Probabilistic Graphical Models

Lecture 2 – Bayesian Networks Representation

CS/CNS/EE 155
Andreas Krause
Announcements

- Will meet in Steele 102 for now
- Still looking for another 1-2 TAs..
- Homework 1 will be out soon. Start early!! 😊
Multivariate distributions

Instead of random variable, have random vector
\[ \mathbf{X}(\omega) = [X_1(\omega), ..., X_n(\omega)] \]

Specify \( P(X_1=x_1, ..., X_n=x_n) \)

Suppose all \( X_i \) are Bernoulli variables.

How many parameters do we need to specify?
Marginal distributions

Suppose we have joint distribution \( P(X_1, ..., X_n) \)

Then

\[
P(X_i = x_i) = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} P(x_1, \ldots, x_n)
\]

If all \( X_i \) binary: How many terms?
Rules for random variables

- Chain rule
  \[ P(x_1 \ldots x_n) = P(x_1) \cdot P(x_2 | x_1) \cdot \ldots \cdot P(x_n | x_1 \ldots x_{n-1}) \]

- Bayes’ rule
  \[ P(x | y) = \frac{P(y | x) \cdot P(x)}{P(y)} \]
  How do we get \( P(y) \)?
Key concept: Conditional independence

- Events \( \alpha, \beta \) conditionally independent given \( \gamma \) if
  \[
P(\alpha \land \beta | \gamma) = P(\alpha | \gamma) \cdot P(\beta | \gamma)
  \]

- Random variables \( X \) and \( Y \) conditionally independent given \( Z \) if for all \( x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z) \)
  \[
P(X = x, Y = y | Z = z) = P(X = x | Z = z) \cdot P(Y = y | Z = z)
  \]
  If \( P(Y = y | Z = z) > 0 \), that’s equivalent to
  \[
P(X = x | Z = z, Y = y) = P(X = x | Z = z)
  \]
  Similarly for sets of random variables \( X, Y, Z \)
  We write: \( P \models X \perp Y | Z \)
Why is conditional independence useful?

- \( P(X_1, \ldots, X_n) = P(X_1) \cdot P(X_2 \mid X_1) \cdot \ldots \cdot P(X_n \mid X_1, \ldots, X_{n-1}) \)

  How many parameters?

  \[
  2^0 + 2^1 + 2^2 + \ldots + 2^{n-1} = 2^n - 1
  \]

- Now suppose \( X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \mid X_i \) for all \( i \)

  Then

  \[
  P(X_1, \ldots, X_n) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_2) \cdot \ldots \cdot P(X_n \mid X_{n-1})
  \]

  How many parameters?

  \( 2^{m-1} \leq 2^n \)  

  Exponential reduction in # params

- Can we compute \( P(X_n) \) more efficiently?  
  Yes (often)
Properties of Conditional Independence

- **Symmetry**
  - $X \perp Y \mid Z \Rightarrow Y \perp X \mid Z$

- **Decomposition**
  - $X \perp Y, W \mid Z \Rightarrow X \perp Y \mid Z$

- **Contraction**
  - $(X \perp Y \mid Z) \land (X \perp W \mid Y, Z) \Rightarrow X \perp Y, W \mid Z$

- **Weak union**
  - $X \perp Y, W \mid Z \Rightarrow X \perp Y \mid Z, W$

- **Intersection**
  - $(X \perp Y \mid Z, W) \land (X \perp W \mid Y, Z) \Rightarrow X \perp Y, W \mid Z$
  - Holds only if distribution is positive, i.e., $P>0$
Key questions

- How do we specify distributions that satisfy particular independence properties?
  - **Representation**

- How can we exploit independence properties for efficient computation?
  - **Inference**

- How can we identify independence properties present in data?
  - **Learning**

Will now see example: Bayesian Networks
Key idea

- Conditional parameterization (instead of joint parameterization)
- For each RV, specify $P(X_i | X_A)$ for set $X_A$ of RVs
- Then use chain rule to get joint parametrization
  \[ P(x_1 \ldots x_m) = \prod P(X_i | x_{A \setminus i}) \]
- Have to be careful to guarantee legal distribution…
  
  \[ P(x_1 y), P(y_1 x) \]
  Does there exist $P(k_i y)$ with above joint cond. distributions
  Not in general
Example: 2 variables

\[ P(I > VH) = 0.8 \]

\[ P(G | I) \]

\[
\begin{array}{c}
\text{\(\sum\)} = 1 \\
\text{\(\sum\)} = 1 \\
\text{\(\sum\)} = 1 \\
0.8 \quad 0.2 \\
0.6 \quad 0.4
\end{array}
\]
Example: 3 variables

\[ P(I, S, G) = P(I) \cdot P(G|I) \cdot P(S|I) \]
Example: Naïve Bayes models

- Class variable $Y$
- Evidence variables $X_1,...,X_n$
- Assume that $X_A \perp X_B \mid Y$ for all subsets $X_A,X_B$ of $\{X_1,...,X_n\}$

