Problem Set #5
Northwestern University EECS 348, Spring 2012

Turn in via Blackboard – it's okay to upload a PDF scan of a hand-written copy. Due Tuesday, May 29th at 11:59PM.

White wire	Red wire	Blue wire	Green wire	**Explosion**
0	0	0	0	1
0	0	1	1	0
0	1	0	0	0
0	1	1	0	1
1	0	0	0	1
1	0	1	1	1
1	1	0	0	1

Table 1: Training data

White wire	Red wire	Blue wire	Green wire	**Explosion**
0	0	1	0	1
0	0	1	1	1
1	1	0	1	1

Table 2: Validation data

1) You're trying to develop a predictor for whether a particular type of explosive device can be disarmed by cutting specific wires within it. The device has four wires of varying colors, and in experiments on recently seized devices, you've observed the training data shown in **Table 1**. (1 indicates the wire has been cut, or an explosion).
    a. (4 points) Draw the decision tree that results when using **Table 1** for training. Always split on the feature (wire) that *minimizes* expected entropy after the split, defined as:

$$EAfter(Explosion \mid Feature) = \sum_{v} P(Feature = v) H(Explosion | Feature = v)$$

where

$H(Explosion | Feature = v) =$

$-\sum_{v'} \mathrm{P}(Explosion = v' | Feature = v) \lg P(Explosion = v' | Feature = v)$

(use the above expressions rather than the mutual information expression from the lecture nodes; the above is similar but removes a constant term and fixes a typo)

So for example:

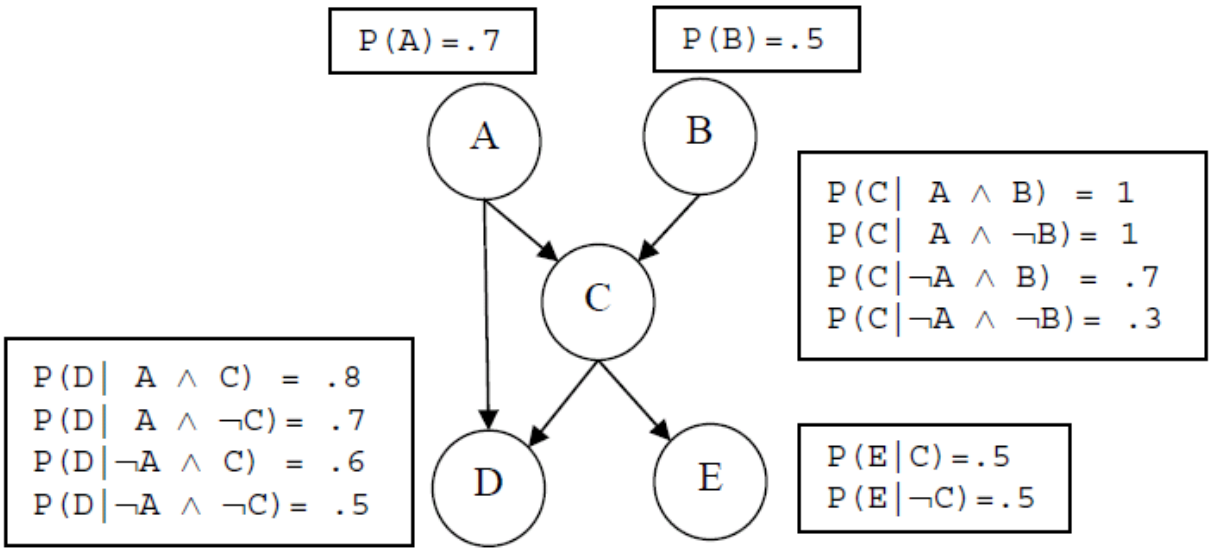
EAfter(explosion | red wire) = - (3/7) ((2/3) lg (2/3) + (1/3) lg (1/3))
- (4/7) ((3/4) lg (3/4) + (1/4) lg (1/4)) = 0.857,

versus

EAfter(explosion | green wire) = - (2/7) ((1/2) lg (1/2) + (1/2) lg (1/2))
- (5/7) ((4/5) lg (4/5) + (1/5) lg (1/5)) = 0.801

So splitting on green wire first is preferable to red wire (though white or blue may be better still). Feel free to use the above expressions (e.g. cutting and pasting into a spreadsheet or search engine) to perform your computations quickly. You should only need to compute about 7 more of these, total.

b. (1 point) Assume you generate the small set of validation data in Table 2. What is your decision tree's accuracy on the validation data?



2) Consider the above Bayesian Network. Here each letter represents a Boolean random variable. The probability distribution for each variable given its parents is shown. In the tables, P(A) indicates the probability that A is 1 (true);to get the probability that A is false, use 1 -P(A).
   a. (1 point) Assume C is known to be true. Will knowing the value of B affect the estimated probability that D is true?
   b. (1 point) What is P(A, ¬B, ¬C, ¬D, E)? Show your work.
   c. (1 points) How many different numbers would you need to specify the full joint distribution of all five variables, if there weren't any conditional independences? How does this compare to the number of parameters above?
3) Pick some problem you're interested in modeling that can be described in about five random variables. So for example, you might choose to predict who will win the next Cubs game, based on the ERA of each starting pitcher, whether the game is home or away, and what the winning percentage of the opponent is.
   a. (1 point) Draw a Bayes Net capturing your domain -- you **don't** need to specify any probability numbers, just the graph.
   b. (1 point) Explain one or two conditional independencies exhibited in your graph. If there aren't any, describe the conditional independence that would be most plausible (or "least implausible") in your domain.