COMP150AIH: Artificial Intelligence in Health Informatics **Fall 2011**

Problem Set 1

Handed Out: September 19,2011 *Due: September 28/30,2011*

- Homework is due either physically, in class, September 28 or, if you'd prefer, electronically (via e-mail) by midnight on Friday the 30th.
- We encourage you to discuss the homework with other members of the class. The goal of the homework is for you to learn the course material. However, you should write your own solution.
- Please keep your solution brief, clear, and legible.
- We encourage you to ask questions before class, after class, or via email. Please don't start e-mailing us questions the night before the homework is due, though.

1. (Shortliffe & Cimino, Chapter 3, Problem 1)

   Calculate the following probabilities for a patient about to undergo CABG surgery:

   (a) The only possible, mutually exclusive outcomes of surgery are death, relief of symptoms (agina and dyspnea), and continuation of symptoms. The probability of death is 0.02, and the probability of relief of symptoms is 0.80. What is the probability that the patient will continue to have symptoms?

   (b) Two known complications of heart surgery are stroke and heart attack, with probabilities of 0.02 and 0.04, respectively. The patient asks what chance he or she has of having *both* complications. Assume that the complications are conditionally independent, and calculate your answer.

2. (Shortliffe & Cimino, Chapter 3, Problem 4)

   You have a patient with cancer who has a choice between surgery or chemotherapy. If the patient chooses surgery, he or she has a 2 percent chance of dying from the operation (life expectancy = 0), a 50 percent chance of being cure (life expectancy = 2 years), and a 48 percent chance of not being cured (life expectancy = 1 year). If the patient chooses chemotherapy, he or she has a 5 percent chance of death (life expectancy = 0), a 65 percent chance of cure (life expectancy = 15 years), and a 30 percent chance that the cancer will be slowed but not cured (life expectancy = 2 years). Create a decision tree. Calculate the expected value of each option in terms of life expectancy.

3. (Shortliffe & Cimino, Chapter 3, Problem 5)

   You are concerned that a patient with a sore throat has a bacterial infection that would require antibiotic therapy (as opposed to a viral infection, for which no treatment is available). Your treatment threshold is 0.4, and based on the examination you estimate the probability of bacterial infection as 0.8. A test is available (TPR = 0.75, TNR = 0.85) that indicates the presence or absence of bacterial infection. Should you perform the test? Explain your reasoning. What additional factors might alter your analysis?

4. For this problem (and the next), we will be working with the `pima-indians-diabetes` UCI dataset.[1] You are encouraged to peruse the UCI documentation regarding this dataset. We have pre-processed the data for discretization purposes, available at `pima-indians-diabetes` which is defined by the following variables

Variable	Notation	Values	Description
pregnant	P	$0, 1, 2, >= 3$	see UCI documentation
age	A	$< 30, 30 - 39, 40 - 49, >= 50$	
heredity	H	low, middle, high	
BMI	BMI	low, normal, overweight, obese	
diabetic	D	true, false	
skin	S	low, middle, high	
glucose	G	low, middle, high	
insulin	I	low, middle, high	
BP	BP	desirable, hypertensive	

Table 1: Discretized Version of `pima-indians-diabetes` Data

Consider the Bayesian network defined by Figure 1.[2]

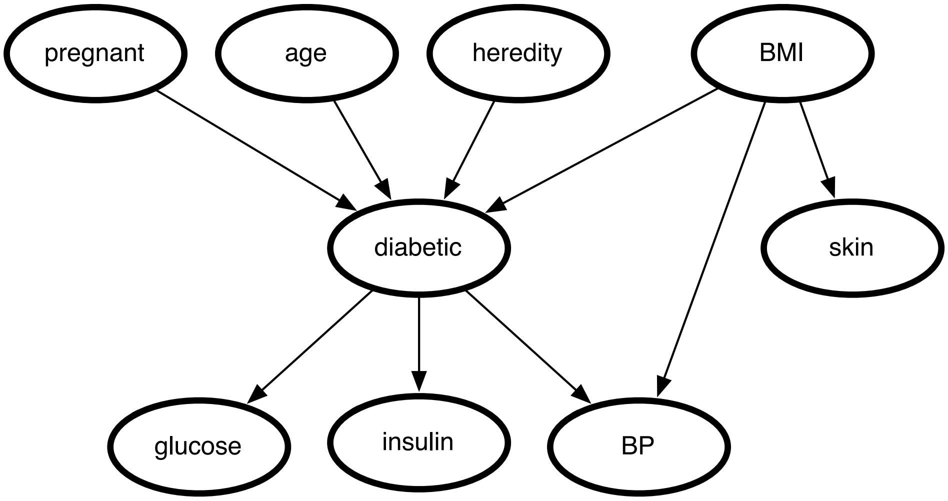

Figure 1: Proposed Bayesian Network for Pima Indians Diabetes Data

(a) Compute the number of parameters required for the conditional probability table associated with each network vertex (random variable). Compute the number of parameters required to represent the joint probability distribution.

[1] The UCI Machine Learning Repository is a good place to get small data sets. The specific page for the Pima Indians Diabetes data is `http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`

[2] Modified from `http://www.cs.utsa.edu/~bylander/cs6243/bayes-example.pdf`

(b) Compute $p(I = middle|P = 1, A = 30 - 39, BMI = overweight)$.

NB: While you are welcome to estimate these parameters from `file` however you choose, I would use variants of the following commands:

```
$ cut -d, -f8 pima-indians-diabetes-discretized.data | sort | uniq -c
 396 age<30
 165 age=30-39
 126 age=40-49
 81 age>=50
  f
$ grep "pregnant=1" pima-indians-diabetes-discretized.data | \
grep "age=30-39" | grep "bmi=overweight"
```

(c) Compute $p(H = high|P = 0, A = 20 - 29, BMI = normal, D = true)$.

(d) Assume causal edge semantics: in other words, assume that an arrow from A to B implies A *causes* B. If we were to add a latent (i.e., unobserved) *fitness* variable, how would you modify the network? Are there any other modifications you would make to the network in Figure 1? Explain your reasoning.

5. For this problem, we are going to experiment with the Naïve Bayes classifier. You are welcome to use any available tools to compute these values – we recommend Weka.[3] The training/testing datasets are available in the course website direction `http://www.cs.tufts.edu/comp/150AIH/hw1` as:

- `pima-indians-diabetes-discretized-training.arff`
- `pima-indians-diabetes-discretized-testing.arff`

(a) Draw the corresponding Bayesian Network. How many parameters are required to represent the resulting distribution?

(b) What is naive about it, in this case (state this in words; what assumptions are you making)? How might you make this *slightly* less naïve? Why not drop the naïvety completely?

(c) Write down the resulting confusion matrix based on the given training and testing split.

(d) Compute the accuracy, sensitivity and specificity. Discuss the relative utility of these three measurements for this dataset. Which factors should be used to determine which may be most appropriate?

bonus Plot the ROC curve.[4] Interpret the result.

[3] `http://www.cs.waikato.ac.nz/ml/weka/`

[4] You can do this via Weka or some other package; don't bother writing the code to generate the ROC (unless you want to, of course).