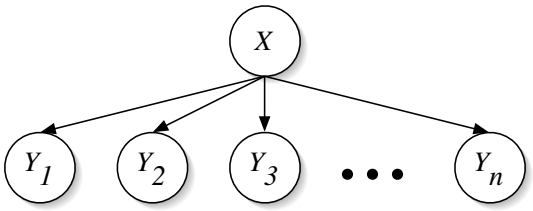
CSE 150. Homework 2 *Summer 2014*

Out: *Fri Aug 8*
Due: *Tue Aug 12*

2.1 Probabilistic reasoning

A patient is known to have contracted a rare disease which comes in two forms, represented by the values of a binary random variable $X \in \{0, 1\}$. Symptoms of the disease are represented by the binary random variables $Y_k \in \{0, 1\}$, and knowledge of the disease is summarized by the belief network:



The conditional probability tables (CPTs) for this belief network are as follows. In the absence of evidence, both forms of the disease are equally likely, with prior probabilities: $P(X=0) = P(X=1) = \frac{1}{2}$. In the first form of the disease $(X = 0)$, all the symptoms are uniformly likely to be observed, with $P(Y_k=0|X=0) = \frac{1}{2}$ for all k. By contrast, in the second form of the disease $(X = 1)$, the first symptom occurs with probability one,

$$P(Y_1=1|X=1) = 1,$$

while the k^{th} symptom (with $k \geq 2$) occurs with probability

$$P(Y_k=1|X=1) = \frac{f(k-1)}{f(k)},$$

where the function $f(k)$ is defined by

$$f(k) = 2^k + (-1)^k.$$

Suppose that on the k^{th} day of the month, a test is done to determine whether the patient is exhibiting the k^{th} symptom, and that each such test returns a positive result. Thus, on the k^{th} day, the doctor observes the patient with symptoms $\{Y_1=1, Y_2=1, \ldots, Y_k=1\}$. Based on the cumulative evidence, the doctor makes a new diagnosis each day by computing the ratio:

$$r_k = \frac{P(X=0|Y_1=1, Y_2=1, \ldots, Y_k=1)}{P(X=1|Y_1=1, Y_2=1, \ldots, Y_k=1)}.$$

If this ratio is greater than 1, the doctor diagnoses the patient with the $X=0$ form of the disease; otherwise, with the $X=1$ form. Compute the ratio r_k as a function of k. How does the doctor's diagnosis depend on the day of the month? Does the diagnosis become more or less certain as more symptoms are observed? Explain.

2.2 Zip codes

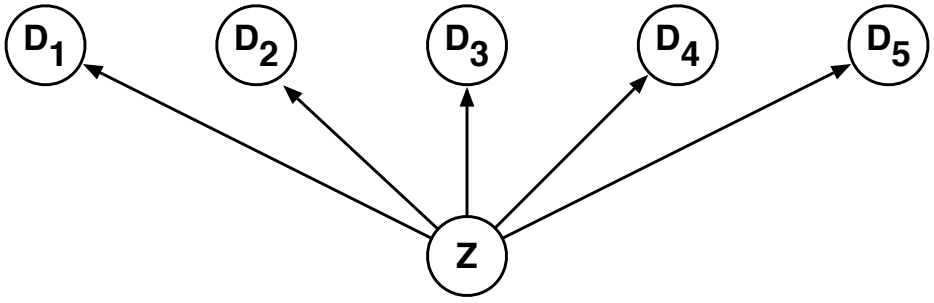
Consider the belief network shown below, where the random variable $Z \in \{0, 1, \ldots, 99999\}$ stores a five-digit zip code and the random variable $D_i \in \{0, 1, \ldots, 9\}$ reveals only the zip code's ith digit. Also, suppose that the zip codes are distributed accorded to their population density:

$$P(Z = z) \;=\; \frac{\text{POP}(z)}{\sum_{z'=0}^{99999} \text{POP}(z')},$$

where $\text{POP}(z)$ denotes the population of zip code z. Note that in this model the conditional probability tables for the random variables D_i are particularly simple:

$$P(D_i = d | Z = z) \;=\; \begin{cases} 1 & \text{if } d \text{ is the } i\text{th digit of } z, \\ 0 & \text{otherwise.} \end{cases}$$

Now imagine a game in which you are asked to guess the zip code, one number at a time, of a random person sampled from the census. The rules are as follows: after each number (0 through 9) that you guess, you'll be told whether the number appears in the zip code, and if it does appear, all the places that it occupies. Given the *evidence* you have at any stage in this game, the critical question is what number to guess next.



