Assignment 2: Bayesian Networks and Decision Networks

CS486/686 – Fall 2008

Out: Oct 2, 2008
Due: Oct 23, 2008

Be sure to include your name and student number with your assignment.

Every year, credit card companies lose millions of dollars due to frauds resulting from lost or stolen cards. Recently, the financial industry has turned to AI for solutions to the fraud detection problem. Intuitively, credit card holders tend to make purchases following a certain pattern. A fraud is likely to happen when this pattern is broken. In this assignment you will implement a simple fraud detection system in 3 steps. First, you will implement the variable elimination algorithm. Second, you will define your fraud detection system as a Bayesian network and compute the likelihood of a fraud in different situations. Third, you will extend your Bayesian network to a decision network which will be used to decide when a transaction should be blocked. Computations in your Bayesian network and your decision network should be done with your implementation of the variable elimination algorithm. However, if you are not able to complete your implementation of variable elimination, you can also do the computations by hand since they are relatively simple.

1. **[0 pts]** Implement the variable elimination algorithm by coding the following 4 functions in the programming language of your choice.

   (a) **restrictedFactor = restrict(factor, variable, value):** function that restricts a variable to some value in a given factor.

   (b) **productFactor = multiply(factor1, factor2):** function that multiplies two factors.

   (c) **resultFactor = sumout(factor, variable):** function that sums out a variable in a given factor.

   (d) **normalizedFactor = normalize(factor, variable):** function that normalizes a factor with respect to a variable. This is useful when the factor is a distribution with respect to the given variable (i.e. sum of the probabilities must be 1).

   (e) **resultFactor = inference(factorList, queryVariables, orderedListOfHiddenVariables, evidenceList):** function that computes $Pr(queryVariables|evidenceList)$ by variable elimination. This function should restrict the factors in factorList according to the evidence in evidenceList. Next, it should sumout the hidden variables from the product of the factors in factorList. The variables should be summed out in the order given in orderedListOfHiddenVariables. Finally, the answer should be normalized to make sure that the probability distribution sums up to 1.

   Tip: factors are essentially multi-dimensional arrays, hence you may want to use this data-structure, but feel free to use a different data-structure.

   Tip: test each function individually using simple examples from the lecture slides. If you wait till the end to test your entire program it will be much harder to debug.

   What to hand in: hand in a printout of your code. Note that there are no marks given for Question 1. However, in Questions 2 and 3, part of the marks will be given for using your variable elimination algorithm to do the computations.

2. **[61 pts]** Suppose you are working for a financial institution and you are asked to implement a fraud detection system. You plan to use the following information:

   - When the card holder is travelling abroad, fraudulent transactions are more likely since tourists are prime targets for thieves. More precisely, 1% of transactions are fraudulent when the card holder is travelling, where as only 0.4% of the transactions are fraudulent when she is not travelling. On average, 5% of all transactions happen while the card holder is travelling. If a transaction is fraudulent, then the likelihood of a foreign purchase increases, unless the card holder happens to be travelling. More precisely, when the card holder is not travelling, 10% of the fraudulent transactions are foreign purchases where as only 1% of the legitimate transactions are foreign purchases. On the other hand, when the card holder is travelling, then 90% of the transactions are foreign purchases regardless of the legitimacy of the transactions.
   - Purchases made over the internet are more likely to be fraudulent. This is especially true for card holders who don't own any computer. Currently, 60% of the population owns a computer and for those card holders, 1% of their legitimate transactions are done over the internet, however this percentage increases to 2% for fraudulent transactions. For those who don't own any computer, a mere 0.1% of their legitimate transactions is done over the internet, but that number increases to 1.1% for fraudulent transactions. Unfortunately, the credit card company doesn't know whether a card holder owns a computer, however it can usually guess by verifying whether any of the recent transactions involve the purchase of computer related accessories. In any given week, 10% of those who own a computer purchase with their credit card at least one computer related item as opposed to just 0.1% of those who don't own any computer.

   (a) **[25 pts]** Construct a Bayes Network that your fraud detection system can use to indentify fraudulent transactions.

