CS440/ECE448 Lecture 18: Bayesian Networks

Slides by Svetlana Lazebnik, 10/2016
Modified by Mark Hasegawa-Johnson, 10/2017
Review: Bayesian inference

- A general scenario:
  - Query variables: $X$
  - Evidence (observed) variables and their values: $E = e$

- **Inference problem**: answer questions about the query variables given the evidence variables

- This can be done using the posterior distribution $P(X | E = e)$

- Example of a useful question: **Which $X$ is true?**
  - More formally: what value of $X$ has the least probability of being wrong?
  - Answer: **MPE = MAP** (argmin $P(error)$ = argmax $P(X=x|E=e)$)
Today: What if $P(X,E)$ is complicated?

• Very, very common problem: $P(X,E)$ is complicated because both $X$ and $E$ depend on some hidden variable $Y$

• SOLUTION:
  • Draw a bunch of circles and arrows that represent the dependence
  • When your algorithm performs inference, make sure it does so in the order of the graph

• FORMALISM: Bayesian Network
Hidden Variables

- A general scenario:
  - Query variables: \( X \)
  - Evidence (observed) variables and their values: \( E = e \)
  - Unobserved variables: \( Y \)

- **Inference problem**: answer questions about the query variables given the evidence variables
  - This can be done using the posterior distribution \( P(X | E = e) \)
  - In turn, the posterior needs to be derived from the full joint \( P(X, E, Y) \)

\[
P(X | E = e) = \frac{P(X, e)}{P(e)} \propto \sum_y P(X, e, y)
\]

- Bayesian networks are a tool for representing joint probability distributions efficiently
Bayesian networks

• More commonly called *graphical models*
• A way to depict conditional independence relationships between random variables
• A compact specification of full joint distributions
Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Constructing a Bayesian network
- Conditional independence ≠ Independence
- Real-world examples
Bayesian networks: Structure

- **Nodes**: random variables

- **Arcs**: interactions
  - An arrow from one variable to another indicates direct influence
  - Must form a directed, *acyclic* graph
Example: N independent coin flips

- Complete independence: no interactions
Example: Naïve Bayes document model

- Random variables:
  - $X$: document class
  - $W_1, \ldots, W_n$: words in the document
Outline

• Review: Bayesian inference
• Bayesian network: graph semantics
• The Los Angeles burglar alarm example
• Constructing a Bayesian network
• Conditional independence ≠ Independence
• Real-world examples
Example: Los Angeles Burglar Alarm

• I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  • Example inference task: suppose Mary calls and John doesn’t call. What is the probability of a burglary?

• What are the random variables?
  • Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

• What are the direct influence relationships?
  • A burglar can set the alarm off
  • An earthquake can set the alarm off
  • The alarm can cause Mary to call
  • The alarm can cause John to call
Example: Burglar Alarm

- Burglary
- Earthquake
- Alarm
  - JohnCalls
  - MaryCalls
Conditional independence and the joint distribution

• Key property: each node is conditionally independent of its non-descendants given its parents

• Suppose the nodes $X_1, \ldots, X_n$ are sorted in topological order

• To get the joint distribution $P(X_1, \ldots, X_n)$, use chain rule:

\[
P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid X_1,\ldots,X_{i-1})
\]

\[
= \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))
\]
Conditional probability distributions

• To specify the full joint distribution, we need to specify a conditional distribution for each node given its parents:
  \[ P(X \mid \text{Parents}(X)) \]
Example: Burglar Alarm

- $P(B)$: Burglary
- $P(E)$: Earthquake
- $P(J|A)$: John Calls
- $P(M|A)$: Mary Calls
- $P(A|B, E)$: Alarm

The diagram illustrates the conditional probabilities and relationships between events such as burglary, earthquake, and the alarm and the calls from John and Mary.
Example: Burglar Alarm

- A “model” is a complete specification of the dependencies.
- The conditional probability tables are the model parameters.
Outline

• Review: Bayesian inference
• Bayesian network: graph semantics
• The Los Angeles burglar alarm example
• Constructing a Bayesian network
• Conditional independence ≠ Independence
• Real-world examples
Constructing Bayesian networks

1. Choose an ordering of variables $X_1, \ldots, X_n$

2. For $i = 1$ to $n$
   - add $X_i$ to the network
   - select parents from $X_1, \ldots, X_{i-1}$ such that $P(X_i \mid \text{Parents}(X_i)) = P(X_i \mid X_1, \ldots, X_{i-1})$
Example

• Suppose we choose the ordering M, J, A, B, E
Example

- Suppose we choose the ordering M, J, A, B, E

\[ \text{MaryCalls} \text{ JohnCalls} \]
Example

• Suppose we choose the ordering M, J, A, B, E
Example

• Suppose we choose the ordering M, J, A, B, E
Example

• Suppose we choose the ordering M, J, A, B, E
Example

• Suppose we choose the ordering M, J, A, B, E
Example

• Suppose we choose the ordering M, J, A, B, E
Example

• Suppose we choose the ordering M, J, A, B, E
Example contd.

