浅析信息学竞赛中概率论的基础与应用

江苏省扬州中学　胡渊鸣

摘 要

随着信息学竞赛内容的不断充实, 涉及概率的问题不断涌入. 具有一定水平的选手都具备了相当的解决概率问题的能力, 但是一旦询问起解法的稍严格的证明, 很多选手就束手无策了.

本文针对作者自身学习初等概率论过程中的体会, 以及和其他选手交流时发现的概率论难点进行重点分析, 内容包括概率论的基本概念、随机变量与期望、概率转移网络上的相关问题、Markov不等式以及Chebyshev不等式等, 从理论和实践两个紧密联系的层面帮助读者理清概率论的脉络, 深化对涉及概率论的解法的理解.

1　概率

1.1　什么是概率

如果读者仔细想过这个问题, 可能会发现这个问题不太好回答. 直观的说, 概率大的事情发生的可能性就大, 因此概率就是对事情发生的可能性的度量.

概率论的公理化体系会涉及很多测度论的内容, 有兴趣的读者不妨去查阅相关书籍. 由于本文只是从信息学竞赛的角度去讨论概率论, 往往并不涉及一些比较奇异的集合(比如可数个集合的并之类), 也不涉及某些连续的随机问题, 所以在信息学竞赛中用到的概率理论可以大大简化. 不过, 在此还是建议大家在闲暇时去研究一下概率论的公理化体系, 相信会对你对概率论的理解有很大帮助.

1.2　概率空间

竞赛中用到的初等概率论有三个重要成分, 即**样本空间**Ω, **事件集合**F和**概率测度**P. 我们常说的**事件**, 其实是Ω**的某个子集**(请读者时刻牢记这一点). 在竞

赛中, 往往可以认为Ω的每个子集都是一个事件(但在超出竞赛的某些领域, 这样做是有问题的), 所有事件的集合记为F(注意这里是集合的集合). 概率测度P, 是**事件集合到实数的一个函数**. 当然, 并不是所有的概率测度都是合理的. 一个合理的概率测度, 需要满足以下3条**概率公理**:

(1) 对于任意的事件A, 有$P(A) \geq 0$(非负性).

(2) $P(\Omega) = 1$(规范性).

(3) 对于事件A和B, 如果$A \cap B = \Phi$, 有$P(A \cup B) = P(A) + P(B)$(可加性).

我们称符合要求的三元组(Ω, F, P)为**概率空间**. 举例来说, 随机掷一个均匀的骰子, 考虑其向上的面, 我们有**样本空间**$\Omega = \{1, 2, 3, 4, 5, 6\}$, "奇数向上"的**事件**是$\{1, 3, 5\}$, **事件集合**$F$为$\Omega$的幂集(所有子集的集合), **概率测度**$P(A) = \frac{|A|}{6}$, 经过验证, 这是一个合理的概率空间.

相信这个概念大家都比较熟悉, 并且也觉得是理所当然的. 在此列出只是考虑到部分读者可能看到后面的内容时会对概率的基础感到困难.

1.3 条件概率

条件概率在OI中其实是非常常用的概念, 但是一不小心就会导致错误. 要解释这个概念并不难, 我们来看一个例子: α大学的学生中有99%是男生, β大学则有99%是女生. 假设两个学校的人数一样多, 我们在两所学校中随机的选出一个学生, 他(她)的性别是男性的概率有多大?

这个问题显然是古典概型的问题, 不难得出男生在所有学生中占50%, 所以选出男生的概率当然是50%.

如果我们与选出的学生进行交谈以后得知他(她)是α大学的, 那么这个概率显然变成99%了. 由此可见, **当我们得到了更多的信息以后, 事件的概率是会改变的**.

记已知事件B发生的情况下, 事件A发生的(条件)概率是$P(A|B)$. 设选出的人是α大学的学生这个事件为U_1, β大学为U_2, 选出的学生是男生的事件为S_1, 女生为S_2. 在上面的这个例子中, 我们可以写作$P(S_1|U_1) = 99\%$.

注意, 很多初学者在理解这个问题的时候, 总是有一个"先后性"的枷锁, 认为事件B一定先于事件A发生. 这在某些情况下是不妥的, 事件只是样本空间的

某个子集, 并没有时间这个属性, 本例中选择学生是一次性完成的. 虽然在某些情况下两个事件之间确实有可能有明显的时间先后关系, 如复习以后考试通过的概率和不复习考试通过的概率明显是不一样的.

