

Homework set 4

David Wolfe

May 5, 2003
Due: May 16, 2003

When you submit this problem set, please submit all solutions to all problems related to the talks, including your own problem.

This problem set has you use the program, *JavaBayes* to help create and investigate a simple belief network. To run it, be sure that the current directory `.` is in your `CLASSPATH`. For example,

```
setenv CLASSPATH /usr/local/lib/java:.
```

Then, go to the directory and run *JavaBayes* by typing,

```
cd ~wolfe/public/85/JavaBayes/Classes
java JavaBayes
```

This will launch two windows. Although documentation for *JavaBayes* can be found via the course homepage, I recommend reading the Console window and playing around. A few things are worth noting:

1. Some buttons are along the top as well as the bottom of the Editor window.
2. You need to hit **Enter** or **Return** to have a change entry you make in a text field take effect.
3. You'll start by Creating nodes, then Editing variables, then Editing functions, then using Query and Observe on the network. Do read the messages in the Console window to learn how to use each button.

-
1. (10 points) Do problems 14.3 and 14.11 from the text
 2. (100 points) In this problem, your goal is to write a program which helps someone make predictions about blood types.

ABO blood group: For ABO blood group genes there are three alleles, A, B and O. OO individuals have type O blood. AO and AA individuals have type A blood. BO and BB individuals have type B blood. AB individuals have type AB blood. For this problem, assume that the frequency of the alleles is 67% O, 26% A and 7% B. (This is an estimate of the frequency in the American white population.)

Rh: A person can be Rh- or Rh+. Rh+ is dominant, and Rh- is recessive. In other words, if your alleles are ++ or +-, you are Rh+, but if your alleles are --, you are Rh-.

There are risks associated with an Rh- mother giving birth to an Rh+ baby. In particular, there is some risk with the first Rh+ baby she has, and much greater risk with the second. Assume that 84% of the population is Rh+, and 16% is Rh-; from this you can compute the probability that an allele is + or -.

This problem consists of a few warm-up exercises, and I recommend you do them both by hand and by using *JavaBayes*. You may want to check with me to make sure you get the warm-ups right before moving on to the later problems.

For the later problems, use *JavaBayes*. Assume that any person whose blood type is unknown has alleles randomly determined with the frequencies given above. Show all your work; in particular, be sure to diagram any Bayesian networks you use.

- (a) (warm-up 10%) Compute, by hand, the probability that a person has blood-type A.
- (b) (warm-up 10%) I have blood-type A. Compute, by hand, the probability that my ABO blood group alleles are both A?
- (c) (warm-up 10%) I have blood-type A. Use *JavaBayes* to determine the probability my mother is of type A.
- (d) (10%) For the last question, and the questions below, you may have noticed that you have several choices for how to represent the states of your network. The fewer states you use, the larger the probability tables you'll need. Sometimes the tradeoffs make few states worthwhile, sometimes it's less work to use more states. Discuss at least two alternatives and justify the choices you make. (You'll probably want more than one state for an individual.)

- (e) (15%) Suppose Alice is pregnant and she is Rh-negative (Rh-). She knows also that her husband is Rh+, and her mother-in-law is as well. (She doesn't know her father-in-law's Rh). What is the probability the fetus is Rh-?
- (f) (15%) Suppose further Alice's first child is born and is Rh+. She's gets pregnant for a second time. What is the probability her second child is Rh+ as well.
- (g) (15%) New scenario: Suppose Alice is pregnant and is Rh-. Tragically, the father abandoned Alice, and Alice doesn't know his Rh factor. But she interviews some of his relatives. After finding out a few of the Rh factors of his relatives, a curious thing happens. She calculates the probability her unborn child is Rh+, then asks one more relative who turns out to be Rh+, and the probability that her unborn child is Rh+ goes *down!* How is this possible? Describe your logic qualitatively, and then quantitatively justify your answer by using JavaBayes.
- (h) (15%) Compose a family tree in which there are two nodes in your network, call them *A* and *B*, which are analogous to the *Gas* and *Radio* on page 445. In particular, you should be able to exercise all four cases mentioned (there are two cases in case 3 of the example.) Confirm the independence assumptions numerically by *observing* node *A* (i.e., forcing the node's value), and checking whether node *B*'s probabilities change as a result of the observation.
3. (**Optional for extra credit**) In this open-ended problem, your goal is to write a program which helps someone do genetic predictions. We'll use an example, which is of historical significance in biology, for it represents the first demonstration of a linkage between two traits in humans that are not on the X- or Y-chromosome. Because of the historical significance of the example, there is lots of information on the linkage that can be understood by a layman.

