



cmsc726: Graphical Models

material from: Michael Jordan, Nir Friedman and Daphne Koller

Describing Data

- The canonical descriptive strategy is to describe the data in terms of their underlying distribution
- As usual, we have a p -dimensional data matrix with variables X_1, \dots, X_p
 - A variable describes sets of collectively mutually exclusive events, i.e. not two variables for HasDisease and Healthy
 - Variable represents a value, not a probability, i.e., Smoker, not Chance of Smoker
 - Clarity Test: knowable in principle, Gas Price: cents per gallon versus cheap, expensive (unless cheap and expensive are defined precisely in terms of price)
- The joint distribution is $P(X_1, \dots, X_p)$
- The joint gives us complete information about the variables
- Given the joint distribution, we can answer any question about the relationships among any subset of variables
 - are X_2 and X_5 independent?
 - What is the most likely value for disease, given test1 is positive?
 - If I want to determine whether someone has disease 1, what test will give me the most information?

Graphical Models

- In the next 3-4 lectures, we will be studying graphical models
- e.g. Bayesian networks, Bayes nets, Belief nets, Markov networks, etc.
- We will study:
 - representation
 - reasoning
 - learning
- Materials based on upcoming books by Nir Friedman and Daphne Koller and Michael Jordan.

Probability Distributions

- Let X_1, \dots, X_p be random variables
- Let P be a joint distribution over X_1, \dots, X_p

If the variables are binary, # of parameters?
we need $O(2^p)$ parameters to describe P

Can we do better?

- **Key idea:** use properties of independence

Independent Random Variables

- X is **independent** Y if
 - $P(X = x|Y = y) = P(X = x)$ for all values x, y
 - That is, learning the values of Y does not change prediction of X
- If X and Y are independent then
 - $P(X, Y) = ?$
 - $P(X, Y) = P(X|Y)P(Y) = P(X)P(Y)$
- In general, if X_1, \dots, X_p are independent, then
 - $P(X_1, \dots, X_p) = ?$
 - $P(X_1) \dots P(X_p)$
 - # of parameters?
 - $O(n)$ parameters

Conditional Independence

- Unfortunately, most of random variables of interest are not independent of each other
- A more suitable notion is that of **conditional independence**
- Two variables X and Y are **conditionally independent** given Z if
 - $P(X = x|Y = y, Z = z) = P(X = x|Z = z)$ for all values x, y, z
 - That is, learning the values of Y does not change prediction of X once we know the value of Z
 - notation: $I(X, Y | Z)$

Example: Naïve Bayesian Model

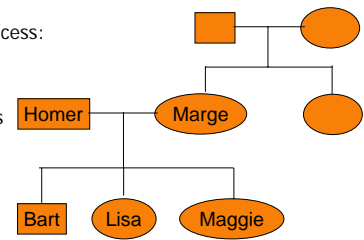
- A common model in early diagnosis:
 - Symptoms are conditionally independent given the disease (or fault)
- Thus, if
 - X_1, \dots, X_p denote whether the symptoms exhibited by the patient (headache, high-fever, etc.) and
 - H denotes the hypothesis about the patients health
- then, $P(X_1, \dots, X_p, H) = P(H)P(X_1|H) \dots P(X_p|H)$,
- This **naïve Bayesian** model allows compact representation
 - It does embody strong independence assumptions

Example: Family trees

Noisy stochastic process:

Example: Pedigree

- A node represents an individual's genotype

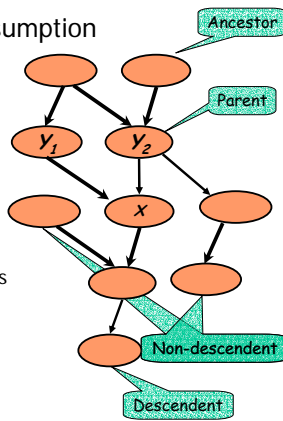


Modeling assumptions:

Ancestors can effect descendants' genotype only by passing genetic materials through intermediate generations

