AI
SOLUTIONS

Computer Science Department
Stanford University
Comprehensive Examination in Artificial Intelligence
Autumn 2000

October 31, 2000

Solution Samples

PLEASE READ THIS FIRST

a. You should write your answers for this part of the Comprehensive Examination in a BLUE BOOK. Be sure to write your MAGIC NUMBER on the cover of every blue book that you use.

b. Be sure you have all the pages of this exam. There are 9 pages in addition to this cover sheet.

c. This exam is OPEN BOOK. You may use notes, articles, or books – but no help from people or computers.

d. Show your work, since PARTIAL CREDIT will be given for incomplete answers. For example, you can get credit for making a reasonable start on a problem even if the idea or arithmetic does not work out. You can also get credit for realizing that certain approaches are incorrect.

e. Points in this exam add up to 60. Points are allocated according to the number of minutes we believe a student familiar with the material should take to answer the questions. If you are somewhat less familiar with the material, a question may easily take you longer than the number of points it's worth. Therefore be careful: **IF YOU ARE TAKING TOO LONG ON A QUESTION, WRITE DOWN WHATEVER YOU HAVE AND MOVE ON.**

1 Search (8 points)

a. Uninformed Search

(i) (1 point) Describe or give an example of a search space where depth-first search will perform much better than iterative deepening search

(ii) (1 point) Describe or give an example of a search space where breadth-first search will perform much better than depth-first search

(iii) (1 point) Describe or give an example of a search space where depth-first search will perform much better than breadth-first search

b. Heuristic Search

(5 points) A* search involves evaluating search paths via $\hat{f} = g + \hat{h}$, where g is the lowest cost path to the current search state, and $\hat{h}$ is the heuristic function for the cost to a goal state. Now assume that the heuristic function $\hat{h}$ is induced from a function between nodes $h'(x, y)$ which provides an optimistic estimate of the cost and which obeys the triangle inequality. (The triangle inequality says that the sum of the costs from x to z and z to y must not be less than the cost from x to y directly.) Prove that the $\hat{f}$-cost along any path in the search tree is nondecreasing.

Answers

a. (i) The search space has only 1 choice at each point: it is a long linear chain, but it is very deep. Depth first will do $O(n)$ work while iterative deepening will do $O(n^2)$.

(ii) The search space has a high branching factor, but there are goal states only 2 moves away from the start state, but for some other moves (including the first one tried) the search state is extremely deep or infinite.

(iii) All paths in the search space are finite, there are no goal states near the start state, but there is a goal state at the end of all branches. (Or: referring to a high branching factor causing BFS to exceed available memory.)

b. The triangle inequality applied to a heuristic $\hat{h}$ says that $\hat{h}(n) \leq h'(n, n') + \hat{h}(n')$ for any nodes n, n' (since the triangle inequality is true of any goal node g). Nondecreasing $\hat{f}$-cost along a path means that the f-cost of a successor node is always at least as large as that of the node itself.

Want: $\hat{f}(n) \leq \hat{f}(n')$ if $n' \in S(n)$, the successors of node n.

I.e., want (rewriting this in terms of g and $\hat{h}$): $g(n) + \hat{h}(n) \leq g(n') + \hat{h}(n')$ if $n' \in S(n)$.

Our aim is to show that this is implied by the triangle inequality. To do this, we simply add $g(n)$ to both sides of the triangle inequality:

$g(n) + \hat{h}(n) \leq g(n) + h'(n, n') + \hat{h}(n')$

But if n' is a successor of n, then $g(n) + h'(n, n') \leq g(n')$ (as h' is optimistic). Hence: $g(n) + \hat{h}(n) \leq g(n') + \hat{h}(n')$, as required.

2 Logic: resolution (12 points)

(This question comes from (the late) Jon Barwise and (our new provost) John Etchemendy. It also appears in the exercises of Russell and Norvig.)

Consider the following statements:

> If a unicorn is mythical, then it is immortal, but if it is not mythical, then it is a mortal mammal. If the unicorn is either immortal or a mammal, then it is horned. The unicorn is magical, if it is horned.

