Knowledge engineering with Bayesian networks (a homework solution guide)

Péter Antal
Outline

• Tasks in the homework
• Knowledge engineering steps
  – Importance of ordering
  – Canonical models
    • Conditional probability decision trees
Homework

• Guide
• Tool
  – BayesCube
• 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 represent 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

• Subtasks:
  – 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.
  – Extend your Bayesian network into a decision network.
  – Investigate the value of further information.

• Optional tasks:
  – Analyse estimation biases.
  – 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.
Consultation

The preliminary approval of your planned homework is mandatory!
Subtasks: importance of causality

- The minimal level contains the following subtasks:
  - 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.
  - 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.
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 $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability $q_i$ for each cause alone
   \[ P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i \]

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>$P(\text{Fever})$</th>
<th>$P(\neg\text{Fever})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.98</td>
<td>0.02 = 0.2 $\times$ 0.1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06 = 0.6 $\times$ 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12 = 0.6 $\times$ 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 $\times$ 0.2 $\times$ 0.1</td>
</tr>
</tbody>
</table>

Number of parameters **linear** in number of parents
Decision trees, decision graphs

Decision tree: Each internal node represents a (univariate) test, the leaves contain 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) \sim (Bleeding=weak, Regularity=irreg)

A.I.: BN homework guide
Subtasks: sensitivity of inference

• The minimal level contains the following subtasks:
  – 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.
  – 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.
Subtasks: sensitivity of inference

• The minimal level contains the following subtasks:
  – 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.
  – 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.
Subtasks: learn model

- The minimal level contains the following subtasks:
  - 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.
  - 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.
Reminder: Value of further information

Current evidence $E$, current best action $\alpha$
Possible action outcomes $S_i$, potential new evidence $E_j$

$$EU(\alpha|E) = \max_a \sum_i U(S_i) \ P(S_i|E, a)$$

Suppose we knew $E_j = e_{jk}$, then we would choose $\alpha_{e_{jk}}$ s.t.

$$EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) = \max_a \sum_i U(S_i) \ P(S_i|E, a, E_j = e_{jk})$$

$E_j$ is a random variable whose value is currently unknown
$\Rightarrow$ must compute expected gain over all possible values:

$$VPI_E(E_j) = \left( \sum_k P(E_j = e_{jk} | E) \ EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) \right) - EU(\alpha|E)$$

($VPI = value of perfect information$)
Subtask: test a decision network

- Investigate the value of further information as follows:
  - select values for some “evidence” variables \( E=e \),
  - using BayesCube calculate the current expected loss/utility \( EU(D|e) \),
  - select a variable “I” as potential “further” information,
  - using BayesCube calculate the conditional probabilities of potential further observations (i.e. the conditional probabilities of potential values of this “further information” variable, \( p(I=i|E=e) \)),
  - using BayesCube calculate the expected losses/utilities corresponding to these potential further observations \( EU(D|e,i) \),
  - calculate the (expected) value of (perfect) information corresponding to this variable “I”, \( \sum_i p(i|e) \cdot EU(D|e,i) - EU(D|e) \).
Optional subtasks: estimation bias

• The minimal level contains the following subtasks:
  – 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.**
  – 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.
Optional subtasks: effect of model uncertainty and sample size on learning

- The minimal level contains the following subtasks:
  - 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.
  - 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.
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.
  - Evaluate value of further information.
  - Investigate the effect of model uncertainty and sample size on learning.