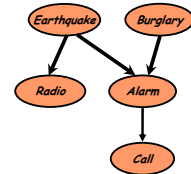
Markov Assumption

- We now make this independence assumption more precise for **directed acyclic graphs** (DAGs)
- Each random variable X , is independent of its non-descendants, given its parents $Pa(X)$
- Formally, $I(X, NonDesc(X) | Pa(X))$



Markov Assumption Example

- In this example:
 - $I(E, B)$
 - $I(B, \{E, R\})$
 - $I(R, \{A, B, C\} | E)$
 - $I(A, R | B, E)$
 - $I(C, \{B, E, R\} | A)$



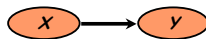
I-Maps

- A DAG G is an **I-Map** of a distribution P if the all Markov assumptions implied by G are satisfied by P (Assuming G and P both use the same set of random variables)

Examples:



x	y	$P(x,y)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25



x	y	$P(x,y)$
0	0	0.2
0	1	0.3
1	0	0.4
1	1	0.1

Factorization

- Given that G is an I-Map of P , can we simplify the representation of P ?

- Example:



- Since $I(X, Y)$, we have that $P(X|Y) = P(X)$
- Applying the chain rule

$$P(X, Y) = P(X|Y) P(Y) = P(X) P(Y)$$

- Thus, we have a simpler representation of $P(X, Y)$

Factorization Theorem

Thm: if G is an I-Map of P , then

$$P(X_1, \dots, X_p) = \prod_i P(X_i | Pa(X_i))$$

Proof:

- By chain rule: $P(X_1, \dots, X_p) = \prod_i P(X_i | X_1, \dots, X_{i-1})$
- wlog. X_1, \dots, X_p is an ordering consistent with G

From assumption: $Pa(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$

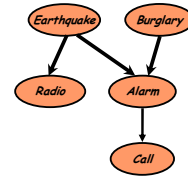
$$\{X_1, \dots, X_{i-1}\} - Pa(X_i) \subseteq \text{NonDesc}(X_i)$$

- Since G is an I-Map, $I(X_i, \text{NonDesc}(X_i) | Pa(X_i))$
- Hence,

$$I(X_i, \{X_1, \dots, X_{i-1}\} - Pa(X_i) | Pa(X_i))$$

- We conclude, $P(X_i | X_1, \dots, X_{i-1}) = P(X_i | Pa(X_i))$

Factorization Example



$$P(C, A, R, E, B) = P(B)P(E)P(R|E)P(A|B, E)P(C|A, R, E)$$

versus

$$P(C, A, R, E, B) = P(B) P(E) P(R|E) P(A|B, E) P(C|A)$$

Consequences

- We can write P in terms of "local" conditional probabilities

If G is **sparse**,

- that is, $|Pa(X_i)| < k$,

\Rightarrow each conditional probability can be specified compactly

- e.g. for binary variables, these require $O(2^k)$ params.

\Rightarrow representation of P is **compact**

- linear in number of variables

Pause...

We defined the following concepts

- The Markov Independences of a DAG G
 - $I(X_i, \text{NonDesc}(X_i) | Pa_i)$
- G is an I-Map of a distribution P
 - If P satisfies the Markov independencies implied by G

We proved the factorization theorem

- if G is an I-Map of P , then

$$P(X_1, \dots, X_n) = \prod_i P(X_i | Pa_i)$$

Conditional Independencies

- Let $\text{Markov}(G)$ be the set of Markov Independencies implied by G
- The factorization theorem shows

$$G \text{ is an I-Map of } P \Rightarrow P(X_1, \dots, X_n) = \prod_i P(X_i | Pa_i)$$

- We can also show the opposite:

Thm:

$$P(X_1, \dots, X_n) = \prod_i P(X_i | Pa_i) \Rightarrow G \text{ is an I-Map of } P$$

Implied Independencies

- Does a graph G imply additional independencies as a consequence of $\text{Markov}(G)$?