From the above, can you prove that the unicorn is mythical? How about magical? Horned? Use resolution for your proofs (using propositional logic is possible and acceptable). Show:

a. (2 points) the basic logical translation of this text

b. (2 points) the translations of these into a form suitable for resolution theorem proving

c. (8 points) For each of the three unicorn properties (mythical, magical, horned), if it can be proved, show your proof. If it can't be proved, explain the justification for being able to conclude that the fact that unicorns have this property doesn't follow from the information given.

Answers

a. All statements about unicorns, so we don't mention them in proposition names.

   (i) Mythical $\rightarrow$ Immortal (i.e., Mythical $\rightarrow \neg$ Mortal)

   (ii) $\neg$ Mythical $\rightarrow$ Mortal $\wedge$ Mammal (which is equivalent to $\neg$ Mythical $\rightarrow$ Mortal, $\neg$ Mythical $\rightarrow$ Mammal)

   (iii) (Immortal $\vee$ Mammal) $\rightarrow$ Horned (which is equivalent to (Mortal $\vee$ Horned) $\wedge$ ($\neg$ Mammal v Horned))

   (iv) Horned $\rightarrow$ Magical

b. The following clauses are the result (with obvious abbreviations):

   (i) $\neg$My $\vee$ $\neg$Mo

   (ii) My $\vee$ Mo

   (iii) My $\vee$ Mam

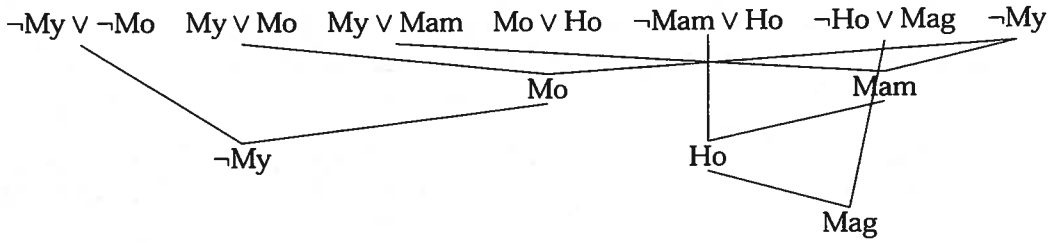
   (iv) Mo $\vee$ Ho

(v) ¬Mam ∨ Ho

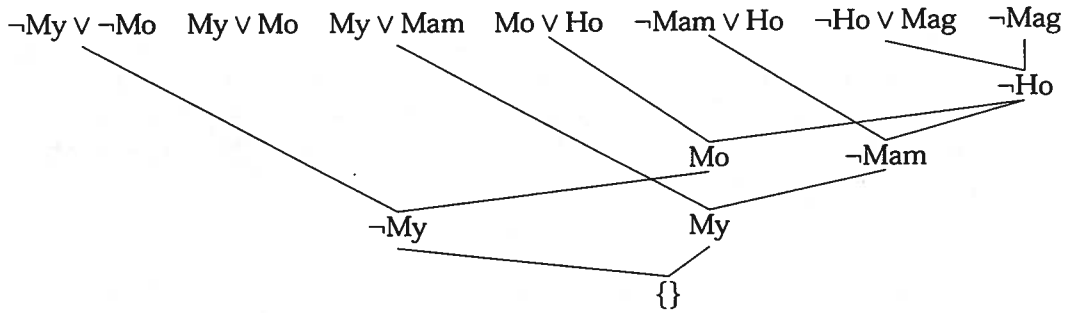
(vi) ¬Ho ∨ Mag

c. We use refutation proofs which will derive a contradiction if a set of sentences is unsatisfiable.