- Deciding conditional independence is hard in noncausal directions
  - The causal direction seems much more natural
- Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed (vs. $1+1+4+2+2=10$ for the causal ordering)
Why store it in causal order? A: Saves memory

• Suppose we have a Boolean variable $X_i$ with $k$ Boolean parents. How many rows does its conditional probability table have?
  • $2^k$ rows for all the combinations of parent values
  • Each row requires one number for $P(X_i = \text{true} \mid \text{parent values})$

• If each variable has no more than $k$ parents, how many numbers does the complete network require?
  • $O(n \cdot 2^k)$ numbers – vs. $O(2^n)$ for the full joint distribution

• How many nodes for the burglary network?
  $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)
Outline

• Review: Bayesian inference
• Bayesian network: graph semantics
• The Los Angeles burglar alarm example
• Constructing a Bayesian network
• Conditional independence ≠ Independence
• Real-world examples
The joint probability distribution

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i)) \]

For example,

\[ P(j, m, a, \neg b, \neg e) = P(\neg b) P(\neg e) P(a \mid \neg b, \neg e) P(j \mid a) P(m \mid a) \]
Independence

• By saying that $X_i$ and $X_j$ are independent, we mean that $P(X_i | X_j) = P(X_i)$
• $X_i$ and $X_j$ are independent if and only if they have no common ancestors
• Example: *independent coin flips*

![Diagram of independent coin flips]

• Another example: Weather is independent of all other variables in this model.

![Diagram of weather, cavity, toothache, catch relationships]
Conditional independence

• By saying that $W_i$ and $W_j$ are conditionally independent given $X$, we mean that $P(W_i|X, W_j) = P(W_i|X)$

• $W_i$ and $W_j$ are conditionally independent given $X$ if and only if they have no common ancestors other than the ancestors of $X$.

• Example: naïve Bayes model:
The meaning of this graph is that $W_i$ and $W_j$ are conditionally independent given $X$. $P(W_i|X, W_j) = P(W_i|X)$ is the meaning of the graph.
Conditional independence ≠ Independence

Being conditionally independent given X does NOT mean that $W_i$ and $W_j$ are independent. Quite the opposite.

Suppose $P(X) = 1/2$, and $w_i = x$, and $w_j = x$. Then $P(W_i) = \frac{1}{2}$, but $P(W_i|W_j) = 1$. 
Conditional independence

Another example: *causal chain*

![Diagram](image)

- The meaning of this graph is that X and Z are conditionally independent given Y: $P(z|x, y) = P(z|y)$.

- Being conditionally independent given Y does NOT mean that X and Z are independent. Quite the opposite. For example, suppose $P(X) = 0.5$, $P(Y|X) = 0.8$, $P(Y|\neg X) = 0.1$, $P(Z|Y) = 0.7$, and $P(Z|\neg Y) = 0.4$. Then we can calculate that $P(Z|X) = 0.64$, but $P(Z) = 0.535$. 
Conditional independence ≠ Independence

Common cause

- Are X and Z independent?
  - No
  - $P(Z|X) \neq P(Z)$

- Are they conditionally independent given Y?
  - Yes
  - $P(Z|X, Y) = P(Z|Y)$

Y: Project due
X: Newsgroup busy
Z: Lab full

Common effect

- Are X and Z independent?
  - Yes
  - $P(Z|X) = P(Z)$

- Are they conditionally independent given Y?
  - No
  - $P(Z|X, Y) \neq P(Z|Y)$

X: Raining
Z: Ballgame
Y: Traffic
Outline

• Review: Bayesian inference
• Bayesian network: graph semantics
• The Los Angeles burglar alarm example
• Constructing a Bayesian network
• Conditional independence ≠ Independence
• Real-world examples
A more realistic Bayes Network: Car diagnosis

- **Initial observation:** car won’t start
- **Orange:** “broken, so fix it” nodes
- **Green:** testable evidence
- **Gray:** “hidden variables” to ensure sparse structure, reduce parameters
Car insurance
In research literature...

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data
Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan
(22 April 2005) Science 308 (5721), 523.
In research literature...

Fig. 3 A parametric, fixed-order model which describes the visual appearance of $L$ object categories via a common set of $K$ shared parts. The $j^{th}$ image depicts an instance of object category $o_j$, whose position is determined by the reference transformation $\rho_j$. The appearance $w_{ji}$ and position $v_{ji}$, relative to $\rho_j$, of visual features are determined by assignments $z_{ji} \sim x_{o_j}$ to latent parts. The cartoon example illustrates how a wheel part might be shared among two categories, bicycle and cannon. We show feature positions (but not appearance) for two hypothetical samples from each category.

Describing Visual Scenes Using Transformed Objects and Parts
E. Sudderth, A. Torralba, W. T. Freeman, and A. Willsky.

In research literature...

Audiovisual Speech Recognition with Articulator Positions as Hidden Variables
Mark Hasegawa-Johnson, Karen Livescu, Partha Lal and Kate Saenko

In research literature...

Detecting interaction links in a collaborating group using manually annotated data
S. Mathur, M.S. Poole, F. Pena-Mora, M. Hasegawa-Johnson, N. Contractor

Social Networks 10.1016/j.socnet.2012.04.002
In research literature...

- **Link**: \( L_{ij} = 1 \) if \( #i \) is listening to \( #j \).
- **Indirect**: \( I_{ij} = 1 \) if \( #i \) and \( #j \) are both listening to the same person.
- **Speaking**: \( S_i = 1 \) if the \( i \)'th person is speaking.
- **Gaze**: \( G_{ij} = 1 \) if \( #i \) is looking at \( #j \).
- **Neighborhood**: \( N_{ij} = 1 \) if they’re near one another.

Detecting interaction links in a collaborating group using manually annotated data

S. Mathur, M.S. Poole, F. Pena-Mora, M. Hasegawa-Johnson, N. Contractor

*Social Networks* 10.1016/j.socnet.2012.04.002
Summary

• Bayesian networks provide a natural representation for (causally induced) conditional independence
• Topology + conditional probability tables
• Generally easy for domain experts to construct