Session 8:
Hands-on Exercises: (Learning)

Marek J. Druzdzel

University of Pittsburgh

School of Information Sciences and Intelligent Systems Program

marek@sis.pitt.edu
http://www.pitt.edu/~druzdzel

Summer School on *"Modeling and Decision Making Using Bayesian Statistics"*
Aalto University, Department of Applied Mechanics, Marine Technology in Espoo, Finland
4th – 8th June, 2012

Course schedule

Day 1 Monday

Session 1: Introduction to Bayesian inference
Session 2: Bayesian networks
Session 3: Building Bayesian networks
Session 4: Hands-on exercises (Bayesian networks)

Day 2 Tuesday

Session 5: Learning Bayesian networks and causal discovery
Session 6: Decision theory and decision analysis
Session 7: Hands-on exercises (learning)
Session 8: Hands-on exercises (decision modeling)

Session overview

Two exercises:

1. **Causal discovery, continuous networks (PC algorithm).**
2. **Learning Bayesian networks for classification (discrete networks), varification of accuracy.**

- **This is group work (ideally pairs of participants).**
- **Individual work is possible, although discouraged (you typically learn more when working in a small group).**
- **I will be there to help you (raise your questions – they may be of interest to everybody).**

What I want you to know after this session

- Know how to construct Bayesian networks from data
- Know how to interpret the learned graph
- How to enter prior knowledge into the learning process
- Know how to verify the diagnostic/classification accuracy of a learned network

Exercise 1

Based on the data file retention.txt, answer the question which of the variables included in the file are likely causes of low student retention in US universities.

Variables in the retention.txt file:

spend: average spending per student (total amount in US$ spent per year per student)

apret: average percentage retention (percentage of students, who finish the first year and stay at the university)

top10: percentage of first year students who were in the top 10% of their high school graduating class (a measure of student quality)

rejr: rejection rate (percentage of applicants who have not been admitted to the university; a measure of university's selectivity: the higher this percentage, the more selective the university)

tstsc: standardized test scores (results of standardized tests for applicants, expressed on a scale 0-100, where 100 is perfect score; a measure of quality of incoming students)

pacc: acceptance percentage (percentage of applicants admitted to the university, who accept the university's admission offer and start their studies at this university (applicants to US universities typically apply to several universities and go to the most preferred university among those that admitted them)

strat: student-teacher ratio (ratio of the total number of students to the total number of faculty at the university (i.e., the number of students for each of the professor; a measure of the quality of teaching at the university)

salar: average faculty salary (average annual faculty salary in US$ (a measure of the quality of the faculty)

Exercise 2

The file house-votes.txt contains a set of data from the Machine Learning Repository at the University of California, Irvine (http://archive.ics.uci.edu/ml/), describing how various US congressmen voted on 16 different questions. It is possible to create a Bayesian network that will classify the congressmen (i.e., whether they are Republican or Democrats) based on their voting pattern.

Create a collection of Bayesian networks for the purpose of this classification and verify their classification accuracies.

Hints:

(1) Because algorithms for learning the structure are not able to learn from data sets with missing values, you can deal with missing data in the following way: (1) remove all records with missing data, (2) learn the structure of the network, and (3) load the full data set again and use this complete set (with missing data) to learn the parameters of the previously learned model.

(2) When verifying a network's accuracy, you should use cross-validation and uniformization of probabilities.