SMU - Bayesian Networks Assignment

Ondřej Hubáček

April 2018

1 Submission and evaluation

1. Each student works individually.
2. The way of submission:
   (a) https://cw.felk.cvut.cz/brute/
   (b) deadline: Mon 30.4.2018 23:59
   (c) archive structure (username.zip):
      i. a file username.pdf with the report
      ii. a file username.bif with the selected model
      iii. 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
   (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_sample_2018.csv.
3. Manually construct a baseline network structure that you find best for the given domain.
4. Check if the following statements hold for your model (draw information flow diagram) and discuss if the findings make intuitive sense:
   (a) Season $\perp\!\!\!\perp$ NoFatalities | NoJourneys
   (b) Weather $\perp\!\!\!\perp$ NoAccidents | RoadCond

   (c) Season $\perp\!\!\!\perp$ Weekend | NoAccidents

5. 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 estimation of the missing values by EM+MLE

6. Learn the quantitative parameters (CPTs) of the baseline network from the training data 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).

7. Evaluate the baseline model and try to improve it using HillClimbing routine (or another structure learning algorithm) and handcrafted knowledge.

8. Train the final network using the whole dataset and report Jensen-Shannon divergence and total variation distance.

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

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

    (a) diagram and description of the baseline network

    (b) information flow diagram and brief discussion if the findings from (4) hold

    (c) interpretation of CPT from (6b)

    (d) how you derived the final network

    (e) Jensen-Shannon divergence and total variation distance from (8)