6.891 Review Problems

5/16/01

1 Bayes' Nets

I am a professor trying to predict the performance of my class on an exam. After much thought, it is apparent that the students who do well are those that studied and do not have a headache when they take the exam. My vast medical knowledge leads me to believe that headaches can only be caused by being tired or by having the flu. Studying, the flu, and being tired are pairwise independent.

a) We will model everything with Boolean variables. F indicates the presence of the flu, T indicates being tired, H - having a headache, S - studying, and E - passing the exam. Which of the following three networks best models the relationships described?
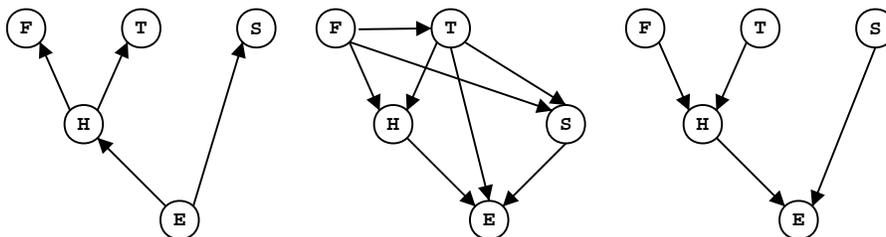
Figure 1: From left to right, models 1, 2, and 3

b) Why were the other two networks unsatisfactory models? Explain the deficiencies of each in terms of the conditional independence and dependence relationships they model. Which one of the remaining models represents an equivalent joint probability table as the best model, given that the description of the relationships was accurate?

c) I found that tiredness and having the flu each have a small impact on the likelihood of studying (small because MIT students are so tough). Draw a network that expresses this connection. Compute its complexity and the complexity of the network you choose in part a). Give two reasons why

the original network is superior, despite the *small* improvement this new network gives in predictive power.

d) Leslie got the flu. Using model 3, compute the probability that she will fail the exam, in terms of values that are available in the conditional probability tables stored at each node.

e) Michael passed the exam. Using model 3, compute the probability that he studied, in terms of values that are available in the conditional probability tables stored at each node.

2   True & False

1. For any A and B $P(A) = P(A|B)P(B) + P(\neg A|B)P(B)$

2. Every continuous function can be computed by some multi-layer perceptron with arbitrary unit functions.

3. Bayes' nets only represent causal relationships between variables.

4. For any A and B $P(A) > P(A, B)$

5. Perceptrons are sensitive to their learning rate when learning linearly separable classes.

6. If we use mean-squared error to train a perceptron, we will always successfully separate linearly separable classes.

7. Over-fitting is only a problem with noisy data.

8. If A and B are independent, $P(A, B) = P(A)P(B)$

9. Multi-layer perceptrons are equivalent to single-layer perceptrons unless they contain non-linear hidden units.

3   The Prisoner's Dilemma

The Prisoner's Dilemma is a well-known problem in game theory. Two thieves (partners in crime) are arrested and held in custody separately. The police offer each the same deal. Inform on your partner and we will reduce your sentence. The following outcomes and costs are possible:

1. If both you and your partner stay quiet, you will both be convicted of misdemeanor offenses (lesser charges). The cost of this is 10.

2. If you turn state's evidence (cooperate with the police), you will be convicted of a misdemeanor and fined. The cost of this is 50.

3. If you do not cooperate, but your partner does, you will be convicted of a felony (a major crime). The cost of this is 100.

The dilemma is that the best course of action is for both of you to stay quiet, but since there is no honor among thieves, you do not trust one another. Then you both will turn state's evidence in order to avoid being convicted of the major crime (which happens if your partner turns state's evidence and you do not).

Consider this twist. Before you are hauled away, you and your partner swear to keep quiet. You believe that there is a 60% chance that he will keep his word. Draw a decision tree that represents your decision (to keep quiet or to cooperate) and the possible outcomes. What decision has the highest expected value? If x represents the probability that your partner will keep quiet, for what value of x is the value of keeping quiet equal to the value of cooperating?

4   Searching

a) I am an explorer looking for a new route to India. Assuming the Earth is flat (I'm not a very bright explorer), and I know the straight-line distances between any two cities on Earth, what would be a good search strategy for finding the shortest route?

b) Just before I am about to leave, I run into Magellan and he informs me that the Earth is spherical and that my straight-line distances are under-estimates. Should I keep the same search strategy? Why or why not?

5   Overfitting

We have discussed many types of machine learning including regression (fitting a function to data points), generative models (Bayes' nets, for example), and classification (perceptrons, multi-layer perceptrons, support vector machines). Pick two types of machine learning and describe how overfitting can occur and how we can modify their error functions to avoid it.

