This problem set is handed out a day after sixth lecture(LEC#6) and due two weeks later.

Goals

This assignment is meant to teach you about methods of reasoning and learning under uncertainty. The main part focuses on Bayes Networks, a graphical formalism that represents probabilistic relationships among random variables. The paper is broken down into 4 sections:

Background: Summary of the basic notions and definitions of Bayesian networks.

Representation: A problem requiring the use of Bayesian networks to represent expert medical knowledge.

Reasoning: A problem requiring the use of Bayesian networks to reason under uncertainty.

Learning: A problem requiring the use of Bayesian networks to learn from a database.

Background

Much of what we know about the world is known with various degrees of certainty. Although many different numerical models have been proposed for representing this uncertainty, many researchers today believe that probability theory—the most extensively studied uncertainty calculus—is an appropriate basis for building computerized reasoning systems.

Mathematical Preliminaries

We take the world of interest to consist of a set of random variables, $X = \{X_1, X_2, \ldots, X_n\}$. Each of the X_i can take on one of a discrete set of values, $\{x_{i1}, x_{i2}, \ldots, x_{ik_i}\}$.[1] Each possible combination of assignments of values to each of the variables represents a possible state of this world. There are $\prod_{i=1}^{n} k_i$ such states. This is clearly

[1] Formulations using continuous variables are also possible, but we will not pursue them here. A continuous variable can, of course, be approximated by a large number of discrete values.

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support

exponential in the number of variables; for example, if each variable is binary, all of the $k_i = 2$ and the number of states is 2^n. Each state may be identified with the particular values that it assigns to each variable. We might have, for example, $S_{12} = \{X_1 = a, X_2 = e, \ldots, X_n = b\}$.

A probability function, P, assigns a probability to each of these possible states. The probability for each state, $P(S_i)$, is the *joint probability* of that particular assignment of values to the variables. All states are distinct, and they exhaustively enumerate the possibilities; therefore,

$$\sum_{j=1}^{\prod_{i=1}^{n} k_i} P(S_j) = 1.0$$

We will often be interested in the probability not of individual states, but of certain combinations of particular variables taking on particular values. For instance, we may be interested in the probability that $\{X_3 = a, X_5 = b\}$. In such cases, we wish to treat other variables as "don't cares." We call such a partial description of the world a *circumstance*,[2] C, and the variables that have values assigned the *instantiation-set* of the circumstance, $I(C)$. The probability of a circumstance C can be computed by summing the probabilities of all states that assign the same values to the variables in the instantiation set $I(C)$ as C. If we re-order variables $\mathcal{X}$ so that the first m are the instantiation set of C, then (in a shorthand notation):

$$P(C) = \sum_{X_{m+1}} \ldots \sum_{X_n} P(S).$$

For example, if the world consists of four binary variables, W, X, Y and Z, then the circumstance $\{X = \text{true}, Z = \text{false}\}$ is given by

$$P(\{X = \text{true}, Z = \text{false}\}) = \sum_{u \in \{\text{true},\text{false}\}} \sum_{v \in \{\text{true},\text{false}\}} P(\{W = u, X = \text{true}, Y = v, Z = \text{false}\})$$

The computational difficulty of calculating the probability of a circumstance comes precisely from the need to sum over a possibly vast number of states. The number of such states to sum over is exponential in the number of "don't cares" in a circumstance.

Defining P for each state of the world is a rather tedious and counter-intuitive way to describe probabilities in the world, although the ability to assign an exponentially large number of independent probabilities would allow any probability

[2]"Circumstance" is not a commonly-used term for partial descriptions of the world. People use terms such as "partially-specified state" and other equally unsatisfying terms.

distribution to be described. It might happen that the world really *is* so complex that only such an exhaustive enumeration of the probability of each state is adequate, but fortunately many variables appear to be *independent*. For example, in medical diagnosis, the probability that you get a strep infection is essentially independent of the probability that you catch valley fever (a fungal infection of the lungs prevalent in California agricultural areas). Formally, this means that $P(\text{strep}, \text{vf}) = P(\text{strep})P(\text{vf})$. When two variables both depend on the same set of variables but not directly on each other, they are *conditionally independent*. For example, if an infection causes both a fever and diarrhea, but there is no other correlation between these symptoms, then $P(\text{fever}, \text{diarrhea}|\text{infection}) = P(\text{fever}|\text{infection})P(\text{diarrhea}|\text{infection})$.

