CIS 7414X Expert Systems: Class Projects

Description

This class project is for you to explore your understanding of Bayesian networks on real (or semi-real) world data sets. You will be required to choose a data set that are most interesting to you. With this data set, you will define a problem of your own. The problem tasks can be about predictions, diagnosis, monitoring, controlling or the any combination of them. When constructing such a problem, please be aware that you must be able to construct a Bayesian network to solve it, and you must be able to extract most of conditional probabilities from the data set. In some cases, you are allowed to make up additional conditional probabilities based on "your expertise". However, the final result of your system will be verified against the data set.

Requirements

You should do the project by yourself. For full credit, the problem should be composed in such a way that

- a Bayesian network can be constructed to solve the problem, and
- conditional probability tables can be extracted from the data

For bonus credit, the problem should be composed in such a way that

- the structure of the Bayesian network (or part of it) can be learned from the data, and/or
- the conditional probabilities can be learned from the data.

You are reqired to submit 3 reports

- A 1 page project proposal describing (**due on October 13rd**)
  - the data set you have chosen,
  - the problem you are going to solve,
  - a preliminary idea of which variables will be modeled in the Bayesian network,
  - a preliminary idea of the causal links among these variables,
  - how you will extract probabilities from the data, and
  - how you can verify the system output
- A 2 page milestone report (**November 10th**) describing
  - the Bayesian network you have constructed,
  - the basic results of your network, and
  - the difficulties you have experienced
- A final report of no more than 8 pages describing
  - the problem,
  - the data set,
  - your analysis of the problem,
  - the Bayesian network you are proposing to solve the problem, and
  - your results: e.g. classification accuracy
- A 5 – 10 minitue presentation (**December 1st**)

Data sets

Among the datasets in the repositories listed below, please select one of that interests you most for your project. If you would like to choose your own dataset, please contact the instructor for permission.

- UC Irvine machine learning repository: `http://archive.ics.uci.edu/ml/`

- The CMU ”Probabilistic Graphical Models” course website (by Carlos Guestrin): `http://www.cs.cmu.edu/~guestrin/Class/10708-F08/projects.html`

- Sam Roweis’s data sets: `http://www.cs.nyu.edu/~roweis/data.html`

Important dates

- **October 6th**: Class Discussion

- **October 13rd**: Project proposal due

- **November 10th**: Milestone report due

- **December 1st**: Final report and presentation due

Appendix: Classifiers and classification accuracy

Most projects can be formulated as a classifier:

$$Clsf : F_1 \times F_2 \times \ldots \times F_n \rightarrow C$$

where F_is are input/evidence variables and C is the output class.

For example, given the values for the pollution level, smoking or not, Xray test, and Dyspnoea, you might want to predict whether a patient has cancer or not. A Bayesin network might have an output regarding the probabilities on the values which C can take given the input values to P, S, X and D:

P	S	X	D	$P(C = T\|P, S, X, D)$	$P(C = F\|P, S, X, D)$
H	T	Pos	T	.8	.2
H	T	Pos	T	.6	.4
H	T	Pos	T	.7	.3
...					
L	F	Neg	F	.2	.8

With this probabilistic output, a simple classifier can be constructed by taking the value with the highest probability as the classifier’s output:

P	S	X	D	$Clsf$	$P(C=T\|P,S,X,D)$	$P(C=F\|P,S,X,D)$
H	T	Pos	T	T	.8	.2
H	T	Pos	T	F	.4	.6
H	T	Pos	T	T	.7	.3
. . .						
L	F	Neg	F	F	.2	.8

On the other hand, in your test data set, you might have:

P	S	X	D	C
H	T	Pos	T	T
H	T	Pos	T	T
H	T	Pos	T	T
. . .				
L	F	Neg	F	F

The classifier's accuracy rate on the test data is then

$$\frac{\#\{\langle P,S,X,D,C=Clsf(P,S,X,D)\rangle\}}{\#\{\langle P,S,X,D,C\rangle\}}$$

the number of the cases in which $Clsf(P,S,X,D)$ matches the test data divided by the total number of cases in the test data.