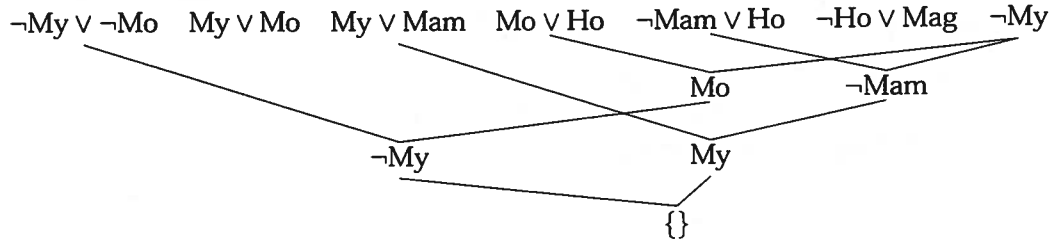
(i) Is it mythical? We add ¬My. No empty clause (i.e., no contradiction) can be derived. Refutation failure is a complete proof procedure for the propositional calculus, i.e., $KB \wedge \neg P \rightarrow False) \leftrightarrow (KB \rightarrow P)$. Hence, the given clauses do not entail that the unicorn is mythical.



(ii) Is it magical? We add ¬Mag. The empty clause (i.e., a contradiction) can be derived. Hence, the given clauses do entail that the unicorn is magical.



(iii) Is it horned? We add ¬Ho. The empty clause (i.e., a contradiction) can be derived. Hence, the given clauses do entail that the unicorn is horned.



3 Probabilistic models (12 points)

a. (1 point) A scientist tells us that 1/3 of all kangaroos have blue eyes, and 1/3 of all kangaroos are left-handed. On the basis of this information alone, what bounds can we place on the proportion of kangaroos that are both blue-eyed and left-handed?

b. (2 points) A loglinear model for the joint distribution of some number of categorical variables $X_1, \ldots, X_n$ takes the following general form:

$$\log P(X_1 = x_1, \ldots, X_n = x_n) = \sum_{C \in \mathcal{C}} \lambda_C(x_C)$$

where each $C \subseteq \{1, \ldots, n\}$ (so $\mathcal{C} \subseteq 2^{\{1,\ldots,n\}}$). That is, x_C is some subset of the x_i variables, and $\lambda_C(\cdot)$ is a function of this subset.

Show that all Bayesian networks ("graphical models") are loglinear models.

c. (9 points) Consider this situation: You saw John talking to his boss. Later you saw John looking upset. John may have been upset because his boss gave him a warning. Or he may be upset because his boyfriend left him.

(i) (1 point) Draw a suitable Bayesian network to represent the causal structure among the 4 random variables B for boss talking to John, W for boss warning John, U for John being upset, and L for his boyfriend leaving him.

(ii) (1 point) Make up conditional probability tables for each node. [HINT: keep the numbers simple. Use quarters, thirds, and halves!!]

(iii) (4 points) Given the information above (B and U), calculate the probability that John's boss warned him.

(iv) (3 points) Calculate the probability that John's boss warned him given the above information and that you know that his boyfriend left him.

Your answers may be approximate, but you should show your work

Answers

a. $0 \leq p \leq 1/3$

b. Suppose the Bayesian network has k nodes. In a Bayes net, due to the Markov assumption, the joint probability can be expressed as follows:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \prod_{i=1}^{k} P(X_i | Pa(X_i))$$

where $Pa(X_i)$ are the parent nodes of X_i in the directed graph. It was accepted to say that the λ_C were these conditional probability functions, and the result is thus in the form required. But this was a little imprecise, since x_C was specified as a set, whereas doing things this way requires that one member of the set be distinguished.

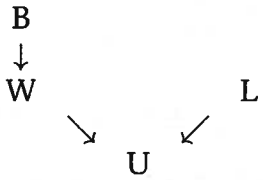
So, for $i = 1, \ldots, k$, let $C_i = \{X_i\} \cup Pa(X_i)$ and let $C_{k+i} = Pa(X_i)$. For $i = 1, \ldots, k$, let λ_{C_i} be the joint probability of the variables given as arguments, and for $i = k+1, \ldots, 2k$, let λ_{C_i} be the inverse of the joint probability of the variables given as arguments.

