Knowledge and data engineering with Bayesian networks (a homework solution guide)

Péter Antal

Outline

- Tasks in the homework
- Knowledge enginering steps
  - Importance of ordering
  - Canonical models
    - Conditional probability decision trees

Homework

- Guide
- Tool
  - http://redmine.genagrid.eu/projects/bayescubedownload/wiki/Wiki
- Manual
- List of illustrative domains

Goal of the homework

To demonstrate and practice this multifaceted nature of Bayesian networks.

- As a probabilistic logic knowledge base, it provides a coherent framework to represents beliefs (see Bayesian interpretation of probabilities).
- As a decision network, it provides a coherent framework to represent preferences for actions.
- As a dependency map, it explicitly represents the system of conditional independencies in a given domain.
- As a causal map, it explicitly represents the system of causal relations in a given domain.
- As a decomposable probabilistic graphical model, it parsimoniously represents the quantitative stochastic dependencies (the joint distribution) of a domain and it allows efficient observational inference.
- As an uncertain causal model, it parsimoniously represents the quantitative, stochastic, autonomous mechanisms in a domain and it allows efficient interventional and counterfactual inference.

Obligatory and optional subtasks

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - Quantify the Bayesian networks.
  - Evaluate it with global inference and „information sensitivity of inference” analysis.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Consultation

The preliminary approval of your planned homework is mandatory!

Documentation

Domain description.	10-100 words
Variable definitions, with definitions of their values.	<20 words/variable
Structure of the Bayesian network.	Explain the (preferably) causal order of the variables and interesting independencies in your model. 50-500 words + figure(s).
Quantify the Bayesian networks.	Illustrate your estimation in your model. 50-200 words + table(s)/figure(s).
Evaluate it with global inference and „information sensitivity of inference" analysis.	20-100 words + table(s)/figure(s).
Compare the structural and parametric differences between the constructed and learnt models.	50-200 words.
Analyse estimation biases.	250-500 words + table(s)/figure(s).
Investigate the effect of model uncertainty and sample size on learning.	500-1000 words + table(s)/figure(s).

The overall documentation can be 3-5 pages (minimal) or 5-10 pages (full).

Submission

- After consultation(!)
- the model XML with its documentation should be sent by **email to your consulent.**
- **Deadlines:**
  - **Soft:** before the last week of the semester (5th of December)
  - **Hard**: before the end of the semester (15th of December).

Subtasks: importance of causality

- The minimal level contains the following subtasks (10 point):
  - **Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.**
  - Consult it.
  - Quantify the Bayesian networks.
  - Evaluate it with global inference and „information sensitivity of inference” analysis.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Subtasks: canonical models

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - **Quantify the Bayesian networks.**
  - Evaluate it with global inference and „information sensitivity of inference” analysis.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Noisy-OR

