CS 70 Discrete Mathematics and Probability Theory
Spring 2011 Demmel HW 12

Due April 22, 8:00am

You may work in groups of up to 3 people (no larger!). Please read the group collaboration policies on bSpace or `www.cs.berkeley.edu/~demmel/cs70_Spr11` before beginning group work. You *must* write up the solution set entirely on your own. You must never look at any other students' solutions (not even a draft), nor share your own solutions (not even a draft).

Please begin your answer to each question on a new sheet of paper, and make sure that each sheet is labeled with your name, section number, GSI name, the assignment number, the question number, and "CS70–Spring 2011".

Turn in each question in a different box in 283 Soda Hall: Question 1 in the box labeled "CS70 - 1", Question 2 in the box labeled "CS70 - 2", etc. Reason: Different problems will be graded in parallel by different readers.

Warning: You risk receiving no credit, or losing points, for any homework that does not conform to the above regulations! Please take the time to write clear and concise solutions; we will not grade messy or unreadable submissions. No late homeworks will be accepted. We will drop the lowest two homework scores.

1. (20 pts.) Which bandit is which?

You walk into a room with 3 slot machines, numbered 1, 2 and 3, where machine i let's you win with probability p_i. You know that $p_1 = .8$, $p_2 = .5$ and $p_3 = .25$, but you don't know which machine is which. So you pick a random machine and play it 7 times. The results are WWLLWLW (W = Win and L = Loss).

(a) What is the probability that the machine you chose is machine 1? Machine 2? Machine 3? Which machine is most likely? Show your work.

(b) Suppose that, after playing 7 times, you want to play a different machine. If you pick a different machine at random and play it, what is the probability of winning?

2. (20 pts.) Which bandit is which (continued)?

We start with the same setup as in Question 1: You walk into a room with 3 slot machines, with probabilities of winning $p_1 = .8$, $p_2 = .5$ and $p_3 = .25$. You pick a random machine, play it 7 times, with results WWLLWLW. Now suppose your friend arrives, picks a different machine at random to play, and wins 3 times in a row. What is the probability that your machine is machine 1? Machine 2? Machine 3?

3. (20 pts.) Which bandit is which (continued)?

Suppose you walk into a room of n slot machines, each with distinct winning probability p_i. You pick one at random, play it m times, and win w times and lose $m-w$ times. What is the probability it is machine i? Which machine is most likely? In other words, explain how you would pick the mostly likely i depending on the values of $p_1, ..., p_n$, w and m. Hint: Explain what happens when w/m is much closer to one value of p_i than any other.

4. (20 pts.) Functions of Random Variables

Let $S = \{1,2,3,4,5,6\}$ be a sample space, with probability function $P(i) = i^2/91$. Let random variables $f(i) = i$ and $g(i) = 6 - i$.

(a) What is $E(f)$? $E(g)$? $V(f)$? $V(g)$?

(b) What is $E(f+g)$? $V(f+g)$? $E(f-g)$ $V(f-g)$?

(c) Let $h(x) = (-1)^x$. What is $E(h(f))$? $E(h(g))$?

(d) Let $h(x) = x^2 - 5x + 6$. What is $E(h(f))$? $E(h(g))$?

(e) Are f and g independent? Are f and $f+g$ independent?

5. (21 pts.) Mendelian inheritance

Let's look at an application of the theory we've been developing, to classical inheritance of genes, as initially described by Gregor Mendel based upon his experiments with pea plants.

Here's a review of Mendel's model. Mendel determined that the height of each pea plant is determined by the genes it has. In his model, each pea plant has two genes that together determine the plant's height; each gene can be one of two possibilities, either h or H. The H gene is dominant, and if H is present in either of the plant's two genes, the plant will be tall. The h gene is recessive, and if both genes are h, the plant will be short.

We can see that there are four possibilities for the combinations of the plant's two genes: HH, Hh, hH, hh. The combinations Hh and hH are indistinguishable, so the standard convention is to write both of these cases as simply Hh. Thus, there are three possibilities for the plant's genotype (the portion of its genetic code related to height): HH, Hh, and hh.

The height of each plant is determined by its genotype. Plants with a genotype of HH or Hh will be tall; plants with the genotype hh will be short.

A new pea plant can be formed by crossing two existing pea plants: its "father" and its "mother". The "child" plant inherits one gene from its father (a gene chosen uniformly at random from its father's two genes) and one gene from its mother (randomly chosen from the mother's two genes). For each parent, it's random which gene the child inherits from that parent, and both possibilities are equally likely. For example, if the father has genotype HH and the mother has genotype Hh, the child might have genotype HH or Hh, each with probability $1/2$. As another example, if the father has genotype Hh and the mother has genotype Hh, the child's genotype could be HH, Hh, or hh; these occur with probabilities $1/4$, $1/2$, and $1/4$, respectively. (Here Hh occurs with probability $2/4$, since it can be obtained either by inheriting a H from the father and h from the mother, or by inheriting a h from the father and H from the mother.)

Mendel deduced that, in a large population of pea plants, the genotypes will be HH, Hh, and hh, in proportions $1/4$, $1/2$, and $1/4$, respectively. Thus, $3/4$ of pea plants (the ones with genotype HH or Hh) will be tall, and $1/4$ will be short. Of the tall pea plants, $1/3$ will have genotype HH and $2/3$ will have genotype Hh.