- We can define a **logic** of independence statements

- Some axioms:

- $I(X; Y | Z) \Rightarrow I(Y; X | Z)$
- $I(X; Y_1, Y_2 | Z) \Rightarrow I(X; Y_1 | Z)$

d-separation

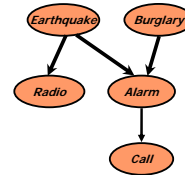
- A procedure $d\text{-sep}(X; Y | Z, G)$ that given a DAG G , and sets X , Y , and Z returns either *yes* or *no*
- **Goal:**
 $d\text{-sep}(X; Y | Z, G) = \text{yes}$ iff $I(X; Y | Z)$ follows from $\text{Markov}(G)$

Paths

- **Intuition:** dependency must “flow” along paths in the graph
- A path is a sequence of neighboring variables

Examples:

- $R \leftarrow E \rightarrow A \leftarrow B$
- $C \leftarrow A \leftarrow E \rightarrow R$



Paths

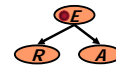
- We want to know when a path is
 - **active** -- creates dependency between end nodes
 - **blocked** -- cannot create dependency end nodes
- We want to classify situations in which paths are active.

Path Blockage

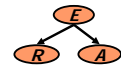
Three cases:

- Common cause

Blocked



Active

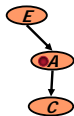


Path Blockage

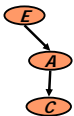
Three cases:

- Common cause
- Intermediate cause
-

Blocked



Active

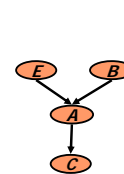


Path Blockage

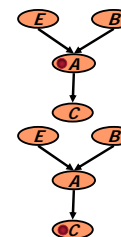
Three cases:

- Common cause
- Intermediate cause
- Common Effect

Blocked



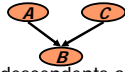
Active



Path Blockage -- General Case

A path is active, given evidence Z , if

- Whenever we have the configuration



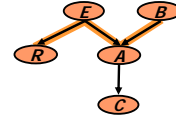
B or one of its descendants are in Z

- No other nodes in the path are in Z

A path is blocked, given evidence Z , if it is not active.

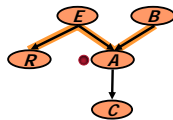
Example

- Blocked (R,B) ?



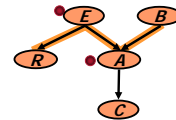
Example

- Blocked (R,B) = yes
- Blocked $(R,B|A)$?



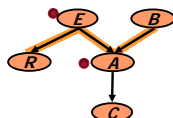
Example

- Blocked (R,B) = yes
- Blocked $(R,B|A)$ = no
- Blocked $(R,B|E,A)$?



Example

- Blocked (R,B) = yes
- Blocked $(R,B|A)$ = no
- Blocked $(R,B|E,A)$ = yes



d-Separation

- X is **d-separated** from Y , given Z , if all paths from a node in X to a node in Y are blocked, given Z .
- Checking d-separation can be done efficiently (linear time in number of edges)
 - Bottom-up phase: Mark all nodes whose descendants are in Z
 - X to Y phase: Traverse (BFS) all edges on paths from X to Y and check if they are blocked

Soundness

Thm:

- If
 - G is an I-Map of P
 - $d\text{-sep}(X; Y | Z, G) = \text{yes}$
- then
 - P satisfies $I(X; Y | Z)$

Informally,

- Any independence reported by d-separation is satisfied by underlying distribution

Completeness

Thm:

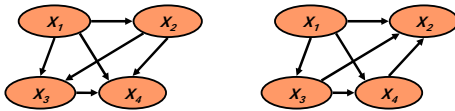
- If $d\text{-sep}(X; Y | Z, G) = \text{no}$
- then there is a distribution P such that
 - G is an I-Map of P
 - P does not satisfy $I(X; Y | Z)$

Informally,

- Any independence not reported by d-separation might be violated by the underlying distribution
- We cannot determine this by examining the graph structure alone

I-Maps revisited

- The fact that G is I-Map of P might not be that useful
- For example, **complete** DAGs
 - A DAG is G is complete if we cannot add an arc without creating a cycle



- These DAGs do not imply any independencies
- Thus, they are I-Maps of any distribution