计算条件概率的公式如下:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

在研究概率问题时, 我们常常把$A \cap B$写成AB, 或者"A, B". 所以, $P(A \cap B)$, $P(AB)$, $P(A, B)$含义是一样的.

其实在考虑条件概率的时候, 我们是**把事件B看做了新的样本空间**, 由于事件和样本空间都是集合(别忘了同时样本空间本身也是一个事件), 这样做是没有问题的. 新的概率测度往往被称为条件概率, 条件概率公式实质上揭示了**两个样本空间上的概率测度的关系**.

1.4 全概率公式

全概率公式在概率问题中有着举足轻重的作用, 但是很多选手没有深入理解它的含义. 如果$B_1, B_2, B_3, ..., B_n$是概率空间的一个划分, 那么有:

$$P(A) = \sum_k P(A|B_k)P(B_k)$$

这个公式是用分类讨论方法研究概率问题时的基础, 几乎所有的概率问题都涉及这个公式. 刚才我们在得出男生在所有学生中占50% 这个结论的时候, 实际上已经用到了全概率公式了: $P(S_1) = P(S_1|U_1)P(U_1) + P(S_1|U_2)P(U_2) = 99\% \times 50\% + 1\% \times 50\% = 50\%$

1.5 Bayes公式

考虑等式$P(A|B)P(B) = P(AB) = P(B|A)P(A)$, 忽略中间一项, 将两边同时除以$P(B)$, 得到"Bayes公式":

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

该公式在计算**逆向概率**的问题上非常有用.

由于通常已知的往往是一系列的条件概率$P(B|A_k)$, 其中$A_1, A_2, ..., A_n$是样本空间的一个划分, 这种情况下还会将上式和全概率公式联合起来, 得到:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_j P(B|A_j)P(A_j)}$$

再来看看上面讲到的那个例子, 现在问题变成:已知选到的同学是男生, 你能否求出他是α大学学生的概率?

现在要求的实际上是$P(U_1|S_1)$利用Bayes公式, 可以求得:

$$P(U_1|S_1) = \frac{P(S_1|U_1)P(U_1)}{P(S_1|U_1)P(U_1) + P(S_1|U_2)P(U_2)} = \frac{99\% \times 50\%}{99\% \times 50\% + 1\% \times 50\%} = 99\%$$

2　随机变量与期望

期望是概率题中又一常出现的概念, 它是随机变量的一个重要属性, 而这一点常常被选手所忽略, 导致期望成了缺乏随机变量这个基础的空中楼阁. 要理解期望到底是什么, 必须先了解随机变量.

2.1　随机变量的定义

随机变量(random variable)是一个有趣的词, 因为它其实并不是随机性的来源(随机性来自样本空间), 也不是变量, 而是定义在样本空间Ω 上的**确定的**实值函数. 不过很多情况下, 我们往往会抛弃样本空间而直接去考虑随机变量的属性(比如说期望), 以至于很多选手只知道期望的概念, 而搞不清随机变量到底是什么. 在这里, 有必要给随机变量下一个清晰的定义:

函数$X : \Omega \to \mathbb{R}$被称为一个随机变量.

在大多数情况下, 有了随机变量以后就可以抛弃对原本样本空间的关注, 而是集中注意于对于每个实值(注意, 本文只考虑离散的情况), 随机变量能够取得该值的概率. 这个过程实际上是将样本空间重新划分的一个过程, 将在这个函数下取得同一实数值的样本空间中的元素合并了. 当然, 读者之前在理解随机变量的时候, 很可能就是直接从其取每个值的概率入手的.

2.2　随机变量的期望

期望(expectation)是对随机变量表现出的平均情况的的一种刻画, 也是在竞赛中最频繁出现的概念. 很多选手是在脱离随机变量的情况下学习这个概念的,

这样做是不可取的. 对于一个随机变量, 定义其期望如下:

$$E[X] = \sum_{\omega} P(\omega)X(\omega) = \sum_{x} xP(X = x)$$

这里, $X = x$表示的是一个事件, 等价于集合$\{\omega|\omega \in \Omega, X(\omega) = x\}$. 前一个式子是从输入(样本空间)的角度入手的定义, 后一个式子则是从不同的输出入手, 将样本空间进行了划分(将输出相同的输入看成一个整体), 即将不同的输出值按概率加权后求和.

