COS 341: Discrete Mathematics

Homework #6 — Fall 2006
Probability I — Due: Wednesday, November 22

Special late policy: Because this homework is due right before Thanksgiving, there will be a special late policy in which the Thursday and Friday of Thanksgiving break count together as a single late "day." To be clear, this table shows how many "days" late your assignment will be counted if turned in on the following dates:

Calendar date	**Number of late days charged**
Wednesday, November 22	0
Friday, November 24	1
Sunday, November 26	2
Monday, November 27	3
Tuesday, November 28	*Not accepted on or after this date*

For this homework only, if you are out of the Princeton area over break, you may mail or email your assignment to Mohammad. If mailed, your homework is considered submitted on the post mark date, and should be sent to this address: Mohammad Mahmoody Ghidary, Princeton University, Department of Computer Science, 35 Olden Street, Princeton, NJ 08540. (It would be wise to send him email at the same time you mail your assignment so that he can look out for it; also, save a photocopy of your work.) Note that mailing your assignment may delay when it is graded and returned to you.

See instructions on the "assignments" web-page on how and when to turn in homework, and be sure to read the collaboration and late policy for this course. Approximate point values are given in brackets. *Be sure to show your work and justify your answers.*

1. You shuffle a deck of cards and deal your friend, who is honest, a 5-card hand.

a. [5] Suppose your friend says, "I have the ace of spades." What is the probablity that she has another ace?

b. [5] Suppose your friend says, "I have an ace." What is the probability that she has another ace?

c. [3] Are your answers to (a) and (b) the same? Explain why.

2. Two new teams, the Frequentists and the Bayesians, have made it to the 2007 World Series. Assume that the Frequentists win each game with probability 2/3 independently of the outcomes of other games. Nevertheless, to keep the Bayesians happy, assume that the teams keep playing, possibly forever, until the Bayesians have indeed won four games. Thus, the series only ends when the Bayesians have won exactly four games. Let R be the random variable denoting the number of the fourth game won by the Bayesians (and therefore the final game of the series); for example, if they win the first, second, fourth and sixth games, then $R = 6$. Let $R = \infty$ if they never manage to win four games.

a. [4] For each finite n, give a simple closed form expression for $\Pr[R = n]$.

b. [4] What is the probability that the Bayesians never win four games, that is, $\Pr[R = \infty]$? (Note that this is *not* necessarily the same as $\lim_{n\to\infty} \Pr[R = n]$.)

c. [4] What is the probability that the Bayesians win an ordinary 7-game series?

3. A gambler has two dice in his pocket. The first die is an ordinary fair die, giving equal probability to each of the six sides. The other die is numbered like an ordinary die; however, when thrown, the probability that this die comes up "6" is 1/3, with the remaining probability distributed equally among the other five faces.

Now consider the following experiment. The gambler picks one of the two dice at random (choosing each with equal probability). Then he throws that selected die twice in a row. Let R_1 be the result of the first throw, and let R_2 be the result of the second throw. Also, let D be the random variable indicating which die was selected; this random variable takes values f for the fair die, and b for the biased die.

a. [4] Suppose for just this part that the first roll R_1 comes up 6. What is the probability that the biased die was chosen? What is the probability that the second roll R_2 will be 2?

b. [4] Show that R_1 and R_2 are *not* independent.

It might seem strange in the last problem that R_1 and R_2 are not independent since if we were talking about a single die, whether biased or not, we would ordinarily assume that the throws of that one die are independent. Indeed, although R_1 and R_2 are not independent, they "become" independent once we know which of the two dice is selected; that is, once a die has been selected, multiple throws of that single die are independent. The random variables R_1 and R_2 are only dependent as long as we remain ignorant about which die will be selected.

Mathematically, we can capture this idea with the concept of *conditional independence*. We say that two events A and B are *conditionally independent* given a third event C if

$$\Pr[A \cap B|C] = \Pr[A|C] \Pr[B|C].$$

That is, once everything is conditioned on C, the events A and B become independent. Similarly, two random variables X and Y are *conditionally independent* given a third random variable Z if

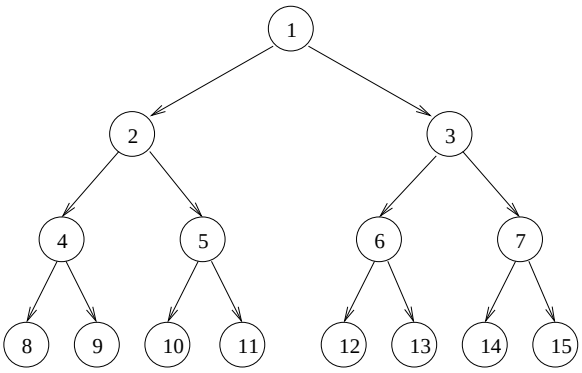
$$\Pr[X = x \cap Y = y|Z = z] = \Pr[X = x|Z = z] \Pr[Y = y|Z = z]$$

for all possible values of x, y and z. In other words, the events $X = x$ and $Y = y$ must be conditionally independent given $Z = z$ for all choices of x, y and z.

