An Introduction to Bayesian Networks: Representation and Approximate Inference

Marek Grześ

Department of Computer Science
University of York
Graphical Models Reading Group

May 7, 2009
Data and Probabilities

Representation of Bayesian Networks

Approximate Reasoning with Gibbs Sampling

References
<table>
<thead>
<tr>
<th>Bronchitis</th>
<th>Cancer</th>
<th>Smoker</th>
<th>Count</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>absent</td>
<td>absent</td>
<td>nonsmoker</td>
<td>3</td>
<td>0.03</td>
</tr>
<tr>
<td>absent</td>
<td>absent</td>
<td>smoker</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>absent</td>
<td>present</td>
<td>nonsmoker</td>
<td>27</td>
<td>0.27</td>
</tr>
<tr>
<td>absent</td>
<td>present</td>
<td>smoker</td>
<td>15</td>
<td>0.15</td>
</tr>
<tr>
<td>present</td>
<td>absent</td>
<td>nonsmoker</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>present</td>
<td>absent</td>
<td>smoker</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>present</td>
<td>present</td>
<td>nonsmoker</td>
<td>18</td>
<td>0.18</td>
</tr>
<tr>
<td>present</td>
<td>present</td>
<td>smoker</td>
<td>35</td>
<td>0.25</td>
</tr>
</tbody>
</table>

[Cuss 09]
Joint and Marginal Probability Distributions, Queries

- $P$ is a probability over $\mathcal{X} = \{X_1, ..., X_n\}$, a joint probability distribution is $P(X_1, ..., X_n)$
Joint and Marginal Probability Distributions, Queries

- $P$ is a probability over $\mathcal{X} = \{X_1, \ldots, X_n\}$, a joint probability distribution is $P(X_1, \ldots, X_n)$

- Inference - query types:
  - conditional probability query: $P(Y|E = e)$,
  - marginal probabilities: $P(X_i)$, $P(X_i, X_j)$, $P(Y)$,
  - likelihoods: $P(E)$,
  - maximum a posteriori query: find the most likely assignment to the variables $Y$ given the evidence $E = e$ : $\arg\max_y P(y|e)$. 
Queries: examples

- $P(Bronchitis|\text{Cancer} = \text{absent}, \text{Smoker} = \text{nonsmoker})$

<table>
<thead>
<tr>
<th>Bronchitis</th>
<th>Cancer</th>
<th>Smoker</th>
<th>Count</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>absent</td>
<td>absent</td>
<td>nonsmoker</td>
<td>3</td>
<td>0.6</td>
</tr>
<tr>
<td>present</td>
<td>absent</td>
<td>nonsmoker</td>
<td>2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

- $P(\text{Cancer}, \text{Smoker})$

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Smoker</th>
<th>Count</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>absent</td>
<td>nonsmoker</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td>absent</td>
<td>smoker</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>present</td>
<td>nonsmoker</td>
<td>45</td>
<td>0.45</td>
</tr>
<tr>
<td>present</td>
<td>smoker</td>
<td>50</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- Straightforward and natural, but complexity is a problem!
Graphical Models

- Graphical models are a marriage between probability theory and graph theory [Jord 99].
- Probability theory to deal with uncertainty.
- Graph theory to deal with complexity (compact representation and efficient reasoning).
A simple Bayesian network

- \( P(\text{Pneumonia, Tuberculosis, Lung Infiltrates, XRay, Sputum Smear}) \).
A simple Bayesian network

- $P(\text{Pneumonia, Tuberculosis, Lung Infiltrates, XRay, Sputum Smear})$.

[Koll 07]
A simple Bayesian network (ctd)

\[
P(P) = P(P) \cdot P(T) \cdot P(I|P,T) \cdot P(X|I) \cdot P(S|T)
\]

| $P$ | $T$ | $P(I|P,T)$ |
|-----|-----|------------|
| $p$ | $t$ | 0.8        |
| $p$ | $t$ | 0.6        |
| $p$ | $t$ | 0.2        |
| $p$ | $t$ | 0.01       |

| $I$ | $P(X|I)$ |
|-----|----------|
| $i$ | 0.8      |
| $\bar{i}$ | 0.6   |

| $S$ | $P(S|T)$ |
|-----|----------|
| $s$ | 0.8      |
| $\bar{s}$ | 0.6    |

[Koll 07]
A simple Bayesian network (ctd)

Product Rule of Probability and Independence

Why we can write this:

Product Rule of Probability and Independence

Why we can write this:


- Repeated application of the product rule of probability:

\[ P(X_1, ..., X_n) = P(X_n|X_1, ..., X_{n-1})...P(X_2|X_1)P(X_1). \]
Product Rule of Probability and Independence

Why we can write this:

Repeated application of the product rule of probability:
\[ P(X_1, ..., X_n) = P(X_n|X_1, ..., X_{n-1})\cdots P(X_2|X_1)P(X_1). \]
This can be seen a fully connected graph (Bayesian network) with each node having incoming links from all lower numbered nodes.
Product Rule of Probability and Independence

Why we can write this:


- Repeated application of the product rule of probability:
  \[ P(X_1, ..., X_n) = P(X_n|X_1, ..., X_{n-1})...P(X_2|X_1)P(X_1) \]
  This can be seen a fully connected graph (Bayesian network) with each node having incoming links from all lower numbered nodes.