Noisy-OR distributions model multiple noninteracting causes

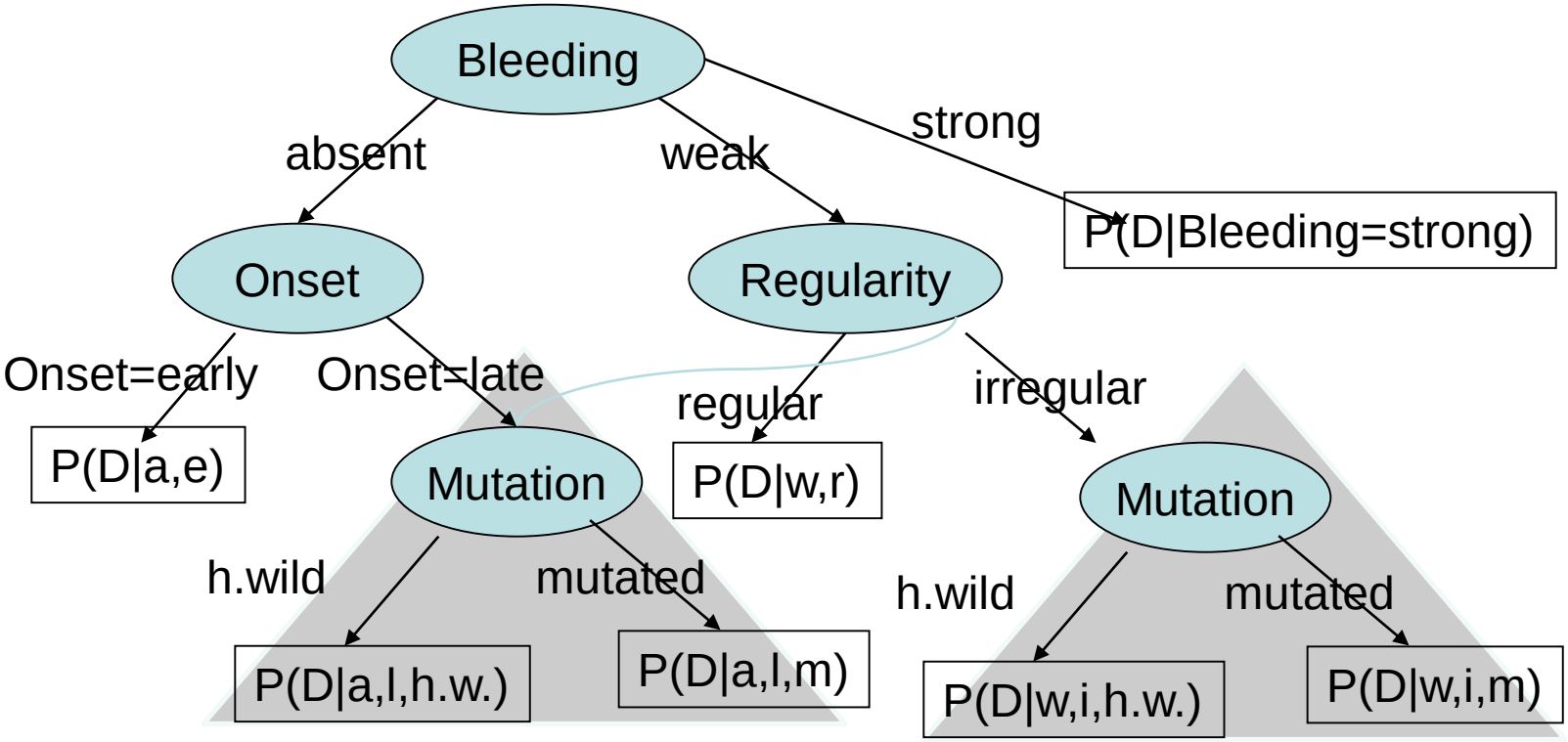
1) Parents $U_1 \dots U_k$ include all causes (can add leak node)
2) Independent failure probability q_i for each cause alone

$$\Rightarrow P(X|U_1 \dots U_j, \neg U_{j+1} \dots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters **linear** in number of parents

Decision trees, decision graphs



Decision tree: Each internal node represent a (univariate) test, the leafs contains the conditional probabilities given the values along the path.
Decision graph: If conditions are equivalent, then subtrees can be merged.
E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)

Subtasks: sensitivity of inference

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - Quantify the Bayesian networks.
  - **Evaluate it with global inference and „information sensitivity of inference" analysis.**
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Subtasks: sensitivity of inference

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - Quantify the Bayesian networks.
  - **Evaluate it with global inference and „information sensitivity of inference" analysis.**
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Subtasks: learn model

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - Quantify the Bayesian networks.
  - Evaluate it with global inference and „information sensitivity of inference” analysis.
  - **Generate a data set from your model.**
  - **Learn a model from your data.**
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Subtasks: estimation bias

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - Quantify the Bayesian networks.
  - Evaluate it with global inference and „information sensitivity of inference” analysis.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - **Analyse estimation biases** (5 point).
  - Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Subtasks: effect of model uncertainty and sample size on learning

- The minimal level contains the following subtasks (10 point):
  - Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
  - Consult it.
  - Quantify the Bayesian networks.
  - Evaluate it with global inference and „information sensitivity of inference” analysis.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
- Optional tasks:
  - Analyse estimation biases (5 point).
  - **Investigate the effect of model uncertainty and sample size on learning**: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning (10 point).

Summary

- The homework takes you through real stages of knowledge engineering and machine learning:
  - Select a domain, create variables (5-10), and specify structure.
  - Quantify the Bayesian network.
  - Analyse estimation biases
  - Evaluate it with „information sensitivity of inference” analysis.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models.
  - Investigate the effect of model uncertainty and sample size on learning.