由于对于样本空间的某些元素, 随机变量的输出值很可能是相同的, 有时我们就可以不从样本空间的角度去考虑随机变量, 而是直接考虑"随机变量取某个特定的值"这个事件. 不过仍然需要注意的是, 样本空间和概率测度是随机变量的基石, 如果发现问题从期望这个高层无法解决, 我们就需要从底层入手了. 后面的几个涉及期望的性质的证明就是从概率空间入手的, 没有概率空间我们也没有办法证明两个随机变量的独立性.

2.3 随机变量的独立性与乘积的期望

随机变量的独立性是两个随机事件在其输出层面上的属性. 对于两个随机事件X_1, X_2和实数$x_1 \in X_1(\Omega), x_2 \in X_2(\Omega)$ 如果有$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$ 就称X_1, X_2 相互独立.

两个独立的随机变量的重要性质是:其积的期望等于期望的积, 证明如下:

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

2.4 期望的线性性质

线性性(可加性)是期望的性质中重要的一项. 不管两个随机变量X_1和X_2是否独立, 总有:

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

这个性质在竞赛中的应用常常表现为将一个“大”的随机变量分成“ 小”的随机变量的和, 那么“大”的随机变量的期望就是每个“小”的随机变量的期望的和.

2.5 应用:Maze(Adapted from Codeforces 123E)

2.5.1 问题描述

给定一棵树, 从根节点S出发, 到叶子节点T点停止, 求DFS算法的期望步数. 每次DFS将从当前点出发未到达过的点“RandomShuffle” 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.

2.5.2 解答

作为涉及期望的第一题, 我们有必要把一些细节搞清楚. 先明明确**样本空间**Ω是从S出发, T终止的所有路径的集合. **随机变量**X作用于Ω上, 表示每条路径

的长度. **概率测度**P为Ω的某个子集到$\mathbb{R}$ 的函数$P(A)=\sum_{\omega\in A}P(\omega)$. 这里$P(\omega)$是走出路径$\omega$的概率, 可以计算出来, 也可以避开计算这一概率.

我们要求的是$E[X]$. 不难发现枚举每条路径在数据规模大到一定程度的时候是不可行的, 必须另辟蹊径. 构造一些"小"随机变量X_e, 其中e是树的某条边, $X_e(\omega)$表示路径ω 中, 经过边e的次数(来回都算). 考虑到对于$\forall\omega\in\Omega, X[\omega]=\sum_e X_e(\omega)$, 简记作$X=\sum_e X_e$ 利用"期望的线性性质", 有$E[X]=\sum_e E[X_e]$. 这时, 问题就转化成了求$E[X_e]$

我们称S到T的路径为主路径, 将边分成2类, 一类是在必经路径上的, 一类是不在必经路径上的. 先考虑必经路径上的边. 这类边走过的次数一定是1. 一方面由于边在必经路径上, 如果不走这条边就无法到达T; 另一方面一旦走过这条边, 就一定会在反向走这条边之前到达终点, 所以这条边必须且只能走一次. 在考虑不在必经路径上的边e, 这类边有两种情况, 一种是没有走到, 另一种是来回走了2次. 为了求出期望走的次数, 我们必须求出走0次和走2次的概率. 考虑必经路径上和这条边最近的点u, 假设这条边处于以点u的儿子v为根的子树中, 而终点处于以点u的儿子w为根的子树中. u不等于w, 否则u不是必经路径中离e最近的点. 这时我们发现, e到底走了0次还是2次完全取决于u 出发以后是先走了v 还是先走了w. 如果先走了v, 那么以v为根的子树中每条边都走了2次; 如果先走了w, 路径根本不可能进入子树v. 而在一个随机的排列中, 一个元素排在另一个元素前面的概率是$\frac{1}{2}$, 所以这条边走了0次和2 次的概率都是$\frac{1}{2}$(注意此时我们其实是在随机变量的输出层面上考虑问题). 于是, e走过的期望次数就是1.

综上, 每条边走过的期望次数都是1. 于是一个很神奇的结论是, 不管树的形状是如何的, 我们走过的期望边数都是$n-1$.

2.6 全期望公式

首先来考虑一个类似于条件概率的问题:如果给定了更多的信息, 如事件A一定发生, 样本空间Ω上的随机变量X会出现何种变化?