- Independent variables:
  If \[ P(X_1, X_2) = P(X_1)P(X_2) \], then \( X_1 \) and \( X_2 \) are independent and \( P(X_2|X_1) = P(X_2) \).
Diagnostic Bayesian Networks

Figure: This is a common structure of diagnostic networks: predisposition nodes at the top (Visit to Asia, Smoking), diseases in the middle (Tuberculosis, Lung Cancer, Bronchitis), and symptoms at the bottom (XRay Result, Dyspnea).
Independence/Dependence in Bayesian Networks

- X is conditionally independent of Y given Z if
  
  \[ P(X = x, Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z); \]
  
  \( (X \perp Y | Z) \).

- A and B are marginally dependent, and A and B are conditionally independent.

- A and B are marginally independent, A and B are conditionally dependent.
X is \textit{conditionally independent} of Y given Z if
\[ P(X = x, Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z); \]
\( (X \perp Y | Z) \).

\( A \) and \( B \) are marginally dependent, and \( A \) and \( B \) are conditionally independent.

\[ 
\begin{array}{c}
\text{C} \\
\downarrow \\
\text{A} & \quad \text{B} \\
\end{array}
\quad 
\begin{array}{c}
\text{C} \\
\downarrow \\
\text{A} & \quad \text{B} \\
\end{array}
\]
Independence/Dependence in Bayesian Networks

▶ X is *conditionally independent* of Y given Z if
\[ P(X = x, Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z); \]
\( (X \perp Y | Z) \).

▶ A and B are marginally dependent, and A and B are conditionally independent.

▶ A and B are marginally independent, A and B are conditionally dependent.
Reasoning

- Computing conditional/marginal distributions.
- Exact inference (e.g., junction tree algorithm) is efficient:
  - chain-like graphs,
  - tree-like graphs.
- Approximate inference (variational methods, sampling, etc.) when:
  - dense graphs,
  - layered graphs,
  - coupled graphs [Jord 97].
Sampling

- We would like to compute marginal, $P(X_i)$, or conditional, $P(Y|E)$, probability.
- Let's assume that it is impossible to do this exactly.
- If we can create a generative model of the distribution we are interested in, we can sample from this distribution and compute desired probability empirically.
Markov Chains

Transition matrix:

\[
\begin{bmatrix}
[1,] & 0.4 & 0.2 & 0.3 & 0.1 \\
[2,] & 0.4 & 0.4 & 0.2 & 0.0 \\
[3,] & 0.6 & 0.2 & 0.1 & 0.1 \\
[4,] & 0.7 & 0.1 & 0.0 & 0.2
\end{bmatrix}
\]

Start state:

\[
\begin{bmatrix}
1 & 0 & 0 & 0
\end{bmatrix}
\]

After 3 time steps:

\[
\begin{bmatrix}
0.473 & 0.237 & 0.204 & 0.086
\end{bmatrix}
\]

Start state:

\[
\begin{bmatrix}
0 & 0 & 1 & 0
\end{bmatrix}
\]

After 3 time steps:

\[
\begin{bmatrix}
0.465 & 0.237 & 0.212 & 0.086
\end{bmatrix}
\]

[Cuss 09]
Markov Chain Monte Carlo - MCMC

- It is a simulation-based method, hence Monte Carlo.
- Sample from a sequence of distributions which gets progressively closer to the target distribution (query).
- Sampling from a Markov chain, $P_{MC}(S_{i+1}|S_i)$, $S_i = \{X^i_1, ..., X^i_N\}$.
- A stationary distribution more formally. $\pi$ the transition matrix for a Markov chain. If $P$ is a distribution such that $P = P\pi$ then $P$ is said to be a stationary distribution.
- The aim of Markov chain Monte Carlo methods is to design a Markov chain whose stationary distribution is the target distribution (query).
- The initial distribution does not matter.
- Initial, burn-in samples can be removed.
Gibbs Sampling - One Variable at a Time

- $\mathcal{X} = \{X_1, \ldots, X_n\}$ is a set of variables and a query is $P(Y|E)$.
- Set up Markov chain as follows:
  - initialise all $X_i \in Y$ arbitrarily (variables in $E$ are known),
  - chose $i$ (randomly or cycle over all variables in $Y$ taking each one in turn, but do not choose variables from $E$),
  - sample from $P(X_i|$\{\(E \cup \mathcal{X} \setminus X_i \setminus E\})$,
  - iterate (no samples rejected).
- It has been proved that this process has $P(Y|E)$ as its equilibrium.
- How to do this in graphical models?
Gibbs Sampling in Graphical Models

- In graphical models, when sampling from \( P(X_i | \{E \cup \mathcal{X} \setminus X_i \setminus E\}) \), the conditional independence can be exploited.

- Markov blanket: shields a variable from the influence of other variables.

- Thus, in BNs we can sample from \( P(X_i | \text{MarkovBlanket}(X_i)) \). We can always do this in Gibbs sampling because we sample only one variable at a time so we assume that all other variables are known (in the current state of the Markov chain).
Gibbs Sampling Does Not Always Work
Working Example

- Time for demo: visualisation of Gibbs sampling in graphical models.
References

“Algorithms for Graphical Models: Lecture Notes”.
2009.

“An Introduction to Graphical Models”.

Learning in Graphical Models.

Introduction to Statistical Relational Learning, Chap. Graphical Models in a Nutshell.