# CSE 515: Statistical Methods in Computer Science

## Homework #4

Due **by email** to lowd at cs by 11:59pm on Friday May 29, 2009

**Guidelines:** You can brainstorm with others, but please solve the problems and write up the answers by yourself. You may use textbooks (Koller & Friedman, Russell & Norvig, Wikipedia, etc.), and lecture notes. Please do NOT use any other resources or references (e.g., example code, online problem solutions, etc.) without asking.

**NOTE:** Justify every answer with explanation and/or calculations as appropriate.

**Submission instructions:** Submit a **PDF** of this assignment to Daniel Lowd.

1. Consider the following Bayesian network: $A \rightarrow B \rightarrow C$. And the following data table, with entries '?1' and '?2' missing at random:

| A | B | C |
|---|---|---|
| F | F | F |
| F | F | ?1 |
| F | T | F |
| T | T | T |
| T | ?2 | T |
| T | F | T |

   (a) Use the data to estimate initial parameters for this network, using maximum likelihood estimation for simplicity.

   (b) Apply the EM algorithm (by hand) to estimate the values of the missing data, reestimate the parameters, etc. until convergence. Show your calculations.

   (c) How many iterations does EM take to converge? Will this always be the case? Explain.

2. Show that adding edges to a Bayesian network never decreases the likelihood.

3. (Based on K&F 19.16.) This problem considers the performance of various types of structure search algorithms. Suppose we have a general network structure search algorithm, $A$, that takes a set of basic operators on network structures as a parameter. This set of operators defines the search space for $A$, as it defines the candidate network structures that are "immediate successors" of any current candidate network structure, i.e., the successor states of any state reached in the search. Thus, for example, if the set of operators is {add an edge not currently in the network}, then the successor states of any candidate network $G$ is the set of structures obtained by adding a single edge anywhere in $G$ (so long as acyclicity is maintained).

   Given a set of operators, $A$ does a simple greedy search over the set of network structures, starting from the empty network (no edges), using a penalized log likelihood scoring function (log likelihood of training data minus a constant penalty $\kappa$ for each edge in the network). Now, consider two sets of operators we can use in $A$. Let $A_{[add]}$ be $A$ using the set of operations {add an edge not currently in the network}, and let $A_{[add,delete]}$ be $A$ using the set of operations {add an edge not currently in the network, delete an edge currently in the network}.

(a) Show a distribution where, for any edge penalty $\kappa > 0$, the answer produced by $A_{[add]}$ has a worse score than the answer produced by $A_{[add,delete]}$. (It's easiest to represent your true distribution in the form of a Bayesian network; i.e., a network from which the sample data is generated.)

(b) Show a distribution where, for any edge penalty $\kappa > 0$, $A_{[add,delete]}$ will converge to a local maximum. In other words, the answer returned by the algorithm has a lower score than the optimal (highest-scoring) network. What can we conclude about the ability of our algorithm to find the optimal structure?

4. Naive Bayes (NB) and logistic regression (LR) have the same form, but naive Bayes is a generative model (learned using maximum likelihood, to maximize $P(x, y)$), while logistic regression is a discriminative model (learned using maximum conditional likelihood, to maximize $P(y|x)$). In this problem, assume both models are learned on the same training data with no prior. The conditional log likelihood (CLL) on a dataset $D$ is defined as:

$$\sum_{(x,y)\in D} \log P(y|x)$$

(a) Under what conditions will NB have a higher CLL than LR on the training data?

(b) Under what conditions will NB have a higher CLL than LR on separate testing data?