Conditional parametrization:

- Specify $P(Y)$
- Specify $P(X_i \mid Y)$

Joint distribution

$$P(x_1,...,x_n,y) = P(y) \prod_{i} P(x_i \mid y)$$
Today: Bayesian networks

- Compact representation of distributions over large number of variables
- (Often) allows efficient exact inference (computing marginals, etc.)

HailFinder
56 vars
~ 3 states each

⇒ ~$10^{26}$ terms
> 10,000 years on Top supercomputers
Causal parametrization

- Graph with directed edges from (immediate) causes to (immediate) effects

\[
\begin{array}{c|c}
E & P(E) \\
\hline
T & 0.01 \\
F & 0.99 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c}
B & P(B) & P(E|B) & P(E|\neg B) \\
\hline
T & 0.33 & 0.97 & 0.03 \\
F & 0.67 & 0.03 & 0.97 \\
\end{array}
\]

Earthquake \quad Burglary

\quad Alarm

JohnCalls \quad MaryCalls
Bayesian networks

- A **Bayesian network structure** is a directed, acyclic graph $G$, where each vertex $s$ of $G$ is interpreted as a random variable $X_s$ (with unspecified distribution).

- A **Bayesian network** $(G, P)$ consists of
  - A BN structure $G$ and ..
  - ..a set of conditional probability distributions (CPDs) $P(X_s \mid \text{Pa}_{X_s})$, where $\text{Pa}_{X_s}$ are the parents of node $X_s$ such that
  - $(G, P)$ defines joint distribution

$$P(X_1, \ldots, X_n) = \prod_{i} P(X_i \mid \text{Pa}_{X_i})$$
Bayesian networks

Can every probability distribution be described by a BN?

\[ P(x_1, \ldots, x_n) = P(x_1) P(x_2| x_1) \cdots P(x_n| x_1, \ldots, x_{n-1}) \]
Representing the world using BNs

True distribution $P'$ with cond. ind. $I(P')$

- Want to make sure that $I(P) \subseteq I(P')$
- Need to understand CI properties of BN $(G,P)$

Bayes net $(G,P)$ with $I(P)$
Which kind of CI does a BN imply?

\[
P(E,B) = \sum_{A,J,M} P(E,B,A,J,M)
\]

\[
= \sum_{A,J,M} P(E)P(B)P(A|E,B)P(U|A)P(M|U)
\]

\[
= P(E)P(B) \sum_{A,J,M} P(A|E,B)P(U|A)P(M|U)
\]

\[
= P(E)P(B)
\]
Which kind of CI does a BN imply?

\[ J \perp M \mid A \]

\[ P(J \mid AM) = \frac{P(J, A, M)}{P(A, M)} \]

\[ P(J, A, M) = \sum_{E, B} P(J, A, M; E, B) \]

\[ = \sum_{E, B} P(E) P(B) P(A\mid E, B) P(J\mid A) P(M\mid A) \]

\[ = P(J\mid A) P(M\mid A) \sum_{E, B} P(E) P(B) P(A\mid E, B) \]

\[ = P(J\mid A) P(M\mid A) \underbrace{\sum_{E, B} P(E) P(B)}_{P(A)} = P(A) \]

\[ \Rightarrow P(J \mid AM) = \frac{P(J, A, M)}{P(AM)} = \frac{P(J\mid A) P(A\mid M)}{P(AM)} = P(J \mid A) \]

\[ \square \]
Local Markov Assumption

Each BN Structure $G$ is associated with the following conditional independence assumptions

$$X \perp \text{NonDescendents}_X \mid \text{Pa}_X$$

We write $I_{\text{loc}}(G)$ for these conditional independences

Suppose $(G, P)$ is a Bayesian network representing $P$. Does it hold that $I_{\text{loc}}(G) \subseteq I(P)$? If this holds, we say $G$ is an I-map for $P$. 
Factorization Theorem

$I_{\text{loc}}(G) \subseteq I(P)$

G is an I-map of P (independence map)

True distribution $P$ can be represented exactly as

$$P(X_1, ..., X_n) = \prod_{i} P(X_i \mid Pa_{X_i})$$

i.e., P can be represented as a Bayes net (G,P)
True distribution $P$ can be represented exactly as a Bayes net $(G, P)$