Minimal I-Maps

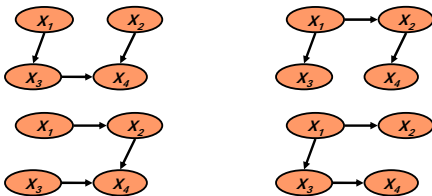
A DAG G is a **minimal I-Map** of P if

- G is an I-Map of P
- If $G' \subset G$, then G' is not an I-Map of P

Removing any arc from G introduces (conditional) independencies that do not hold in P

Minimal I-Map Example

- If is a minimal I-Map
- Then, these are **not** I-Maps:



Constructing minimal I-Maps

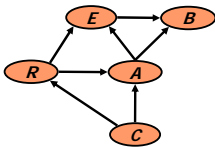
The factorization theorem suggests an algorithm

- Fix an ordering X_1, \dots, X_n
- For each i ,
 - select Pa_i to be a minimal subset of $\{X_1, \dots, X_{i-1}\}$, such that $I(X_i; \{X_1, \dots, X_{i-1}\} - Pa_i | Pa_i)$
- Clearly, the resulting graph is a minimal I-Map.

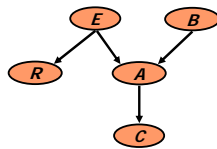
Non-uniqueness of minimal I-Map

- Unfortunately, there may be several minimal I-Maps for the same distribution
 - Applying I-Map construction procedure with different orders can lead to different structures

Order: C, R, A, E, B



Original I-Map



Choosing Ordering & Causality

- The choice of order can have drastic impact on the complexity of minimal I-Map
- Heuristic argument: construct I-Map using **causal** ordering among variables
- Justification?
 - It is often reasonable to assume that graphs of causal influence should satisfy the Markov properties.

P-Maps

- A DAG G is P-Map (**perfect map**) of a distribution P if
 - $I(X; Y | Z)$ if and only if $d\text{-sep}(X; Y | Z, G) = \text{yes}$

Notes:

- A P-Map captures all the independencies in the distribution
- P-Maps are unique, up to DAG equivalence

Unfortunately, some distributions do not have a P-Map ☹

Bayesian Networks

- A Bayesian network specifies a probability distribution via two components:
 - A DAG G
 - A collection of conditional probability distributions $P(X_i | Pa_i)$
- The joint distribution P is defined by the factorization

$$P(X_1, \dots, X_n) = \prod_i P(X_i | Pa_i)$$
- Additional requirement: G is a minimal I-Map of P

DAG Summary

- We explored DAGs as a representation of conditional independencies:
 - Markov independencies of a DAG
 - Tight correspondence between $Markov(G)$ and the factorization defined by G
 - d-separation, a sound & complete procedure for computing the consequences of the independencies
 - Notion of minimal I-Map
 - P-Maps
- This theory is the basis for defining Bayesian networks

CPDs

- So far, we focused on how to represent independencies using DAGs
- The "other" component of a Bayesian networks is the specification of the **conditional probability distributions** (CPDs)
- We start with the simplest representation of CPDs and then discuss additional structure

Tabular CPDs

- When the variables of interest are all discrete, the common representation is as a table:
- For example $P(C|A,B)$ can be represented by

A	B	$P(C=0 A, B)$	$P(C=1 A, B)$
0	0	0.25	0.75
0	1	0.50	0.50
1	0	0.12	0.88
1	1	0.33	0.67

Tabular CPDs

Pros:

- Very flexible, can capture any CPD of discrete variables
- Can be easily stored and manipulated

Cons:

- Representation size grows exponentially with the number of parents!
- Unwieldy to assess probabilities for more than few parents

Structured CPD

- To avoid the exponential blowup in representation, we need to focus on specialized types of CPDs
- This comes at a cost in terms of expressive power
- There are several types of structured CPDs
 - See BNB ch. 4 for details (not required)
 - Noisy-Or
 - Mixed Continuous and Discrete

Causal Independence

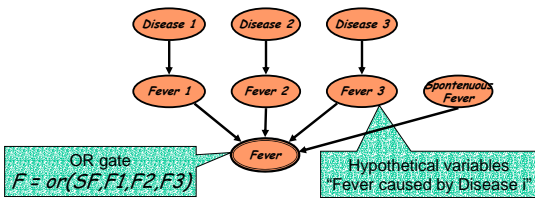
- Consider the following situation



- In tabular CPD, we need to assess the probability of fever in eight cases
- These involve all possible interactions between diseases
- For three disease, this might be feasible... For ten diseases, not likely...