   What to hand in: Show the graph defining the network and the Conditional Probability Tables associated with each node in the graph. This network should encode the information stated above. Your network should contain exactly six nodes, corresponding to the following binary random variables:

   - OC – card holder owns a computer.
   - $Fraud$ – current transaction is fraudulent.
   - $Trav$ – card holder is currently travelling.
   - FP – current transaction is a foreign purchase.
   - IP – current purchase is an internet purchase.
   - CRP – a computer related purchase was made in the past week.

   The arcs defining your Bayes Network should accurately capture the probabilistic dependencies between these variables.

   (b) **[12 pts]** What is the prior probability (i.e., before we search for previous computer related purchases and before we verify whether it is a foreign and/or an internet purchase) that the current transaction is a fraud? What is the probability that the current transaction is a fraud once we have verified that it is a foreign transaction, but not an internet purchase and that the card holder purchased computer related accessories in the past week?

   What to hand in: Indicate what queries (i.e., $Pr(variables|evidence)$) you used to compute those probabilities. Whether you answer the queries by hand or using the code you wrote for Question 1, provide a printout of the factors computed at each step of variable elimination (as done in the lecture slides). Use the following variable ordering when summing out variales in variable elimination: $Trav$, FP, $Fraud$, IP, OC, CRP. Note that a maximum of two thirds of the marks are earned if you answer correctly the question by doing the computations by hand instead of using your program.

   (c) **[12 pts]** After computing those probabilities, the fraud detection system raises a flag and recommends that the card holder be called to confirm the transaction. An agent calls at the domicile of the card holder but she is not home. Her spouse confirms that she is currently out of town on a business trip. How does the probability of a fraud changes based on this new piece of information?

   What to hand in: Same as for Question 2b.

(d) **[12 pts]** Suppose you are not a very honest employee and you just stole a credit card. You know that the fraud detection system uses the Bayes net designed earlier but you still want to make an important purchase over the internet. What can you do prior to your internet purchase to reduce the risk that the transaction will be rejected as a possible fraud?

What to hand in: Tell me the action taken and indicate by how much the probability of a fraud gets reduced. Follow the same instructions as for Question 2b.

3. **[39 pts]** Extend your Bayesian network to become a decision network which will be used to decide when a transaction should be blocked. Use the following information:

- For each legitimate transaction processed, the credit card company earns a profit of roughly 0.5% of the transaction's value through interest charges and merchant charges. For instance, on each transaction of $1000, the credit card company expects a profit of roughly $5. Assuming that the credit card company covers fraudulent transactions, it will suffer a loss equal to the value of each fraudulent transaction. However, if a fraudulent transaction is blocked, there are no loss. In the event where a legitimate transaction is blocked, customers get annoyed (and sometimes cancel their credit card), which is considered to be equivalent to an expected loss of $10.

(a) **[15 pts]** Extend your Bayesian network to a decision network. Assume transactions of $1000.

What to hand in: Show the graph of the decision network and the utility table (conditional probability tables are the same as for your Bayesian network in Question 2). This network should encode the information stated above. In addition to the 6 chance variables defined in Question 2, your network should include a decision node and a utility node:

- B — decision to block (or not) a transaction.
- U — utility.

The arcs into the decision node should encode informational dependencies and the arcs into the utility node should encode utility dependencies.

(b) **[12 pts]** Should a $1000 transaction be blocked when it is a foreign transaction that wasn't done over the internet and computer related accessories were purchased in the past week?

What to hand in: Indicate the decision made as well as the query (i.e., $E(decision|evidence)$) you used to evaluate the decisions. Whether you answer the queries by hand or using the code you wrote for Question 1, provide a printout of the factors computed at each step of variable elimination (as done in the lecture slides). Use the following variable ordering when summing out variables in variable elimination: $Trav$, FP, $Fraud$, IP, OC, CRP. Note that a maximum of two thirds of the marks are earned if you answer correctly the question by doing the computations by hand instead of using your program.

(c) **[12 pts]** What is the expected value of the information gained when calling a customer at home to verify whether she is travelling in the case of a foreign transaction that wasn't done over the internet and computer related accessories were purchased in the past week?

What to hand in: Same as for Question 3b.