Your goal is to predict the probability that a woman's child will have nail patella syndrome given the family history. First, you'll need some background.

Genetics definitions: A human has 23 *chromosomes*, each with two strands of DNA. A *gene* is a segment of the chromosome which is a fundamental unit of heredity, and occupies some location or *locus* (plural *loci*) on the chromosome. Each gene can have several different forms, or *alleles*. There are often just two possible alleles (normal and abnormal), but often an alleles can have more than two forms. A gene is recessive if both (abnormal) alleles are required to cause the abnormality. A gene is dominant if only one allele is required. Each person passes only one allele to an offspring.

Genetic mutations: In the simplest case, one thinks of a child as receiving one DNA strand from each parent. For a particular gene, this means you get one allele from your mother and one from your father. But there are a number of complicating factors, called genetic mutations.

For this problem, you need only consider one types of mutation called cross-over mutation. In cross-over mutations, you don't get a full strand of DNA from one parent, but rather part(s) of each strand. For each gene, however, you get only one allele from the parent in question.

ABO blood group: For ABO blood group genes there are three alleles, A, B and O. OO individuals have type O blood. AO and AA individuals have type A blood. BO and BB individuals have type B blood. AB individuals have type AB blood. For this problem, assume that the frequency of the alleles is 67% O, 26% A and 7% B. (This is an estimate of the frequency in the American white population.)

Nail patella syndrome: Nail patella syndrome is a dominant disorder resulting in malformed nails and kneecaps, sometimes with elbow and kidney abnormalities. To make this problem a little more interesting, however, **assume nail patella syndrome is recessive** The disease occurs in 1 in 50,000 people. If you locate the Online Mendelian Inheritance in Man (OMIM), you can read more about the disease. Denote + as a normal allele, and - as abnormal. In actuality, the disease is dominant, so those with the disease are either +- or -. But for this problem, we are pretending the disease is recessive, so only - *would* have the disease.

Genetic distance: There are a number of ways to measure the genetic distance between two genes, but the goal of the measure is to approximate the probability that a cross-over mutation occurs between the two loci. In particular, you can find the *recombination percentage* between ABO blood group and nail patella syndrome in the OMIM article mentioned above.

The recombination percentage between two different genes is the probability that the genes come from the same segment of DNA. A recombination percentage of 50% means that two genes are not linked (i.e., they are on separate chromosomes, or they are so far apart on one chromosome that enough crossovers will almost surely occur between the loci that they might as well be on different genes.) A smaller recombination percentage means the genes are linked.

For this problem we'll make the following assumptions:

- As mentioned above, assume (wrongly) that the nail patella syndrome is recessive.
 - The recombination rate between nail patella syndrome and ABO blood group is 10%.
 - The frequency of occurrence of the alleles in the ABO blood group and of the nail patella abnormalities are given above.
- (a) Assume someone wants to use inference to answer questions such as, “my maternal grandfather has nail patella syndrome, and I know the ABO blood group of most members of my immediate family. What is the probability my son will be born with nail patella syndrome?”
Carefully decide how many states you'll use to represent each individual in the family tree. There are a number of possible choices to make, so choose carefully.
- (b) Write a program which takes family tree information and node location information, and outputs a JavaBayes network. The network should have all the genetic information required for the queries of the type mentioned above, and should complete the probability tables in the network.
- (c) Compose an interesting family tree, and make a few interesting queries to the network outputted by your program.