{E}[\mathrm{X}]| \geq c\sigma) \leq \frac{1}{c^2}$$

应用:最小圆覆盖算法的分析
众所周知的线性算法, 会不会超时呢?

- 利用Markov不等式来看看
- 对于$n = 200{,}000$

$$P(X \geq n^2) \leq \frac{E[x]}{n^2} = \frac{1}{n} = 5 \times 10^{-6}$$

- 仅是最乐观的估计, 已经相当小
  - 小于每个人死于交通事故的概率