如果将这个受约束的随机变量记作$X|A$, 那么对于$\forall x\in X(\Omega)$, 我们有:

$$P((X|A)=x)=\frac{P(X=x,A)}{P(A)}$$

这就是Ω上的随机变量X在新的样本空间A上, 形成的新的随机变量$X|A$的全部信息.

对于随机变量X, Y, 一个让人有点摸不着头脑的公式是

$$E[E[X|Y]] = E[X]$$

根据上面的分析, $X|Y = y$是在样本空间$Y = y$上面的一个随机变量, 于是$E[X|Y = y]$是一个确定的数值. 如果不明确指定Y的取值, 则$E[X|Y]$是一个新的随机变量, 其期望表示在Y 的各种特定输出y的前提下, $E[X|Y = y]$按照$P(Y = y)$加权的和.

这个公式的意义类似于全概率公式, 使得我们可以对期望问题进行分类讨论. 全概率公式也是这个公式的基础, 下面的推导中第(5)到第(6) 步就用到了全概率公式.

为了便于理解这个公式, 证明如下:

$$\begin{aligned}
E[X|Y] &= \sum_{y\in Y(\Omega)} E[X|Y=y]P(Y=y) &(3)\\
&= \sum_{y\in Y(\Omega)} \sum_{x\in X(\Omega)} xP(X=x|Y=y)P(Y=y) &(4)\\
&= \sum_{y\in Y(\Omega)} \sum_{x\in X(\Omega)} xP(X=x,Y=y) &(5)\\
&= \sum_{x\in X(\Omega)} x \sum_{y\in Y(\Omega)} P(X=x,Y=y) &(6)\\
&= \sum_{x\in X(\Omega)} x \sum_{y\in Y(\Omega)} P(X=x|Y=y)P(Y=y) &(7)\\
&= \sum_{x\in X(\Omega)} xP(X=x) &(8)\\
&= E[X] &(9)
\end{aligned}$$

举个现实中的例子:在全年级的学生(样本空间Ω)中等概率(概率测度$P(A) = \frac{|A|}{|\Omega|}$)任选一人, 询问其上次考试的成绩$X$(随机变量), 求$E[X]$. 第一种方法是直接询问每个同学的成绩然后算出平均数作为$E[X]$. 第二种方法是请求每个班主任算出本班的同学的平均成绩, 然后再按照选到的同学属于每个班的概率(即该班人数除以年级总人数)加权求和.

第二种方法其实是构造了一个新的随机变量Y, $Y(\omega)$表示的是ω所在的班级编号. 这种方法算出来的就是$E[E[X|Y]]$. 显然两种方法求出来的期望是相等的, 这也说明了全概率公式的正确性.

读者可能会注意到其实这里的Y只涉及判断相等的操作, 其值域不一定是实数. 在后面的例子中会再次提到这一点.

3 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$\sum_v G[u, v] \le 1$.

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).

熟悉**Markov链**的同学可能会觉得这个模型和Markov链非常相似, 不过考虑到OI中可能遇到的问题类型, 作者将Markov链进行了一些调整. 这个模型具有很强的普适性.

下面来看两个例题.

3.1 例题:Museum(Codeforces 113D)

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.

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. ($n \leq 22$)

3.1.1 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.

3.1.2 解法一: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$x_{(a,b)}^0$ 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.

3.1.3 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.

这个方法比上个方法强大许多, 可以通过全部数据, 时间复杂度$O(n^6)$, 不会超时. 相信不少选手知道这个算法并且能够正确实现甚至AC, 但是经过调查, 发现其实大部分选手并没有深入理解这个方程中的未知数的含义.

3.2 例题:走迷宫(SDOI 2012)

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.

3.2.1 解法

本问题和上述模型十分接近, 建模不存在难点. 要注意的是, 终点T的出边应该全部删去. 这道题的解法具有一定的技巧. 我们用$E[X_u]$表示从点u出发到达终点T期望需要走的步数. 特别的, $E[X_T] = 0$.

于是, 直觉上, 我们有$E[X_u] = \sum_v G[u,v]E[X_v] + 1$.

这个公式可谓是相当简单, 枚举下一步朝那个方向走, 然后按概率加权再加上走的这一步的1个步数.

