Bayesian Networks - Representation

CSE 446: Machine Learning
Emily Fox
University of Washington
March 6, 2017

Learning from structured data
TrueSkill: A Bayesian Skill Rating System

Herbrich et al., 2007

ICU Monitoring

Beinlich et al., 1989

Aleks, Russell, et al., 2008
Digging in:
Learning with and without context/structure

Without context: Handwriting recognition

Character recognition, e.g., kernel SVMs
Without context: Webpage classification

Company website

University website

Personal website

... 

With context: Handwriting recognition

"c" more likely to come after "a" than another "a"
With context: Webpage classification

Company pages tend to point to each other

Modeling structured relationships via Bayesian networks
Today – Bayesian networks

- Provided a huge advancement in AI/ML
- Generalizes naïve Bayes and logistic regression
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

Bayesian network representation

Compact representation of a probability distribution.

Directed Acyclic Graph

Vertices: Random Variables
Edges: Conditional dependencies
“probabilistic relationships”
Bayesian network probability factorization

One CPT (conditional probability table) for each variable

\[ P(\text{variable} | \text{parents of variable}) \]

implies the factorization:

\[ P(X) = \prod_{j=1}^{d} P(X[j] | \text{parents}(X[j])) \]

What a Bayesian network represents (in detail) and what does it buy you?
Causal structure

• Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches

• How are these connected?

(Not a true causal model)

Possible queries

• Inference
  \[ P(F=t \mid N=t) \]
  - infer state of some variable

• Most probable explanation
  \[ \max_{f,a,s,h} P(f,a,s,h \mid N=t) \]

• Active data collection
  - What variable should I observe next?
  \[ H=? \quad s=? \]
CarStarts? Bayesian network

- 18 binary attributes
- Inference
  \[ P(\text{BatteryAge}|\text{Starts}=f) = \frac{P(\text{BA}, S=f)}{P(S=f)} \]
  \[ = \sum_{a,f,l,c,...} p(a,f,l,c,...) \]
  (everything other than BatteryAge and Starts)
- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN - more than 3^{54} = 58149737003040059690390169 terms

Factored joint distribution – A preview

- \( P(F) \) and \( P(A) \)
- \( P(S|F,A) \)
- \( P(F,A,S,H,N) = P(F)P(A)P(S|F,A)P(H|S)P(N|S) \)
- \( 2^5 = 32 \) terms... 31 params
- Will see later why this fact holds
What are these probabilities?
Conditional probability tables (CPTs)

- Flu
- Allergy
- Sinus
- Headache
- Nose

\[
P(F) = 0.05 \\ P(F = \text{t}) = 0.95
\]

\[
P(A) = 0.2 \\ P(A = \text{t}) = 0.8
\]

\[
P(S | F, A) =
\begin{array}{c|cc}
S = \text{t} & S = \text{f} \\
F = \text{f}, A = \text{f} & 0.1 & 0.9 \\
F = \text{f}, A = \text{t} & 0.7 & 0.3 \\
F = \text{t}, A = \text{t} & 0.4 & 0.6 \\
\end{array}
\]

\[
P(H | S) = 2 \text{ params} \\
P(N | S) = 2 \text{ params}
\]

Number of parameters

- Flu
- Allergy
- Sinus
- Headache
- Nose

\[
P(F) \leq 1 \text{ param}
\]

\[
P(A) \leq 1 \text{ param}
\]

\[
P(S | F, A) \leq 4 \times (2-1) = 4 \text{ params}
\]

\[
P(H | S) \leq 2 \text{ params} \\
P(N | S) \leq 2 \text{ params}
\]

\[
P(F, A, S, H, N) \rightarrow 31 \text{ params}
\]

\[10 < 31\]

- more bias
- less flexible
+ need less data to learn
+ more accurate on smaller datasets

10 total params
Factorization speeds up inference

Exploit distributivity:

\[ P(F = x_F|N = t) \propto \sum_{x_A, x_S, x_H} P(F = x_F, A = x_A, S = x_S, H = x_H, N = t) \]

\[ = \sum_{x_A, x_S, x_H} P(F = x_F)P(A = x_A)P(S = x_S | F = x_F, A = x_A)P(H = x_H | S = x_S)P(N = t | S = x_S) \]

\[ = P(F = x_F) \sum_{x_A} P(A = x_A) \sum_{x_S} P(S = x_S | F = x_F, A = x_A) \sum_{x_H} P(H = x_H | S = x_S) \]

variable elimination algorithm

Key: Independence assumptions

Flu \perp N \mid S

- Flu only "causes" Nose through Sinus
- If you tell N=t, this changes prob. of F, but if I first tell you S=t, N doesn't affect prob. of F

Knowing sinus separates variables from each other
Marginal and conditional independence

(Marginal) Independence

- **Flu** and **Allergy** are (marginally) independent

\[
\begin{align*}
\mathcal{F} & \perp \mathcal{A} \\
P(A, F) & = P(A)P(F) \\
P(A|F) & = P(A)
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>Flu = t</th>
<th>Flu = f</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(F)</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Allergy = t</th>
<th>Allergy = f</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(A)</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Flu = t</th>
<th>Flu = f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy = t</td>
<td>0.4 x 0.2 = 0.8</td>
<td>0.4 x 0.8</td>
</tr>
<tr>
<td>Allergy = f</td>
<td>0.8 x 0.2</td>
<td>0.8 x 0.6</td>
</tr>
</tbody>
</table>
Conditional independence

• Flu and Headache are not (marginally) ind.
  \[ P(H=t | F=t) \neq P(H=t) \]

• Flu and Headache are independent given Sinus infection
  \[ P(H=t | F=t, S=t) = P(H=t | S=t) \]

• More generally:
  \[ X \perp Y | Z \]
  \[ P(X,Y | Z) = P(X | Z) P(Y | Z) \]
  \[ P(X | Y, Z) = P(X | Z) \]
What is a Bayes net assuming?