Causal Independence

- Simplifying assumption:
 - Each disease attempts to cause fever, **independently** of the other diseases
 - The patient has fever if one of the diseases "succeeds"
- We can model this using a Bayesian network fragment



Noisy-Or CPD

- Models $P(X|Y_1, \dots, Y_k)$, X, Y_1, \dots, Y_k are all binary
- Parameters:
 - p_i -- probability of $X=1$ due to $Y_i=1$
 - p_0 -- probability of $X=1$ due to other causes
- Plugging these in the model we get

$$P(X=0 | Y_1, \dots, Y_k) = (1-p_0) \prod_i (1-p_i)^{Y_i}$$

$$P(X=1 | Y_1, \dots, Y_k) = 1 - P(X=0 | Y_1, \dots, Y_k)$$

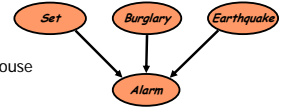
Noisy-or CPD

- Benefits of noisy-or
 - "Reasonable" assumptions in many domains
 - e.g., medical domain
 - Few parameters.
 - Each parameter can be estimated independently of the others
- The same idea can be extended to other functions: noisy-max, noisy-and, etc.
- Frequently used in large medical expert systems

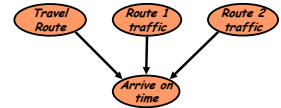
Context Specific Independence

- Consider the following examples:

- Alarm sound depends on
 - Whether the alarm was set before leaving the house
 - Burglary
 - Earthquake



- Arriving on time depends on
 - Travel route
 - The congestion on the two possible routes

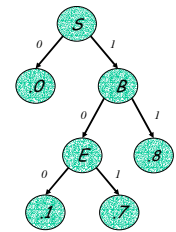


Context-Specific Independence

- In both of these example we have **context-specific independencies (CSI)**
 - Independencies that depends on a particular value of one or more variables
- In our examples:
 - $Ind(A : B, E | S = 0)$
Alarm sound is independent of B and E when the alarm is not set
 - $Ind(A : R_2 | T = 1)$
Arrival time is independent of traffic on route 2 if we choose to travel on route 1

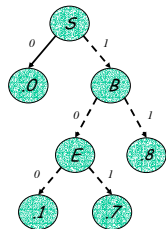
Representing CSI

- When we have such CSI, $P(X / Y_1, \dots, Y_k)$ is the same for several values of Y_1, \dots, Y_k
- There are many ways of representing these regularities
- A natural representation: decision trees
 - Internal nodes: tests on parents
 - Leaves: probability distributions on X
- Evaluate $P(X / Y_1, \dots, Y_k)$ by traversing tree



Detecting CSI

- Given evidence on some nodes, we can identify the "relevant" parts of the trees
 - This consists of the paths in the tree that are consistent with context
- Example
 - Context $S = 0$
 - Only one path of tree is relevant
- A parent is independent given the context if it does not appear on one of the relevant paths



Decision Tree CPDs

Benefits

- Decision trees offer a flexible and intuitive language to represent CSI
- Incorporated into several commercial tools for constructing Bayesian networks

Comparison to noisy-or

- Noisy-or CPDs require full trees to represent
- General decision tree CPDs cannot be represented by noisy-or

Mixed Discrete and Continuous

Conditional Gaussian CPDs

- A model for networks that combine discrete and continuous variables
- If X is continuous
 - Y_1, \dots, Y_k are continuous
 - Z_1, \dots, Z_l are discrete

Conditional Gaussian (CG) CPD:

- For each joint value of Z_1, \dots, Z_l define a different linear-Gaussian parameters
- Resulting multivariate distribution: mixture of multivariate Gaussians
 - Each assignment of values to discrete variables selects a multivariate Gaussian over continuous variables

CPD Summary

- Many choices for representing CPDs
- Any "statistical" model of conditional distribution can be used
 - e.g., any regression model
- Representing structure in CPDs can have implications on independencies among variables