CS 5100, Fall 2011

Assignment 5—Probabilistic Reasoning and Machine Learning

Assigned: November 17, 2011
Due: December 1, 2011 in class (hard copy)

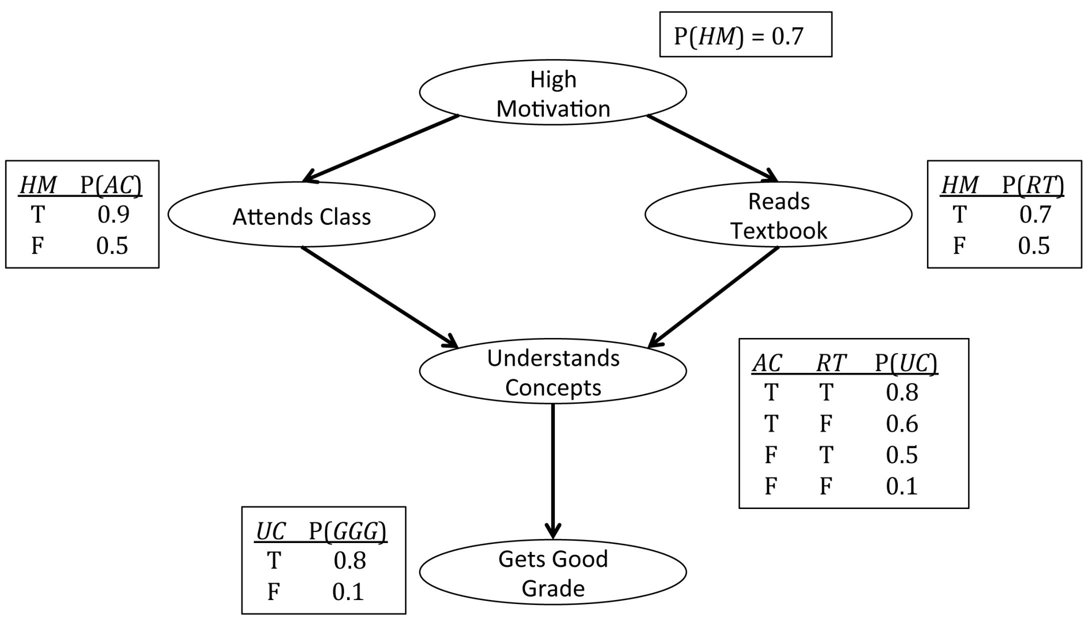
Part I: Bayesian Networks

a. The Bayesian network below represents the factors that influence whether a student will get a good grade. Based on this Bayesian network, calculate (showing your work):

   i. the probability that a student gets a good grade, GIVEN THAT the student has high motivation

   ii. the (unconditional) probability that a student gets a good grade

b. The Bayesian network specifies 11 probability values. How many probability values would have to be provided if we were specifying the entire joint probability distribution for this scenario, without the conditional independence assumptions embodied in the Bayesian network? **NOTE: $P(E)$ means probability that E is true.**



c. Given the belief model described below, provide a diagram of a Bayesian network that represents these beliefs, including the CPTs (conditional probability tables) that contain the probabilities. You can make up plausible numbers for the probabilities to fill in the CPTs. Your numbers should make sense given the problem (i.e., a greater probability value when common sense clearly requires it).

A cardiology doctor must decide whether to recommend heart surgery (HS) for a patient with heart disease. This will depend on three things: if the patient's condition is life threatening (LT), if the patient is in good health overall (GH), and if the patient is able to pay for the surgery (AP). Each of these conditions depends on some other factors:

1. LT depends on whether the patient has a total blockage (TB)
2. GH depends on whether the patient has other serious health problems (OP)
3. AP depends on whether the patient is employed (EM), and whether the patient has health insurance that will cover at least part of the costs (IN)

The unconditional probabilities of some factors are given below:

$P(TB) = 0.2$

$P(OP) = 0.6$

$P(EM) = 0.4$

$P(IN) = 0.9$

Part II: Naïve Bayes Classifiers and Decision Tree Learning

Suppose there are 100 graduate students in CCIS, 50 male and 50 female.

a. What is the "information value" (also called entropy or disorder) of the gender variable?

Now consider two different properties we can use to classify students, such that knowing a students value for these properties could make their gender more predictable:

Degree program—there are 25 Ph.D. students, including 20 males and 5 females; and there are 75 M.S. students, including 30 males and 45 females

Home region—there are 40 students from North America: 10 male and 30 female; 45 students from Asia: 25 male and 20 female; and 15 students from Europe (all male)

b. Would you rather know a student's degree program or their home region in order to make a prediction about gender? Show the calculation of information gain (IG) for each of the two properties to justify your answer.

c. Using the data above as training data, draw the naïve Bayes classifier that would be learned, including the required CPTs (conditional probability tables).