CSE 573 Homework 3

Due In Class

November 13, 2006

Please type or clearly write your answers to the following questions, being concise when possible. You may discuss the questions with other classmates but your answers must be written up individually.

1. (10 pts) Russell & Norvig, 13.11
   Suppose you are given a bag containing n unbiased coins. You are told that $n-1$ of these coins are normal, with heads on one side and tails on another, whereas one coin is a fake, with heads on both sides.

   (a) Suppose you reach into the bag, pick out a coin uniformly at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is the fake coin?

   (b) Suppose you continue flipping the coin for a total of k times after picking it and see k heads. Now what is the conditional probability that you picked the fake coin?

   (c) Suppose you wanted to decide whether the chosen coin was fake by flipping it k times. The decision procedure returns FAKE if all k flips come up heads, otherwise it returns NORMAL. What is the (unconditional) probability that this procedure makes an error?

2. (10 pts) Russell & Norvig, 14.7
   This exercise is concerned with the variable elimination algorithm in Figure 14.10.

   (a) Section 14.4 applies variable elimination to the query

   $$P(Burglary|JohnCalls = true, MaryCalls = true)$$

   Perform the calculations indicated and check that the answer is correct.

   (b) Count the number of arithmetic operations performed, and compare it with the number performed by the enumeration algorithm.

   (c) Suppose a network has the form of a *chain*: a sequence of Boolean variables $X_1, \ldots, X_n$ where $Parents(X_i) = \{X_{i-1}\}$ for $i = 2, \ldots, n$. What is the complexity of computing $Parents(X_1|X_n = true)$ using enumeration? Using variable elimination?

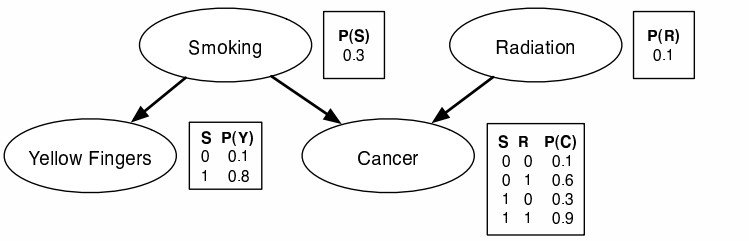

Figure 1: A Bayes Net, B

3. (10 pts) Consider the Bayesian network, B, in Figure 1 where all variables are Boolean, and have associated conditional probability tables as shown.

   (a) What is the Markov blanket of *Cancer*n in B?

   (b) Is *Radiation* independent of *Yellow Fingers* in B? Is *Radiation* independent of *Yellow Fingers* given *Cancer* in B?

   (c) What is the probability of *Cancer* given *Radiation* and *Yellow Fingers*? Given *Radiation* and not *Yellow Fingers*?

   (d) Using likelihood weighting to compute the probability distribution of *Yellow Fingers* and *Radiation* given *Smoking* and not *Cancer*. What weight would you give to the sample (*Smoking = True, Radiation = False, Yellow Finger = True, Cancer = False*)?

   (e) Convert B to an equivalent Markov network, using one potential function per maximal clique.

   (f) Let G be the graph of this Markov network. What is the Markov blanket of *Yellow Fingers* in G?

   (g) According to G, is *Smoking* independent of *Radiation*?

4. (20 pts) In this problem, you will use Alchemy to experiment with MCMC in the context of *entity resolution*, the problem of determining which records in a database refer to the same objects. For example, different atoms may contain the strings `ICDM-2006, Sixth ICDM` and `IEEE ICDM'06`, all of which refer to the same conference.

   You are given a *citation database* where each citation (or bib entry) contains the attributes *Author, Title* and *Venue*, as well as a model for performing entity resolution in the form of an MLN. The model assumes that each attribute consists of one or more words, and defines the predicate *HasWordAuthor(author,word)* which is true iff *author* contains *word*. The predicates *HasWordTitle(title,word)* and *HasWordVenue(venue,word)* are similarly defined.

The goal is to determine, for each pair of constants c_1, c_2 of the same type, if $c_1 = c_2$. Thus the *query predicates* will be the equality predicates, *SameBib, SameAuthor, SameTitle* and *SameVenue*, and the evidence predicates will be the set of all non-equality predicates (*e.g. Author, Title, Venue, HasWordAuthor, HasWordTitle* and *HasWordVenue*).

You will use Gibbs sampling to calculate the conditional probabilities of the query atoms given the evidence.

(a) Use the `infer -p` option in Alchemy to compute the conditional probabilities of all 4 query predicates given all non-equality predicates as evidence. Plot the average conditional log-probability (CLP) of $SameBib(b1, b2)$ after { 10, 100, 1000, 10,000, 100,000, 1,000,000} sampling steps. The CLP of a predicate is the average log-probability of the actual truth values of all possible ground atoms.

Does the CLP appear to have converged? Run this procedure 2 more times, and show the results of all 3 runs on a single graph. The x-axis on your graph will be the number of samples, the y-axis will be average CLP, and each run is one line on the graph. For this question, initialize the starting values uniformly at random by specifying the `-mcmcWalksatType 0` option.

(b) For the same query and evidence in (a), compare the convergence of the Gibbs sampler when it is initialized

- Uniformly at random (using your results from part (a))
- According to a satisfying assignment found by WalkSat (by specifying `-mcmcWalksatType 1`)

Construct the same graph described in (a). Does convergence appear to have been affected by the initialization method?

(c) Now suppose we remove the following sets of predicates from the evidence one at a time. (Do this by commenting out any rule in the MLN that contains one of the predicates in the set.)

i. {*Author, HasWordAuthor*}
ii. {*Title, HasWordTitle*}
iii. {*Venue, HasWordVenue*}

What effect do you expect to see on the ability to predict when two citations refer to the same entry after removing each of the different sources of evidence?

For each set of evidence, repeat the procedure outlined in part (a). You may run one trial per set instead of three. Graph the results for each set of evidence as you did in (a). For this comparison, use whatever initialization method you found to converge faster in part (b) and note which you used.