Local Markov Assumption: A variable X is independent of its non-descendents given its parents.

\[ E \perp A \mid B,C \]
\[ E \perp D \mid B,C \]
\[ F \perp B \mid E \]

Allows you to read off some simple conditional independence relationships.

Explaining away example

Local Markov Assumption: A variable X is independent of its non-descendents given its parents.

\[ F \perp A \text{ (cond. ind. given } \phi = \text{mag.}) \]
\[ F \perp A \mid S \text{ ?? don't know} \]
\[ P(F=t \mid A=t, S=t) \neq P(F=t \mid S=t) \]

Suppose \( P(F=t \mid S=t) \) is high but \( P(F=t \mid S=t, A=t) \) is lower because \( A=t \) explains away sinus inflammation.
Naïve Bayes revisited

\[ x_{[1]} \perp x_{[2]}, \ldots, x_{[d]} \mid y \]

\( \text{true for any feature and set of remaining features} \)

\[ p(y, x_{[1]}, \ldots, x_{[d]}) = p(y) \prod_{j=1}^{d} p(x_{[j]} \mid y) \quad \text{Naïve Bayes} \]

Local Markov assumption of this graph:

\[ x_{[j]} \perp x_{[1]}, \ldots, x_{[j-1]}, x_{[j+1]}, \ldots, x_{[d]} \mid y \]

\( \text{Naïve Bayes!} \)

Factorization of the joint distribution
Joint distribution

Why can we decompose?
Markov Assumption!

The chain rule of probabilities

• $P(A,B) = P(A)P(B|A)$

• More generally:
  - $P(X[1],...,X[d]) = P(X[1])P(X[2]|X[1]) \ldots P(X[d]|X[1],...,X[d-1])$
Chain rule & joint distribution

Local Markov Assumption: A variable X is independent of its non-descendents given its parents

Order of expansion matters! Use topological order

The Representation Theorem – Joint distribution to BN

If cond. ind. in Bayes net are subset of cond. ind. in P

Joint distribution factorizes:

\[
P(X) = \prod_{j=1}^{d} P(X[j] \mid \text{parents}(X[j]))
\]
Bayesian networks recap

- **Representation benefits**
  - Compact representation for probability distributions
  - Exponential reduction in number of parameters
  - Lower variance parameter estimates from limited data

- **Inference benefits**
  - Efficient computation of $P(X|e)$ (i.e., fast probabilistic inference)
  - Involves variable elimination algorithms

- **Other important topics**
  - Structure learning: What graph structure to use?
  - Understanding how evidence can be incorporated and how this changes conditional independence statements (d separation)

---

Hidden Markov models:
A Bayesian network for time series
Example: Motion Capture Segmentation

Hidden Markov model


Markov transition dynamics:
\[ \Pr(x_t = \text{[Jumping Jacks]} | x_{t-1} = \text{[Squats]}) = A_{\text{Jumping Jacks, Squats}} \]

\[ A = \begin{pmatrix}
1 & 2 & 3 \\
0.5 & 0.5 & 0 \\
0 & 1 & 0 \\
\vdots & \vdots & \vdots \\
0 & 0 & 1
\end{pmatrix} \]
Hidden Markov model


Markov transition dynamics:
\[ \Pr(x_t = \square | x_{t-1} = \square) = A_{\square, \square} \]

Conditionally independent emissions:
\[ \Pr(y_t | x_t = \square) = N(\mu_{\square}, \Sigma_{\square}) \]

Joint distribution factorization:

Latent Markov chain structure enables
- Efficient computation of marginals \( p(x_t | y_1, \ldots, y_T) \) via forward-backward alg.
- Most-probable sequence via Viterbi
- Parameter learning via Baum-Welch (EM for HMMs)

GMMs vs. HMMs

Gaussian mixture model

Hidden Markov model

True mode sequence
HMM applications

Example applications:
• Parsing EEG recordings
• Discovering behaviors in videos
• Speech segmentation
• Volatility regimes in financial time series
• Genomics
• ...

Incorporating evidence:
Bayes ball algorithm for analyzing conditional independencies

OPTIONAL
Conditional independence in Bayes nets

- Consider 4 different junction configurations

\[
\begin{array}{cccc}
\text{(a)} & \text{(b)} & \text{(c)} & \text{(d)} \\
\end{array}
\]

- Conditional versus unconditional independence:

\[
\begin{align*}
P(x, y, z) &= p(x)p(z)p(y|x, z) \\
p(x, z) &= p(x)p(z) \Rightarrow \forall y \quad p(x, y, z) \\
p(x, z | y) &= p(x)p(z)p(y|x, z) \\
&\neq p(x | y)p(z | y) \\
&\neq p(y|x)p(z | y) \\
\end{align*}
\]

Verifying $y = \text{earthquake}$, $z = \text{burglar}$, and $y = \text{car alarm}$:

If alarm ($y = 1$), an increase in earthquake ($p(x | y)$), means $p(z | y)$ lower.

Bayes ball algorithm

- Consider 4 different junction configurations

\[
\begin{array}{cccc}
\text{(a)} & \text{(b)} & \text{(c)} & \text{(d)} \\
\end{array}
\]

- Bayes ball algorithm:

Start ball at one end or other.
If ball passes to a node (straight arrows) then, not cond./marg. ind.
If ball bounces back (walls + curved arrows), the nodes are cond./marg. ind.
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H

V structure.
C not observed. Ball bounces away.
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H

V structure.
C observed. Ball can pass through

Ball gets stuck here
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H

V structure.
Descendent of F observed.
Ball can pass through
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H