Learning with Bayesian Networks

Advanced Methods for Knowledge Representation – Assignment #3

Fall 2013

1 Submission and evaluation

1. Each student works individually, not a teamwork assignment.
2. The submission has three parts:
   (a) a *bnet* structure that in your opinion best represents and generalizes the input data,
   (b) a commented source code (m files),
   (c) a report that describes the solution and answers the assignment questions,
3. The way of submission:
   (a) the system *https://cmp.felk.cvut.cz/upload*,
   (b) deadline: 3.1.2014 midnight,
   (c) the archive file structure (the name is *student_username.zip*):
      i. a file *bnet.mat* with *bnet* network specified above,
      ii. a file *student_username.pdf* with the report,
      iii. a directory *matlab* with m files that underlie the solution.
4. Up to 15 points can be obtained for this assignment:
   (a) 4 points for predictive accuracy reached by *bnet* on unseen data,
   (b) 11 points for the report and the functional source code,
   (c) there is a 3 point penalty for each commenced week of delay.

2 Task

1. Get familiar with BN Toolbox for Matlab (*http://bnt.googlecode.com/svn/trunk/docs/usage.html*).
2. Download and study the input dataset lung_txt.csv (available at the course webpage)

(a) try to understand the relationships among variables in the domain of pneumonia diseases, *read_lung.m* gives a short description,

(b) read the data into Matlab: *samples=read_lung('lung_txt.csv')*,

(c) note frequent missing values (NaN), take them as missing completely at random.

3. Manually construct a BN structure that you find best for the given domain and interpret it

   (a) the interpretation is meant as a brief justification of the proposed network, mentioning possible doubts, alternatives, etc.

4. Think about dealing with the input data, in particular, discuss

   (a) the asset of splitting on train and test data to obtain a model that does not overfit the input data,

   (b) the ways of missing data treatment, (a) omit samples with missing values, b) use them to test only, c) possibilities of learning with missing values in BNT, d) estimation of missing values by kNN or incrementally by a BN learnt with fully known samples).

5. Learn quantitative parameters (CPTs) of the network ad 3 from data (the term data refers to a dataset that originates ad 4a) and interpret them

   (a) the interpretation shall prove that you can read the parameters and understand their meaning,

   (b) it is enough to analyze and explain one node/CPT with a proper number of parents (2-3).

6. Modify the network structure generated ad 3 given the data

   (a) the start graph is the structure ad 3, the data serve to further improve it (MCMC is the most straightforward way).

7. Take the data and generate a network structure from scratch (learn also its CPTs).

8. Compare the models obtained ad 5, 6 and 7, select the best one

   (a) compare the Bayesian networks intuitively,

   (b) compare them in terms of likelihood (resp. BIC and Bayesian score) calculated for the input data or its subset, the individual criteria shall be used properly with respect to the choice ad 4a,

   (c) summarize the reasons for selection of the optimal model, rename it as *bnet* and save it into the file *bnet.mat*

   - use commands *bnet=my_opt_bnet; save('bnet','bnet');*
   - the model shall be causal (you are supposed to change edge directions intuitively), however keep the Markov equivalence class, discuss the changes in report,

- verify that the model sorts the variables topologically (a lower triangular DAG matrix is zero), if not, reorganize the variables,
- check that the model is applicable before saving, e.g., the array *bnet.names* contains the variable names in the actual order (the array is filled automatically by the command *mk_bnet*, the content can be checked by e.g., *bnet.names('Age')*),
- do not change variable names (neither abbreviations nor uppercase/lowercase changes are permitted), keep strictly the variable names contained in the vector *names* in *read_lung.m*,

(d) *bnet* will be tested on unseen cases sampled from the identical distribution as the provided input data, the same will also be the process that introduces missing observations.

9. Use the network selected ad 8 for illustrative inference

   (a) propose a few samples with missing values and carry out inference,

   (b) comment/explain the computed distributions over missing values given the evidence (the context given by the observed variable states),

   (c) describe the application of the given network to predict all the missing values in all the samples.

10. Write the report and submit it altogether with your m files and bnet file

    (a) the report must answer the items 3-9 of Task,

    (b) it shall also include any observations beyond the items 3-9, write a brief summary.