SMU - Bayesian Networks Assignment

Ondřej Hubáček

April 2017

1 Submission and evaluation

1. Each student works individually.
2. The way of submission:
   (a) https://cw.felk.cvut.cz/brute/
   (b) deadline: Sun 16.4.2017 23:59
   (c) archive structure (username.zip):
      i. a file username.pdf with the report
      ii. a file username.png with the baseline network structure
      iii. a file username.bif with the selected model
      iv. a directory src with the .py files that underline the solution
3. Up to 13 points can be obtained for this assignment:
   (a) 10 points for the report and the functional source code
   (b) 3 points for performance reached by the model – the joint distribution that underlines your network will be compared with the original one, the network will be tested on unseen data too
   (c) there is a 3 point penalty for each commenced day of delay

2 Task

1. Get familiar with pgmpy https://github.com/pgmpy/pgmpy
2. Study the input dataset crash_txt.csv
3. Manually construct a baseline network structure that you find best for the given domain and interpret it
4. Think about dealing with the input data, in particular focus on:
   (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 the missing data treatment – implement either:

i. estimation of the missing values by EM+MLE

ii. estimation of the missing values by kNN

5. Learn the quantitative parameters (CPTs) of the baseline network from the data (the term data refers to the 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. Evaluate the baseline model and try to improve it using:

   (a) HillClimbing routine (or another structure learning algorithm)

   (b) handcrafted knowledge

7. Save the final network into the file username.bif

8. Write a brief (max 2 pages) report containing:

   (a) description of baseline network

   (b) how you dealt with missing data

   (c) interpretation of CPT from 5b)

   (d) how you derived the final network