Final Examination for 6.034 (Spring 2008)

Name:

The following table needs to be changed to become consistent with the exam.

Q1 30%	Q2 20%	Q3 30%	Q4 20%

Good luck!

Question #1

Consider a Bayesian approach to a real medical diagnosis problem. In an urban health clinic in Lima, Peru, many people have tuberculosis (TB). Thus, if you are running a health clinic there, you would be very interested in determining which of your patients has TB. However, it's too costly to test everyone who comes in, and many who show up for treatment of unrelated problems, say a broken bone, are not very likely candidates for TB. However, if doctors notice that you are coughing, bringing up phlegm on each cough, and you have had this cough for more than two weeks, then there is a pretty high chance that in fact you may be suffering from TB.

These investigators were interested in what demographic factors predict the risk of a patient's having TB, and compiled the following table:

	minibus		**commute**		**family TB**		**TB history**		**not home**		**crowded**	
	yes	no	≥ 1 hr	$<$ 1hr	yes	no	yes	no	yes	no	yes	no
TB	14	2	8	8	7	10	4	13	16	1	6	11
no TB	51	36	20	67	40	85	24	101	87	38	50	75

Note: not all people answered all questions. The features are: (1) whether the person commutes to work by **minibus**, (2) their **commute** time, (3) previous contact with **TB** cases in the **family**, (4) a **history** of **TB**, (5) occupation away from **home**, and (6) over**crowded** conditions.

This table actually lists six *contingency tables* showing the number of times that, say for the first table, having TB is associated with riding a minibus to work or not, compared to similar numbers for those not having TB. This is a more compact representation than the *feature tables* you saw in lecture, which give the values for each feature and the outcome for each (in this case) patient. That table would have 142 rows and one column for each of the above features plus one for TB; the entries in each cell would only be 0 or 1. Nevertheless, from these tables, we can easily compute the R's we described in Prof. Barzilay's lecture on naïve Bayesian classifiers. For example, $R_{\text{minibus}}(1,1) = 14/16$, $R_{\text{minibus}}(0,1) = 2/16$, $R_{\text{minibus}}(1,0) = 51/87$, and $R_{\text{minibus}}(0,0) = 36/87$. This gives us the basis for using the prediction algorithm to suggest for each patient, depending on his or her answers to these questions, whether they have TB or not.

(a) In using Naive Bayesian classification, which of the six features above would give the greatest contribution in the prediction algorithm for a patient with TB?

(b) For a new patient for whom the prediction algorithm based on the first five features computes $S(1) = 0.004$ and $S(0) = 0.01$, what would be the resulting $S(0)$ and $S(1)$ after updating for the fact that the patient actually does *not* live in an overcrowded condition?

(c) Is your result in (b) consistent with the intuition that overcrowding is a risk factor for TB? Why?

(d) Suggest a set of numerical entries for the **crowding** table that would be consistent with the intuition suggested above, and strongly enough that it would "tip the balance" in favor of a prediction of TB in the case described in (b).

(e) The prediction algorithm we have described does not explicitly take the *prior probability* of TB into account. How could we modify the prediction algorithm to make sure that in the absence of any other evidence, $S(1)/(S(0) + S(1)) = P_0(\text{TB})$, the prior probability of TB, which in our case is 12%?

Question #2

Consider the following joint probability distribution on three binary variables, A, B and C:

A	B	C	prob
T	T	T	.112
T	T	F	.448
T	F	T	.028
T	F	F	.012
F	T	T	.048
F	T	F	.192
F	F	T	.112
F	F	F	.048
			1.000

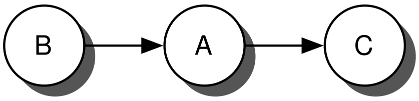
(a) What are the marginal probabilities of each of the individual variables, A, B and C?

(b) Are any of the pairs of variables $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$ (unconditionally) independent? *Show your calculation or give an explicit argument, not just a YES/NO answer.*

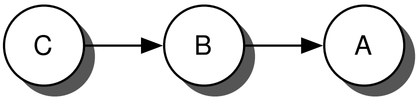
(c) Remember the definition of conditional probability, $P(x|y) = P(x, y)/P(y)$, and thus $P(x, y|z) = P(x, y, z)/P(z)$. Can any of the pairs of variables from part (b) be made independent conditional on the third variable (the one not in the pair)? *Again, show your calculations.*

(d) Of the ten possible Bayes network structures below (next page, each identified by a number), which ones are consistent with the observations you have made above? Please list your answers here:

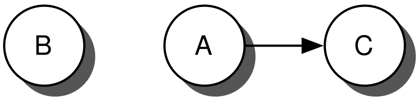
1:



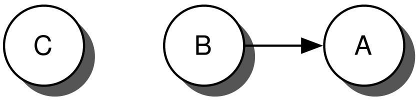
2:



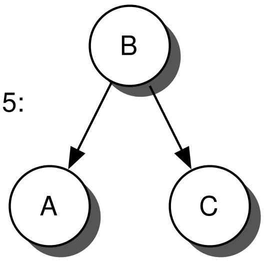
3:



