University of Pittsburgh
CS 2740 Knowledge Representation
Professor Milos Hauskrecht

Handout 16
November 24, 2008

Term project assignment

Due on: Thursday, December 11, 2008 at noon

A probabilistic model for medical diagnosis

Please read the book chapter by Harry Pople, *Heuristic methods for imposing the structure on ill-structured problems, the structuring of medical diagnostics*, distributed during the class.

This chapter was published in 1982 in the book *Artificial Intelligence in Medicine*, edited by Peter Szolovits. The work described in the chapter has close ties to the University of Pittsburgh. Internist was one of the early knowledge based system for representing and reasoning with medical knowledge and was developed at the University of Pittsburgh. Caduceus was a planned refinement of Internist that to my knowledge has never been implemented. The Internist later morphed into the QMR system which was commercially sold as a tool to support diagnosis in internal medicine.

The Internist and Caduceus projects did not use probabilistic models to model uncertainty. Instead they relied on uncertainty scores referred to as evoking strengths and frequency weights (see page 137). If you are surprised by this, please note that the paper was published in 1982 which is well before Bayesian belief networks were proposed.

The goal of this project is to take some of the ideas in Poples paper and try to represent them with a probabilistic model.

Part a. Book chapter abstract

Please write a short abstract of the paper outlining its key ideas. Also please write one paragraph with your opinion about what you liked (or disliked) about the paper and the reasons why.

Part b. Hierarchies

One of the ideas proposed in the paper is the use of hierarchies to organize diseases into categories. Two types of hierarchies are considered (see Figure 3 in the paper): nosology (organization with respect to the organ involvement) and etiology that organizes the diseases with respect to the agent causing the disease. The hierarchical structures in the paper do not have any probabilistic interpretation. Propose a probabilistic model based on the BBN that would reflect the relationship among diseases and their categories. Assume each disease and category is represented as a binary variable. Explain your solution.

Part c. Evidence

Assume a set of evidence variables representing findings, lab tests, symptoms that can be either true or false. Assume these evidence variables are caused by diseases. Explain how would you combine the evidence variables into the probabilistic model you have proposed in part b. Use relations expressed in Figure 4 to illustrate your solution. Please note that some of the evidence variables should be more appropriately linked to a disease category rather than a specific disease similarly to the graphs in Figure 4.

Does your model support the following inferences:

- $P(HepatobiliaryInvolvement = T | jaundice = T)$,
- $P(FibroticHepatocellularInvolvement = T | Jaundice = T)$,
- $P(FibroticHepatocellularInvolvement = T | Jaundice = T, Pallor = T)$.

Please explain how you would calculate the above probabilities from the parameters of your model.

Part d. Intermediate variables

The medical domain is more complex and the disease variables organized in the disease hierarchy together with the evidence variables may not be sufficient to model the complexity and richness of the domain. In Figure 5, Pople considers intermediate variables that are related to important medical conditions such as portal hypertension in Figure 5a, or hyperbilirubinemia in Figure 5b. These variables may be structured into hierarchies of their own. Propose a probabilistic model that allows these intermediate variables and their hierarchies to be included in the model and illustrate your solution on the problem in Figure 5b.

Does your model support the following inferences?

- $P(HepatobiliaryInvolvement = T|jaundice = T)$,
- $P(HepatocellularInvolvement = T|Jaundice = T)$.

Please explain how you would calculate the above probabilities from the parameters of your model.

Part e. Differential diagnosis.

The paper argues the task a physician really solves when diagnosing a patient is that of the differential diagnosis. Differential diagnosis consists of a set of diseases that are considered to be alternatives in explaining the evidence E. Probabilistically the differential diagnosis task would be represented by a conditional probability distribution where the outcome variable is the identity of one of the diseases in the differential, that is, $P(d = d_1|E), P(d = d_2|E), P(d = d_k|E)$ such that $d_1, d_2, \ldots d_k$ are diseases in the differential. Please note that if you have followed the instructions in previous parts, every disease in your model is represented as a binary random variable, so the model really does not support the differential diagnosis model and the distribution associated with the differential.

Suggest a solution for calculating the distribution for the differential diagnosis from your probabilistic model. The solution does not have to be an exact solution it may be an approximation. Please explain why you believe your solution is a good approximation. If you see any cases in which your approximation may fail please explain them as well.

Part f. Multiple diseases.

One point the paper argues is that in some instances a combination of multiple diseases may better explain the findings. Let E be evidence and D_1 and D_2 two variables modeling two diseases. One possibility to decide in between the hypotheses with one or two diseases is to compare $P(D_1 = true|E)$ vs. $P(D_1 = true, D_2 = true|E)$ and choose the option that gives a higher probability. What is wrong with this suggestion? Is there a better solution to solve the problem and can help us to decide if a simple or more complex hypothesis should be preferred?