Returning to the example, in this terminology, R_1 and R_2 are conditionally independent given D.

4. Every human male has exactly one Y-chromosome which he inherits directly from his father. This implies that, in principle, every man has the identical Y-chromosome as his father, his father's father, his father's father's father, etc. However, in reality, small changes occur with low probability in every generation.

To model this situation, let us consider the male descendents of a single man. To simplify matters, we assume that each man has exactly two sons. Thus, we can arrange all of the men in a tree like this one:



In the tree, each of the men have been numbered. Man 1 has two sons, Men 2 and 3; Man 2 has two sons, Men 4 and 5, etc.

Now let us suppose that Man 1's *father* has a unique marker (DNA sequence) on his Y-chromosome. This marker will ordinarily be passed from father to son, but let us suppose that there is a small probability p that the marker vanishes, due to mutation, when passed from father to son. Thus, if a father has the marker, then each of his sons will inherit it with probability $1 - p$ (independent of whether anyone else inherits it). Once the marker vanishes, it never comes back, so if a man does not have the marker, none of his descendents will either.

Let R_i be the event that Man i inherits the marker. Let d_i be the depth of Man i in the family tree; for instance, $d_1 = 0$, $d_2 = d_3 = 1$, $d_4 = d_5 = d_6 = d_7 = 2$, etc.

a. [3] For each i, what is the probability that Man i inherits the marker?

b. [3] Suppose that Man i is a direct ancestor of Man j. What is the probability that Man j inherits the marker, given that Man i has inherited it? What is the probability that Man i has inherited the marker, given that Man j has inherited it?

c. [3] What is the probability that Man 11 has inherited the marker, given that Man 8 did *not* inherit it?

d. [3] Are R_{13} and R_{14} independent? Why or why not?

e. [4] For which of the following choices of i is it the case that events R_{13} and R_{14} are conditionally independent given R_i?

1 2 3 4 6

5. There are n contestants, numbered $1, 2, \ldots, n$, participating in the preliminary round of a spelling bee, which is intended to eliminate all weak spellers. A sequence of m words $W_1, W_2, \ldots, W_m$ are selected independently and uniformly at random from a very large master catalog of difficult words. (The words are not necessarily distinct, although they almost certainly will be since the catalog is so large.) All contestants are presented with this same set of words, and any contestant who misspells even a single word is immediately eliminated from the spelling bee. For simplicity, we assume that every contestant either does or does not know how to spell any particular word; thus, we ignore the possibility that the same contestant may sometimes spell the same word correctly, and sometimes incorrectly if presented with the same word more than once.

Let p_i be the fraction of words in the master catalog that contestant i spells incorrectly. Since each word W_j is chosen uniformly at random from C, this is equivalent to saying that the probability that contestant i misspells word W_j is exactly p_i.

Let E_i be the event that contestant i survives this preliminary round by correctly spelling all m words.

a. [3] Let i and j be two different contestants. Are the events E_i and E_j necessarily independent?

b. [4] What is the probability that contestant i survives the preliminary round? Give an exact answer, and then show that your answer is at most $e^{-p_i m}$.

c. [4] Let ϵ be a small positive number (like 1% or 5%). Let us call contestants ϵ-*weak* if the fraction of words in the master catalog that they do not know how to spell is greater than ϵ. Show that the probability that the set of survivors of this preliminary round includes even one of the ϵ-weak spellers is at most $ne^{-\epsilon m}$.

d. [4] Suppose there are 500 contestants, and the organizers of the spelling bee want to be 95% confident (i.e., they want the probability to be at least 95%) that they have eliminated all spellers who know how to spell fewer than 90% of the words in the master catalog. Using the bound from the last part, how many spelling words should they include in this round?

(As a side note, when the contestants are replaced by prediction rules, and the spelling words are replaced by training examples, this analysis becomes the foundation for understanding a wide range of machine learning methods.)

6. [10] Let X and Y be independent random variables, and let f and g be any functions with domains containing the ranges of X and Y (respectively). Show that the random variables $f(X)$ and $g(Y)$ are also independent. Do not assume any special properties about f and g (for instance, that they are surjections, injections, etc.).