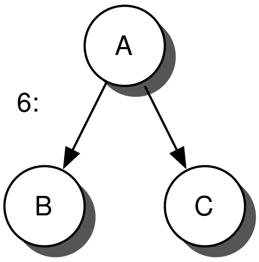
4:



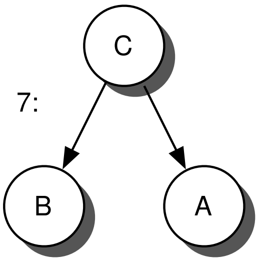
5:



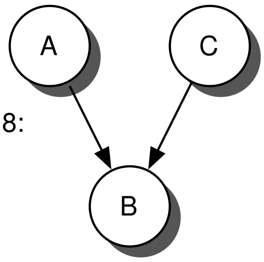
6:



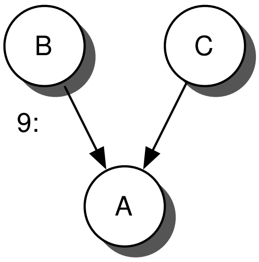
7:



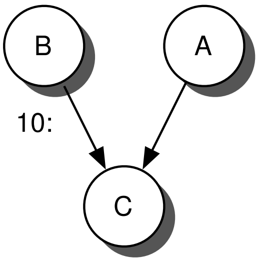
8:



9:



10:



Question #3

Party A is trying to determine if they should move their convention after party B. This would allow them to select their vice presidential candidate after party B, potentially giving them an advantage and allowing them to select the best candidate to offset the other party's selection. Party B has narrowed their selection to two possibilities, one from Connecticut, and the other from Massachusetts. Party A has narrowed their selection to three possibilities, one from New Mexico, one from Indiana, and one from Virginia.

If party B selects the candidate from Connecticut, the utility of each choice for party A is as follows:

New Mexico: 90	Indiana: 100	Virginia: 20

If party B selects the candidate from Massachusetts, the utility of each choice for party A is as follows:

New Mexico: 70	Indiana: 30	Virginia: 80

Party A has determined that there is a 60% chance that party B will select the candidate from Connecticut.

1. What is the expected value of knowing which candidate party B is going to select? Be sure to include ALL of your work including any and all decision trees you construct to answer this problem.

2. If party A has their convention first, who will they select as their vice presidential candidate and why?

3. If party A has their convention second, who will they select as their vice presidential candidate and why?

Question #4

Assume that we have three variables: A, B, and C. The domain of each variable initially is $\{1,\ 2,\ 3,\ 4\}$. Let's look at this problem as a constraint satisfaction problem and we are going to run full constraint propagation on this set of variables with these initial domains. We have the following three binary constraints:

- Constraint between A and B : $A < B$
- Constraint between B and C : $B = C$
- Constraint between A and C : $A < C + 1$

What are the domains of A, B, and C after full constraint propagation of the above constraints?

Question #5

In the table below are the alpha values that we obtained by running the perceptron algorithm with an initial weight vector of all zeros and a learning rate of one. Using these values, write the resulting output from the dual-form perceptron algorithm.

Iteration	Number of mis-classification	Label	Data point
1	8	-1	[1 2 3]
2	3	-1	[1 4 1]
3	2	+1	[1 7 4]
4	3	+1	[1 5 4]
5	0	+1	[1 6 5]

Question #6

Your goal is to design a part-of-speech tagger. For instance, given a sentence "The cats sleep", the tagger should output the following sequence "The_determiner cats_noun sleep_verb". The tag of the word depends on the word itself, the previous word's part of speech tag. You are given a corpus of sentences annotated with correct part-of-speech tags. Propose a method for part-of-speech tagging and specify its complexity. (Note that just applying a supervised learner will not be sufficient due to the dependence on the previous part-of-speech tag.)

Question #7

The idea behind bidirectional search is to simultaneously search forward from the initial state and backward from the goal, and stop when the two searches meet in the middle. Searching backwards means generating predecessors successively starting from the goal node.

- Assume that for both forward and backward searches we use a breadth-first search. Assume that the process of testing for intersection of the two frontiers can be done in constant time. For a tree of height d with a branching factor of b, compute the following:
  - Time complexity
  - Space complexity
- Assume that one direction of a bidirectional search is performed using a breadth-first search. What would be a good choice for the other direction? Why?

- Consider a space where the start space is number 1 and the successor function for state n returns two states, numbered $2n$ and $2n + 1$. Would bidirectional search be appropriate for this problem? If so, describe in detail how it would work.

Question #8

Indicate whether the following statement is true. If your answer is positive provide justification. Otherwise, draw an example of a feature space which demonstrates that the statement is wrong.

- Any feature space that is separable by a decision tree of height one[1] is always separable by a perceptron.

- Any feature space that is separable by a decision tree of height k is separable by a k–nearest neighbors.

- Any feature space that is separable by a perceptron is separable by a decision tree.

[1] The height of the tree is measured by the number of edges from the root to the most distant leave. For instance, a tree of height one has two levels of nodes.

- Any feature space that is separable by decision tree of height one is separable by perceptron.

- Any feature space that separable by k–nearest neighbors is separable by $k + 1$–nearest neighbors.

- Any feature space that separable by decision tree of height k is separable by decision tree of height $k + 1$.