可是, 这毕竟只是直观上的一个公式, 虽然它是正确的, 但是我们必须从理论上去证明它, 否则这是不能使人信服的. 证明的时候需要用到全期望公式$E[X_u] = E[E[X_u|Y_u]]$, Y_u是一个随机变量, 表示**从u出发以后第一步到达的节**

点(正如介绍全期望公式时提到的, 这里需要稍微扩展一下随机变量的概念, 这个随机变量是Ω到V的映射). 不难发现$E[X_u|Y_u = v] = E[X_v]$, 带入$E[X_u]$ 就能得到结果. 读者不妨仔细想一下这个问题里面的样本空间和概率测度具体是什么.

有了n个等式, 我们就可以通过解方程解决本题.

4 两个常用不等式

4.1 Markov不等式

对于非负随机变量X和正实数a, 总有以下不等式成立:

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

这个不等式说明了随机变量表现出一个较大值的概率和它的期望的关系, 在我们对一个随机变量只知道其期望的时候非常实用. 注意, 大多数情况下等号是取不到的, 而且左边会远小于右边.

4.1.1 应用:对最小圆覆盖随机增量算法运行时间的估计

知道这个算法的同学应该都明白这个算法的期望时间复杂度是$\Theta(n)$的. 但是, 总是有些同学会担心, 虽然期望复杂度是线性的, 某些情况下算法还是会退化成平方的, 导致超时. 这里, 我们来利用Markov不等式分析一下这种情况发生的概率. 用随机变量X表示算法的执行时间, 有

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

也就是说, 算法的时间复杂度退化成$\Theta(n^2)$的概率是$\Theta(\frac{1}{n})$级别的. 对于$n = 200,000$, $\frac{1}{n} = 5 \times 10^{-6}$, 仅仅是最粗略的估计, 算法退化的概率已经小于每个人死于交通事故的概率了! 不严格的说, 如果你还是不相信这个算法时间复杂度的稳定性, 你这辈子就别出门了. 如果再考虑别的因素, 退化的概率比上面算出的数值往往还会小几个数量级.

这个例子说明, 如果我们给出了算法的期望复杂度, 往往这个算法发生退化的概率是微乎其微的, 所以大家遇到某些运行时间含有随机因素的算法(随机快排, Treap等)大可放心使用.

4.2 Chebyshev不等式

我们称随机变量X的方差$Var[X]$为$E[(X-E[X])^2]$, 标准差σ为$\sqrt{Var[X]}$.

如果我们知道了随机变量的方差, 就可以进一步对随机变量以大于某一幅度偏移其期望的概率进行估计. 对于随机变量X和正实数a, 有:

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

上式也可以利用变量代换$a=c\sigma$写成:

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

上式表明在仅知道标准差的情况下, 随机变量偏移其期望c倍标准差的概率小于等于$\frac{1}{c^2}$.

其实, 这个不等式从直观角度进一步诠释了方差, 即"方差越小的随机变量偏离其期望的幅度越小".

5 总结

本文涉及的初等概率论的内容其实并不复杂, 但是很多初学者学习的时候往往理不清脉络, 忽视其中的逻辑关系, 认为很多简单的定义无需注意, 直接凭自己靠不住的直觉去理解. 本文从"概率" 这个概率论的基石出发, 通过离散型随机变量的讨论, 重点研究了随机变量基本的数字特征:期望, 并应用于解题过程中. 最后, 通过与均值(用一阶统计量——期望表示)和分散程度(用二阶统计量——方差表示)相关的关于随机变量估计的两个不等式说明两个统计特征的用途.

本文涉及的题目不多, 一些内容是选手们知道结论但是可能理解并不深入的细节问题. 很多概率论中的问题具有相当的哲学高度, 如偶然与必然、局部与全局、有限与无限相互之间的关系. 读者在做题之余, 不妨仔细去想想概率论的一些最为基础的问题, 知其然更知其所以然, 这样一定会利用概率论有效提高思辨能力.

由于作者才疏学浅, 如果文章中有错误还请读者们不吝赐教. 文中很多定理和不等式的证明读者可以在任何一本介绍概率论的书上找到, 在此限于篇幅, 不在赘述.

致谢

感谢CCF提供了展示自我的平台.
感谢倪震祥老师长期以来对我的指导.
感谢张煜承, 许昊然, 贾志鹏同学对我的论文提出宝贵意见.
感谢父母, 学校对我的培养.

参考文献

[1] 苏淳, 《概率论》, 科学出版社.

[2] 斯蒂芬·弗莱彻·休森,《数学桥——对高等数学的一次观赏之旅》, 上海科技教育出版社.

[3] http://tsinsen.com/A1475, Codeforces 113D Museum.