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \text{Pa}_{X_i})$$

$I_{\text{loc}}(G) \subseteq I(P)$

$G$ is an I-map of $P$ (independence map)
Proof: I-Map to factorization

Ordering $\Pi : \{i, \ldots, n\} \rightarrow \{1, \ldots, n\}$ topological

If: $\forall X_j$ descendant of $X_i$

$\Pi(j) > \Pi(i)$

Can find topological ordering in linear time

$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_{\Pi(i)} | x_{\Pi(i-1)}, \ldots, x_{\Pi(1)})$

Chain rule
True distribution $P$ can be represented exactly as a Bayes net $(G, P)$

$$P(X_1, ..., X_n) = \prod_{i} P(X_i \mid \text{Pa}_{X_i})$$

$I_{\text{loc}}(G) \subseteq I(P)$

$G$ is an I-map of $P$ (independence map)
The general case

\[ P(x_1, \ldots, x_n) = \prod P(x_i | Pa_{x_i}) \Rightarrow \forall x_i: \forall N \leq \text{NonDesc}_{x_i} \]

\[ x_i \perp N | Pa_{x_i} \]

\[ Y = Pa_{x_i} \]

\[ Z = \text{NonDesc}_{x_i} \setminus (Y \cup N) \]

\[ D = \text{Desc} (x_i) \]

\[ P(x_i | y, n) = \frac{P(x_i, y, n)}{P(y | n)} \]

\[ P(x_i, y, n) = \sum_{z \in D} P(x_i, y, n, z) \cdot D \]

\[ = \sum_{z \in D} P(x_i | y) \prod_{x \in D} P(x | Pa_x) \prod_{x' \in (N \cup Z \cup Y)} P(x' | Pa_{x'}) \]

\[ P(y, n) = \sum_{x_i} P(x_i, y, n) = \sum_{x_i} \prod_{x' \in (N \cup Z \cup Y)} P(x' | Pa_{x'}) \cdot \sum_{x_i} P(x_i | y) \]

\[ \Rightarrow P(x_i | y, n) = P(x_i | y) \]

\[ \square \]
Factorization Theorem

\[ I_{\text{loc}}(G) \subseteq I(P) \]

G is an I-map of P (independence map)

True distribution P can be represented exactly as Bayesian network (G,P)

\[ P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \text{Pa}_{X_i}) \]
Defining a Bayes Net

- Given random variables and known conditional independences
- Pick ordering $X_1, \ldots, X_n$ of the variables
- For each $X_i$
  - Find minimal subset $A \subseteq \{X_1, \ldots, X_{i-1}\}$ such that $X_i \perp X_{\neg A} \mid A$, where $\neg A = \{X_1, \ldots, X_n\} \setminus A$
  - Specify / learn $\text{CPD}(X_i \mid A)$

Ordering matters a lot for compactness of representation! More later this course.
Adding edges doesn’t hurt

**Theorem:**
Let $G$ be an I-Map for $P$, and $G'$ be derived from $G$ by adding an edge. Then $G'$ is an I-Map of $P$ ($G'$ is strictly more expressive than $G$)

**Proof:** \(\text{wtr: }\mathcal{I}_{\text{loc}}(G') \subseteq \mathcal{I}_{\text{loc}}(G)\)

Then \(\mathcal{I}_{\text{loc}}(G') \subset \mathcal{I}(P)\) since \(\mathcal{I}_{\text{loc}}(G) \subset \mathcal{I}(P)\)

\(\text{wtr: }X \perp \text{Nondesc}(X;G) \mid \text{Pa}(X;G)\)

\(\Rightarrow X \perp \text{Nondesc}(X;G') \mid \text{Pa}(X;G')\)

\(X \perp N \mid Y \Rightarrow X \perp N \mid Y, Z\)

holds by: Weak Union property of C.I. D
Additional conditional independencies

- BN specifies joint distribution through conditional parameterization that satisfies Local Markov Property
- But we also talked about additional properties of CI
  - Weak Union, Intersection, Contraction, ...
- Which additional CI does a particular BN specify?
  - All CI that can be derived through algebraic operations

Local Markov prop. $I_{\text{loc}}(G) \leq I(G)$
What you need to know

- Bayesian networks
- Local Markov property
- I-Maps
- Factorization Theorem
Tasks

- Subscribe to Mailing list
  https://utilits.its.caltech.edu/mailman/listinfo/cs155

- Read Koller & Friedman Chapter 3.1-3.3

- Form groups and think about class projects. If you have difficulty finding a group, email Pete Trautman