Assignment 4: EM and Decision Making Under Uncertainty

CS486/686 – Fall 2015

Out: November 18th, 2015
Due: December 4th, 2015 12pm noon

Submit your assignment via LEARN (CS486 site) in the Assignment 4 Dropbox folder.
No late assignments will be accepted

1. **[60 pts]** Expectation Maximization

   Olertawo is a university town in New Zealand in which there is a strange medical condition called *Dunetts Syndrome*. *Dunetts Syndrome* comes in two forms: mild and severe (as well as not being there at all). It is known that about half the population of Olertawo have *Dunetts*, and that about half of those with *Dunetts* have it in severe form. *Dunetts Syndrome* has three observable symptoms, *Sloepnea*, *Forienndit*is, and *Degar spots*, all of which may be present if the patient has *Dunetts*, but with varying frequencies. In particular, it is well known that *Forienndit*is is present much more often when the condition is in its mild form (it is not as common in severe cases), whereas *Degar spots* are present much more often when the condition is in its severe form (and not as common in the mild cases). *Sloepnea* is present in either form of *Dunetts Syndrome*. However, about 10% of the population have a gene (called *TRIMONO-HT/S*) that makes it so they hardly ever show symptom *Sloepnea* (whether they have *Dunetts Syndrome* or not), but does not affect the other two symptoms. Symptoms *Sloepnea*, *Forienndit*is, and *Degar spots* are sometimes present (but much less often) even if the person does not have *Dunetts Syndrome*.

   (a) construct a Bayesian network (BN) for the domain explained above. Assign priors to each CPT based on the prior information given above. You don't need to be precise - just make rough guesses as to what the CPTs may be based on the description above.

   (b) you are given a dataset from 2000 patients, giving the existence of the three symptoms *Sloepnea*, *Forienndit*is and *Degar spots*, and whether they have gene *TRIMONO-HT/S* or not. About 5% of the data also has a record of wether the patient actually had *Dunetts Syndrome* or not. Using the Expectation Maximization (EM) algorithm, learn the CPTs for your BN. Run EM until the likelihood of the complete data (the sum of all your "weights" over *Dunetts Syndrome*) only changes by 0.01 or less. Start EM from your prior model, but add a small amount δ of random white noise to each parameter. Do this by adding a different randomly generated number $r \in [0, \delta]$ to each probability in each CPT, and then renormalising. Repeat the EM learning for a range of settings of δ from ~ 0 to $4.$, and do 20 trials for each setting with different randomizations at the start of each trial (but not at the start of each EM iteration). Use at least 20 values of δ evenly spread in $[0, 4)$. The idea is to evaluate the sensitivity of EM to the initial guess. As δ gets bigger, your initial CPTs will become more and more random, and you should get increasing numbers of trials with low accuracy. If your initial guess is not very good, you may even find that adding a small amount of noise helps.

   (c) To validate your learned model, you are given 100 test instances in which it is known whether the patient had *Dunetts Syndrome*. Make a prediction of whether *Dunetts Syndrome* is present for each example in the test set based on your learned model using EM and compare with the actual values of *Dunetts Syndrome*.

   The datasets are available on the course webpage. `trainData.txt` is a file of 2000 training examples with five columns. The first three give the presence/absence of each symptom (in order *Sloepnea*,*Forienndit*is,*Degar*

spots,*TRIMONO-HT/S* with 0 indicating the symptom is not present, and 1 indicating it is). The fourth column gives the presence/abscence of gene *TRIMONO-HT/S* (with 0 indicating the gene is not present, and 1 indicating it is). The fifth column is -1 if there is no record of *Dunetts Syndrome*, and $0, 1, 2$ if *Dunetts Syndrome* is recorded as being not present, mild or severe, respectively. `testData.txt` is the 100 examples for testing, has an additional column giving the severity of *Dunetts Syndrome*(the last column: 0=none, 1=mild, 2=severe).

What to hand in:

- A drawing of your Bayesian network showing all CPTs
- A printout of your code for doing EM on this model.
- A graph showing the prediction accuracy, for each value of δ, giving mean accuracy and standard deviation (error bars) over the 20 trials. Plot both the accuracy before and after running EM.

2. **[40 pts]** Decision Networks

In this question, you will construct a decision network for deciding whether to study for a course. Suppose that there are two types of courses: *hard* and *easy*, and that you can either *study* or *party*. If you *study*, then you pass an easy course with probability 0.9, and you pass a hard course with probability 0.6. If, on the other hand, you *party*, then you pass an easy course with probability 0.6 and a hard course with probability 0.35. All courses are either *pass* or *fail* (only two possible grades). You can assume you will know the difficulty of a course before making decisions, so you don't need a distribution over the difficulty.

(a) First, you need to elicit your own preferences (your own trade-off for studying versus passing). Assume the best possible situation is one in which you get to *party*, and you *pass*, and the worst possible situation is one in which you *study*, but *fail*. Formulate a standard gamble to elicit the utility of all four possible outcomes $o \in \{study, party\} \times \{pass, fail\}$, and assign each a utility in the range $[0, 1]$. A *standard gamble* means to assign the best outcome a value of 1, the worst outcome a value of 0, and then evaluate the two remaining outcomes, o, by selecting a number p at which you are indifferent between getting o with certainty and getting the best outcome with probability p and the worst outcome with probability $1 - p$.

(b) Draw a decision network for this problem, clearly labeling the decision node, the chance nodes, and the utility node, and showing the conditional probability table (CPT) clearly.

(c) Given the decision network and your elicited utility function, compute the value of studying or partying for both hard and easy courses. Based on this value, what policy should you follow?

(d) Some courses have four assignments, a midterm, and a final exam. These six assessments follow each other sequentially as assignment 1, assignment 2, midterm, assignment 3, assignment 4, final. Assume that marks for each assessment are allocated immediately upon hand-in, so that the decision to study or party for the next assessment can be based on the grade of the previous ones. Draw the decision network for this sequential decision making problem.

What to hand in:

- A description of your standard gamble, and a table showing your utility function for the four possible outcomes $o \in \{study, party\} \times \{pass, fail\}$.
- A drawing of your decision network, and your CPTs.
- The value for studying and partying for both types of courses, and the policy you should follow for both types of courses.
- The decision network for the sequential version of the problem