6   Planning

If GraphPlan terminates with a successful, 3-action plan in the first iteration, what constraints are there on the order in which the actions must be executed?

7   GraphPlan Vs. Partial Order Planner

Consider the following domain:

| op: fizz	op: fuzz
pre:	pre:
post: C, A	post: D, $\neg A$

The goal is $C \wedge D$. What solution is returned by POP? What solution is returned by GraphPlan? (If there are multiple answers, give them.) What does this example reveal about the expressive power of the solution descriptions in the two algorithms?

8 More GraphPlan

It has been suggested that the first phase of GraphPlan be used as a heuristic function for forward search in the following way: Given a state s and goal g, run the graph-construction phases of GraphPlan until all the components are present and not mutex in the last layer. Let n be the number of action layers in the graph. We will let n be the heuristic value for s. Is this an admissible heuristic?

9 Bayesian Networks

Consider this network:
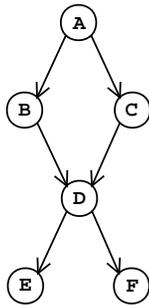
Which of the following conditional independence assumptions are true?

1. A and E are independent.

2. A and E are independent given D.

3. B and C are independent.

4. B and C are independent given A.

5. B and C are independent given D.

6. A and E are independent given B.

7. A and E are independent given F.

8. B and C are independent given E.

Is it possible to represent the joint distribution described by the network in these networks? If so, which one(s)?
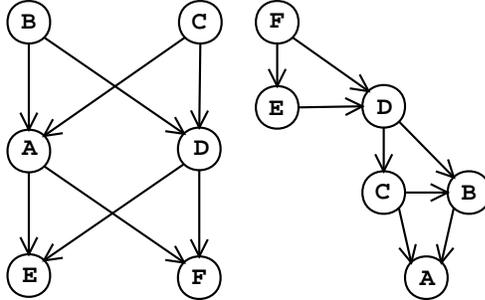
10 Automated Inference

Consider computing $P(A|L)$ using the recursive belief propagation method of section 15.3. What recursive calls to the algorithm would be made?
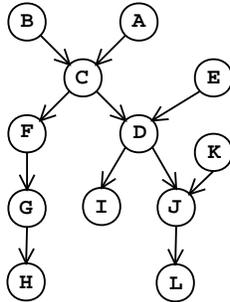
11 Sampling

When is sampling an appropriate model for probabilistic reasoning?

12 Medical Decisions

Your ankle hurts. The doctor thinks there might be a bone chip in there ($P = 0.3$). If there is a bone chip and it's not removed, you'll have major pain for a long time (U = -100). If you have surgery, whether or not there was a chip to be removed, you'll have moderate pain for a short time (U = -10).

Should you have the surgery? (Draw the decision tree.)

What would be the expected value of perfect information about whether you have a bone chip?

The doctor has been practicing a new laparascopic technique that will give much better information about the presence or absence of bone chips. It will

be a bit annoying and uncomfortable (U = -2). It will come out positive or negative with $P(\text{positive}|\text{bone chip}) = x$ and $P(\text{positive}|\neg\text{bone chip}) = y$.

For what values of x and y would it be worthwhile having the laparascopic test done? (Draw and evaluate the appropriate decision tree.)

13 Markov Decision Processes

Consider a 2-state MDP with $R(s_1) = 1$ and $R(s_2) = 0$.

The transition probabilities for the two actions (A_1 and A_2) are:

A_1	s_1	s_2
s_1	0.1	0.9
s_2	0.6	0.4

A_2	s_1	s_2
s_1	0.5	0.5
s_2	0.9	0.1

Use a decision tree to compute the optimal action strategy starting in state s_1, assuming you get to live for 2 steps.

What is the optimal 2-step strategy assuming you have no idea what your starting state is? ($P(s_1) = P(s_2) = 0.5$)

What is the optimal infinite-horizon value function and policy assuming a discount factor of 0.9?

14 Some Other Questions

1. What's the danger of estimating $P(A|B) = \frac{\#\text{ of occurences of A \& B together}}{\#\text{ of occurences of B}}$?

2. It's typical to include arc reversal as a possible step in Bayesian network structure search. It seems redundant because you could always delete the arc and then add it in the other direction. Why should we include it anyway?

3. Give an example of a Bayesian network with 4 nodes that is simpler than a 3-node network that encodes the same probability distribution on three nodes.

4. Suppose that a training set contains only a single example, repeated 100 times. In 80 of the 100 cases, the single output value is 1; in the other 20, it is 0. What will a single, sigmoid-unit perceptron predict for this example, assuming that it has been trained and reaches a global optimum? (Adapted from AIMA 19.6.)

5. AIMA 14.12

6. Nilsson 3.6, 3.7

7. Nilsson 16.8, 16.9, 16.10, 22.9