Here's the thing. Given a pea plant, you can directly measure whether it is tall or short. However, there is no easy way to determine its genotype directly. Of course, based upon the pea plant's height, you can draw some inferences about its potential genotypes, but there is no way to observe genotypes directly (without sophisticated technology that was not available to Mendel). In this problem, we're going to develop a procedure for probabilistically inferring the genotype of a pea plant, by crossing it with other plants and measuring the heights of its children.

So suppose we have a particular pea plant, let's call it Penelope; we want to infer Penelope's genotype. Let the random variable X denote Penelope's genotype. X is hidden: we cannot observe X directly. Based upon

the overall frequency of genotypes in the population at large, before we observe anything, our best estimate for X can be summarized by the prior distribution: $\Pr[X = HH] = 1/4$, $\Pr[X = Hh] = 2/4$, $\Pr[X = hh] = 1/4$. Now we're going to repeatedly pick another plant at random from a large population of pea plants, cross that other plant with Penelope, and look at the height of the child. Let the random variable Y_i be an indicator r.v. for the event that the ith child obtained in this way is tall. In other words, $Y_i = 1$ if Penelope's ith child is tall, and $Y_i = 0$ if Penelope's ith child is short. Assume that the genotypes of all the other plants crossed with Penelope are independent.

(a) Calculate the conditional probabilities

$$\begin{aligned} &\Pr[Y_1 = 1|X = HH] && \Pr[Y_1 = 0|X = HH] \\ &\Pr[Y_1 = 1|X = Hh] && \Pr[Y_1 = 0|X = Hh] \\ &\Pr[Y_1 = 1|X = hh] && \Pr[Y_1 = 0|X = hh] \end{aligned}$$

(b) What is the probability that Penelope's first child is tall? In other words, calculate $\Pr[Y_1 = 1]$.

(c) Suppose we measure and find that Penelope's first child is tall. What is the posterior distribution for X, given this observation? In other words, calculate the conditional probabilities

$$\Pr[X = HH|Y_1 = 1], \qquad \Pr[X = Hh|Y_1 = 1], \quad \text{and} \quad \Pr[X = hh|Y_1 = 1].$$

(d) In part (c), you determined a method for updating the prior distribution to the posterior distribution after observing the event that Penelope's first child is tall, under the assumption that the prior distribution is $(1/4, 2/4, 1/4)$. Now let's generalize this to an arbitrary prior distribution. Suppose the prior distribution is $\Pr[X = HH] = p$, $\Pr[X = Hh] = q$, $\Pr[X = hh] = 1 - p - q$. With this prior, calculate the posterior distribution

$$(\Pr[X = HH|Y_1 = 1], \quad \Pr[X = Hh|Y_1 = 1], \quad \Pr[X = hh|Y_1 = 1]),$$

as a function of p and q. (This provides a general update rule for updating your estimate of the distribution of X, after observing a tall child.)

Hint: The new prior distribution $(p, q, 1 - p - q)$ only applies to Penelope. The remaining plants (from which we draw the plant to cross with Penelope) still have the distribution $(1/4, 1/2, 1/4)$.

(e) In part (d), you developed an update rule for the case where the child is observed to be tall. Now develop a general update rule for the case where the child is observed to be short. Suppose the prior distribution is $\Pr[X = HH] = p$, $\Pr[X = Hh] = q$, $\Pr[X = hh] = 1 - p - q$. With this prior, calculate the posterior distribution

$$(\Pr[X = HH|Y_1 = 0], \quad \Pr[X = Hh|Y_1 = 0], \quad \Pr[X = hh|Y_1 = 0]),$$

as a function of p and q.

Hint: The new prior distribution $(p, q, 1 - p - q)$ only applies to Penelope. The remaining plants (from which we draw the plant to cross with Penelope) still have the distribution $(1/4, 1/2, 1/4)$.

(f) Suppose that, after measuring, we find Penelope's first two children are both tall. Calculate the conditional distribution for X, given that Penelope's first two children are both tall: i.e., calculate the posterior distribution

$$(\Pr[X = HH|Y_1 = 1, Y_2 = 1], \quad \Pr[X = Hh|Y_1 = 1, Y_2 = 1], \quad \Pr[X = hh|Y_1 = 1, Y_2 = 1]).$$

Plot this distribution.

Hint: Apply the update rule from part (d) to the distribution you calculated in part (c).

(g) Another possibility is that, after measuring, we find that Penelope's first child is tall and Penelope's second child is short. What is the conditional distribution for X, given these measurements? Plot this distribution.

6. (25 pts.) Probabilistically Buying Probability Books

Chuck will go shopping for probability books for K hours. Here, K is a random variable and is equally likely to be 1, 2, or 3. The number of books N that he buys is random and depends on how long he shops. We are told that

$$\Pr[N = n | K = k] = \frac{c}{k}, \qquad \text{for } n = 1, \ldots, k$$

for some constant c.

(a) Compute c.

(b) Find the joint distribution of K and N.

(c) Find the marginal distribution of N.

(d) Find the conditional distribution of K given that $N = 1$.

(e) We are now told that he bought at least 1 but no more than 2 books. Find the conditional mean and variance of K, given this piece of information.

(f) The cost of each book is a random variable with mean 3. What is the expectation of his total expenditure? *Hint:* Condition on events $N = 1, \ldots, N = 3$ and use the total expectation theorem.