Then:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \prod_{i=1}^{2k} \lambda_{C_i}(X_{C_i})$$

Taking logs of both sides we now have a loglinear model in the form required.

c. (i)



(ii) A possible answer

	B	~B
	0.1	0.9

	L	~L
	0.2	0.8

	W	~W
B	0.3	0.7
~B	0.1	0.9

	U	~U
W, L	0.9	0.1
W, ~L	0.8	0.2
~W, L	0.7	0.3
~W, ~L	0.1	0.9

(iii) We are interested in $P(W|B,U) = P(W,B,U)/P(B,U)$. From first principles:

$$P(B,L,W,U) = P(B)P(L)P(W|B)P(U|W,L) = 0.1 \times 0.2 \times 0.3 \times 0.9 = 0.0054$$
$$P(B,\neg L,W,U) = 0.1 \times 0.8 \times 0.3 \times 0.8 = 0.0192$$
$$P(B,L,\neg W,U) = 0.1 \times 0.2 \times 0.7 \times 0.7 = 0.0098$$
$$P(B,\neg L,\neg W,U) = 0.1 \times 0.8 \times 0.7 \times 0.1 = 0.0056$$

So,

$$P(W|B,U) = \frac{0.0054 + 0.0192}{0.0054 + 0.0192 + 0.0098 + 0.0056} = \frac{0.0246}{0.04} = 0.615$$

(iv) And

$$P(W|B,U,L) = \frac{0.0054}{0.0054 + 0.0098} = \frac{0.0054}{0.0152} \approx 0.355$$

The lower probability estimate here shows the phenomenon of "explaining away".

4 Learning (14 points)

a. (2 points) A circuit has two input values A and B whose values are either +1 for 'true' or -1 for 'false'. Design a perceptron (Linear Threshold Unit) network which computes the function not (A or B). Draw the network and indicate clearly all weights and threshold values (assume the network outputs 1 if the dotproduct of the weights and the inputs is greater than some threshold t specified by each unit).

b. (3 points) We are trying to predict whether Comps questions are easy or difficult ($D = +$ if they are difficult) based on two features:

L Whether they are long (1 = long)

M Whether they have a lot of math (1 = yes)

For training data, we have examined 12 Comps questions, and have collected the following statistics, which we show twice: on the left are counts for the diferent data patterns, and on the right the data is shown in a contingency table showing $-:+$ counts for each combination of the classificatory variables.

L	M	D	count
0	0	−	4
0	0	+	1
0	1	−	0
0	1	+	3
1	0	−	1
1	0	+	2
1	1	−	1
1	1	+	0

		M 0	M 1
L	0	4:1	0:3
	1	1:2	1:0

Draw a decision tree for this data (using information gain for node splitting, and no stopping criterion or pruning, so that the tree is grown so long as there is some classificatory feature that appears to have information about the target feature).

c. (4 points) A Naive Bayes classifier for this problem predicts the target feature from the prior and independently from the classificatory features as follows:

$$\text{Choose } \hat{d} = \arg\max_{d=\{-,+\}} P(d)P(L = l|d)P(M = m|d)$$

(That is, it calculates the expression shown for both $d = +$ and $d = -$ and chooses the value of d that gives the expression higher probability.) What classificatory decisions would a Naive Bayes classifier make for each combination of classificatory variables?

d. (3 points) Suppose we have 3 test data instances, whose correct classification we know, as follows:

L	M	D
0	0	−
1	1	+
0	1	+

What is the decision of each classifier on each datum? Which does better overall?

e. (2 points) Is the better performance of one learner reasonable or surprising here?

7 8

Adj → beautiful
N → city

(iii) *Sydney* sydney
is $\lambda P \lambda x.P(x)$
a
beautiful $\lambda x.$beautiful(x)
city $\lambda x.$city(x)

(iv) S(rel(subj)) → NP(subj) VP(rel)
NP(val) → PN(val) |
NP($\lambda x.$(adj$(x) \wedge$ n(x))) → (Det) Adj(adj) N(n)
VP(rel(obj)) → V(rel) NP(obj)

(v) This grammar just ignores the article *a*, but it would need a (probably different) semantic translation when *a* is used in a subject or object NP. The complement NP is here a property, and not quantificational as in most uses of NPs in natural languages. (Also, *is* is just an identity function, but is treated as a higher order function to make parallel to other verbs. Not all adjectives are intersective.)