Let's work an example. Suppose that after three guesses—the numbers 6, 7, 4—you've learned that the number 7 does *not* appear, and that the numbers 6 and 4 appear as follows:

$$\underline{6} \quad \underline{\quad} \quad \underline{6} \quad \underline{\quad} \quad \underline{4}$$

Now consider your next guess: call it d. In this game the best guess is the value of d that maximizes

$$P\Big(D_2 = d \textbf{ or } D_4 = d \;\Big|\; D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4, 6, 7\}, D_4 \notin \{4, 6, 7\}\Big).$$

In other works, pick the value of d that is most likely to appear in the missing fields of the zip code. For any value of d we can compute this probability as follows:

$$\begin{aligned} & P\Big(D_2 = d \textbf{ or } D_4 = d \;\Big|\; D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4, 6, 7\}, D_4 \notin \{4, 6, 7\}\Big) \\ &= \sum_z P\Big(Z = z, D_2 = d \textbf{ or } D_4 = d \;\Big|\; D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4, 6, 7\}, D_4 \notin \{4, 6, 7\}\Big), \quad \boxed{\textbf{marginalization}} \\ &= \sum_z P(D_2 = d \textbf{ or } D_4 = d | Z = z)\, P(Z = z | D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4, 6, 7\}, D_4 \notin \{4, 6, 7\}), \quad \boxed{\textbf{product rule \& CI}} \end{aligned}$$

where in the third line we have exploited the conditional independence (**CI**) of the digits D_i given the zip code Z. Inside the sum there are two terms, and they are both easy to compute. In particular, the first term is more or less trivial:

$$P(D_2 = d \textbf{ or } D_4 = d | Z = z) \;=\; \begin{cases} 1 & \text{if } d \text{ is the second or fourth digit of } z \\ 0 & \text{otherwise.} \end{cases}$$

And the second term we obtain from Bayes rule:

$$P(Z = z | D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4,6,7\}, D_4 \notin \{4,6,7\})$$

$$= \frac{P(D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4,6,7\}, D_4 \notin \{4,6,7\} \mid Z = z) P(Z = z)}{P(D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4,6,7\}, D_4 \notin \{4,6,7\})} \quad \boxed{\textbf{Bayes rule}}$$

In the numerator of Bayes rule are two terms; the left term is equal to zero or one (depending on whether the evidence is compatible with the zip code z), and the right term is the prior probability $P(Z = z)$, as determined by the population density. The denominator of Bayes rule is given by:

$$\begin{aligned} & P(D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4,6,7\}, D_4 \notin \{4,6,7\}) \\ &= \sum_z P(Z = z, D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4,6,7\}, D_4 \notin \{4,6,7\}), \quad \boxed{\textbf{marginalization}} \\ &= \sum_z P(D_1 = 6, D_3 = 6, D_5 = 4, D_2 \notin \{4,6,7\}, D_4 \notin \{4,6,7\} \mid Z = z)\, P(Z = z), \quad \boxed{\textbf{product rule}} \end{aligned}$$

where again all the left terms inside the sum are equal to zero or one. Note that the denominator merely sums the prior probabilities of zip codes that are compatible with the observed evidence.

Now let's consider the general problem. Let E denote the evidence at some intermediate round of the game: in general, some numbers will have been guessed correctly and their places revealed in the zip code, while other numbers will have been guessed incorrectly and thus revealed to be absent. There are two essential computations. The first is the *posterior* probability, obtained from Bayes rule:

$$P(Z = z | E) \;=\; \frac{P(E | Z = z)\, P(Z = z)}{\sum_{z'} P(E | Z = z')\, P(Z = z')}.$$

The second key computation is the *predictive* probability, based on the evidence, that the number d appears somewhere in the zip code:

$$P\Big(D_i = d \text{ for some } i \in \{1,2,3,4,5\} \Big| E\Big) \;=\; \sum_z P\Big(D_i = d \text{ for some } i \in \{1,2,3,4,5\} \Big| Z = z\Big) P\Big(Z = z \Big| E\Big).$$

Note in particular how the first computation feeds into the second. Your assignment in this problem is implement both of these calculations. **Turn in a printout of your source code along with answers to the following questions.** Note that you may program in the language of your choice.

(a) Download the file *zipcode.txt* that appears with the homework assignment. The file contains a list of zip codes and their populations from the 2010 census. (Note that some zip codes do not appear in this list; for this assignment you should assume that missing zip codes have zero probability.) From the data given compute the prior probability $P(z) = \text{POP}(z)/\text{POP}_{\text{total}}$. **As a sanity check, print out the <u>five</u> most populated zip codes, and confirm that they correspond to heavily populated areas.**

(b) Consider the following stages of the game. For each of the following, indicate the best next guess—namely, the number d that is most likely (probable) to be among the missing digits. Also report the probability $P(D_i = d \text{ for some } i \in \{1, 2, 3, 4, 5\} | E)$ for your guess d.

(i) **First guess:** ____ ____ ____ ____ ____

(ii) **None correctly guessed:** ____ ____ ____ ____ ____
Incorrectly guessed: 0, 4

(iii) **Correctly guessed:** 8 ____ ____ 7 ____
Incorrectly guessed: none

(iv) **Correctly guessed:** 9 2 ____ ____ 9
Incorrectly guessed: 5, 6

(v) **Correctly guessed:** ____ ____ 7 0 3
Incorrectly guessed: 4, 5, 8, 9