Ph.D. Open, 18-20 March 2010, Probabilistic Graphical Models

Take-home assignments

Assignment 1: Data analysis

File *retention.txt* contains the data that we looked at in the lecture (there are 170 records in the file. The first line contains the names of the eight variables (described below). Please, look at the data carefully using a statistical package of your choice. Write down any interesting observations. Are the data normal and are the interactions linear? As a minimum, you should

(a) generate descriptive statistics and plot histograms for the following three columns: *apret*, *tstsc*, and *salar*.

(b) perform linear regression of *apret* on *tstsc* and *salar* separately and then of *apret* on both *tstsc* and *salar*.

The meaning of the columns is as follows:

spend average spending per student (in dollars)
apret average retention rate (i.e., percentage of students making it through the studies)
top10 percentage of incoming freshmen who were among the top 10% students in their high schools
rejr school's rejection rate (percentage of applicants denied admission)
tstsc average test scores of incoming freshmen
pacc percent of admitted applicants who accept university's offer
strat student-teacher ratio
salar average faculty salary (in dollars)

Each row is for one of the 170 colleges for which the data was measured.

Can you draw any causal conclusions from the data?

Assignment 2: Causal discovery

Your task is to analyze the retention.txt data set from Assignment 1 using **GeNIe**'s (available from http://genie.sis.pitt.edu/) causal discovery algorithms and, by this, verify the findings in [Druzdzel & Glymour 1994, available at http://www.pitt.edu/~druzdzel/psfiles/kdd94.pdf] concerning student retention in US colleges. We performed the original study on 1992 *US News and World Report* data. The data in the *retention.txt* data file are for the year 1993.

There are some small differences in variables under study. Because freshmen retention and senior graduation rates differed so little (most students who drop out do so in their freshmen year), I eliminated the variable freshmen retention. What used to be *apgra* is in the 1993 data set called *apret* (this is the percentage of student retention over the four years). (We rerun the 1992 data set and verified that the conclusion regarding the impact of the quality of incoming students on the

retention rate is the same.) Otherwise all variables are the same (but the measurements are for the year 1993).

Do the 1993 data support Druzdzel & Glymour's conclusions? If not, why not. Can you find anything else going on in the data (i.e., is there any structure in the data)? What are the causal graphs suggested by **GeNIe**? What causes student retention? Please, feel free to help **GeNIe** using your common sense knowledge of interactions between the variables included in the data set (not necessarily the knowledge inputed by Druzdzel & Glymour). For example, you might know something about the time precedence. Perhaps there are pairs of variables for which we are reasonably sure that there is/there is not an arc between them? Check the sensitivity of your result to the significance level. Druzdzel & Glymour made the assumption that the data are normally distributed and linearly dependent (this is an implicit assumption used in the PC algorithm in **GeNIe**). Please, feel free to check whether the assumption is indeed valid in the data using the graphing capabilities in **GeNIe**.

You do not need to make this a lot of work, but please, run the PC algorithm at least a few times with different values of parameters and look critically at the results. Does discretizing the data and running the Greedy Thick Thinning algorithm lead to any more insights?

Here are brief instructions to save you time in paging through the on-line manuals (http://genie.sis.pitt.edu/wiki/):

(1) Open the data file through *File menu*/*Open Data File*.
(2) Choose *Learn New Network* from *Data* menu.
(3) Choose the PC algorithm and set the significance level (PC algorithm uses classical statistical independence tests and here is where you set the significance level for these tests).
(4) You can enter the background knowledge by clicking the Background Knowledge button. Here you can force and forbid causal connections and also order variables in temporal tiers. A variable that is in a later tier cannot cause variables in earlier tiers.
(5) The result of the algorithm is a pattern that you can edit before proceeding with parameter learning. To edit a double-edged or red arc in the pattern graph, right-click on it.

In case of any problems with or any questions about the software, please get in touch with my team through the **GeNIe** and **SMILE**© Forum (http://genie.sis.pitt.edu/forum/).

Assignment 3 (Extra Credit, only for those willing to do extra work)
States of a joint probability space constructed over a finite set of discrete random variables are combinations of outcomes of these individual variables. Can anything be said about the probability distribution over probabilities of individual elements of this space?