The independencies among variables in a domain support a convenient graphical notation, called a *Bayes network*. In it, each variable is a node, and each probabilistic dependency is drawn as a directed arc to a dependent node from the node it depends on. This notion of probabilistic dependency does not necessarily correspond to what we think of as causal dependency, but it is often convenient to identify them. When there is no directed arc from one variable to another, then we say that the second does not depend on the first.

The probability of a state is a product over all variables of the probability that the variable takes on its particular value (in that state) given that its parents take on their particular values:

$$P(X_1 = v_1, \ldots, X_n = v_n) = \prod_{i=1,\ldots,n} P(X_i|\pi(X_i))$$

The right hand side is an abbreviation for

$$\prod_{i=1,\ldots,n} P(X_i|\pi(X_i)) = \prod_{i=1,\ldots,n} P(X_i = v_i|X_j = v_j, \ldots, X_l = v_l)$$

where $\pi(X) = \{X_j, \ldots, X_l\}$. As described above, to find the probability of a circumstance, we must still sum over all of the states that are consistent with the circumstance—i.e., a number of states exponential in the number of “don’t cares.”

Software

The program to use for the exercises is Bayesware Discoverer. A free Student Edition is available from http://bayesware.com. The data for the last exercise are available from the course homepage, as well the related paper by Wolberg et al.

Representation

Question 1. Your expert supplies you with the following assessment: ”I think that disease D_1 causes symptoms S_1 and, together with disease D_2, causes symptom S_2. 80% of patients affected by disease D_1 will manifest symptom S_1, which is present in the 10% of the patients without disease D_1. D_1 and D_2 cause the occurrence of symptom S_2 in 60% of cases, when only D_1 is present S_2 occur in the 70% of patients, when only D_2 is present S_2 occur in the 80% of patients, when neither D_1 or D_2 occur symptom S_2 occurs in the 10%. Disease D_1 occurs in the 20% of the population, while disease D_2 occurs in the 10%.

1. Draw the network capturing the description.
2. What kind of graph is this and why?
3. Assume that the variables can take two values, 1 for present, 0 for absent. Write the conditional probability tables associated to each dependency in the network.
4. Write down the formula to calculate the joint probability distribution induced by the network.
5. Using these table, calculate the marginal probability $p(S_1 = 1)$.

Reasoning

Question 2. Consider the network in Figure 1.

1. How are termed the networks with this kind of structure.
2. List the assumptions of conditional independence encoded by this network.
3. Assume that the distributions of the network as are as follows:

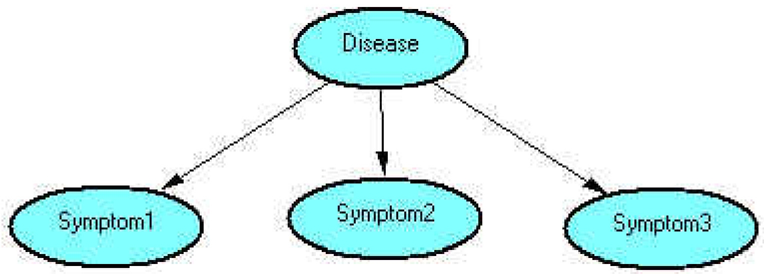

Figure 1: Bayesian network for Reasoning exercise.

Disease

True	False
0.100	0.900

Symptom 1

Disease	True	False
True	0.800	0.200
False	0.100	0.900

Symptom 2

Disease	True	False
True	0.700	0.300
False	0.200	0.800

Symptom 3

Disease	True	False
True	0.900	0.100
False	0.200	0.800

(a) List the probability distributions of each symptom given that the Disease is True.

(b) Compute the posterior probability distribution of the Disease variable given that Symptom1 is True, Symptom2 is True and Symptom 3 is False.

(c) Has any of the symptoms enough evidential power to make True the Disease variable by itself, or you always need a combination of variables to make positive your diagnosis?

Learning

Question 3. Consider the database Wisconsin Breast Cancer. The database is as following:

1. Number of Instances: 699 (as of 15 July 1992)

2. Number of Attributes: 10 plus the class attribute

3. Attribute Information: (class attribute has been moved to last column)

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

The original description of the database is in:

Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences*, *87*, 9193–9196.

Assess whether the conclusions of the paper are consistent with the structure of the learned

1. Learn the Bayesian network from the database.

2. Encode the model suggested by the original paper as a Bayesian network.

3. List the differences between the two models, if any.

4. How important are these differences and how they change the predictive power of the model, i.e. the ability to predict the malignancy of the tumor given the evidence. *hint*: Try to propagate different instances of the symptom variables and assess how different are the posterior distributions of the variable `Class`.