Overview of Probabilistic Graphical Models
Representation: Part I

Aditya Ramesh
_@adityaramesh.com

John Cavazos Lab
University of Delaware
Newark, Delaware 19716

This presentation is based **heavily** on the material by Daphne Koller ([1]), Nir Friedman ([1], [3]), and David Sontag ([2]). I have **directly copied** or otherwise incorporated parts of their work into many of these slides, citing the sources where appropriate. If you find this presentation useful, I highly recommend that you take some time to read their work.

Outline

Event Spaces

- We can formalize the notion of an *event* by defining a space of possible outcomes Ω.
- Further, we define an *event space* S so that we can attach probabilities to specific outcomes.
- The event space S must satisfy the following:
  - $\emptyset, \Omega \in S$, where $\emptyset$ is the *empty event* and Ω is the *trivial event*.
  - Closure under union: $\alpha, \beta \in S \rightarrow \alpha \cup \beta \in S$.
  - Closure under complementation: $\alpha \in S \rightarrow \Omega - \alpha \in S$.

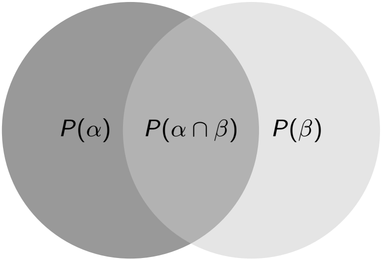
Probability Distributions

- A probability distribution over (Ω, S) is a mapping $P : S \to \mathbb{R}_0^+$. It must satisfy the following:
  - $P(\Omega) = 1$ — that is, the trivial event is given maximal probability.
  - If $\alpha, \beta \in S$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$. If two outcomes are mutually disjoint, the probability of their union is the sum of their probabilities.
  - In this context, we view probabilities as *subjective degrees of belief* rather than as frequencies.

Conditional Probabilities

- How are our beliefs about an outcome affected in light of new evidence?
- This notion can be formalized using the following definition, which relates the areas of the shaded regions in the diagram.

$$P(\beta \mid \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$$



Chain Rule

- From the definition of the conditional distribution, we have

$$P(\alpha \cap \beta) = P(\alpha)P(\beta \mid \alpha).$$

- More generally,

$$P(\alpha_1 \cap \ldots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 \mid \alpha_1) \cdots P(\alpha_k \mid \alpha_1 \cap \ldots \cap \alpha_{k-1}),$$

  but the order in which we choose to pull out variables does not matter.
- The fact that the chain rule allows us to decompose a joint distribution into *factors over smaller subsets of variables* becomes crucial later on.

Bayes Rule

- Decomposing the joint distribution in the defintion of the conditional probability using the chain rule gives us Bayes rule:

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}.$$

- Swapping an event on the left side of the pipe symbol with an event on the right works similarly when we are conditioning on several events:

$$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}.$$

Random Variables

- A random variable associates each outcome in Ω with a value in a discrete or continuous set, e.g. the assignment $Grade = A$ is shorthand for the event $\{\omega \in \Omega : f_{Grade}(\omega) = A\}$.
- Instead of one random variable, we can also have a random *vector* $X(\omega) = \{X_1(\omega), \ldots, X_n(\omega)\}$. Conditioning, the chain rule, and Bayes rule all apply [2].
- When dealing with categorical variables, we use x^i to denote the assignment of x to the ith state of X, where $i \in [1, |Val(X)|]$.

Independence and Conditional Independence

- A distribution P satisfies $(\alpha \perp \beta)$ if and only if $P(\alpha \cap \beta) = P(\alpha)P(\beta)$.
- A distribution P satisfies $(\alpha \perp \beta \mid \gamma)$ if and only if $P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma)P(\beta \mid \gamma)$.
- Note the relationship between independence of variables and factorization of the joint distribution — it will play a very prominent role later on.

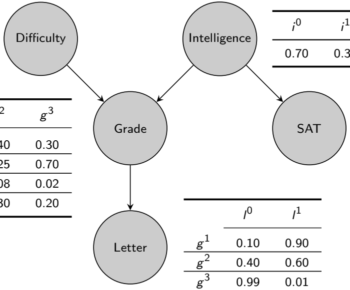
What is a Bayesian Network?

- A Bayesian Network is a directed acyclic graph (DAG) that we use to encode the factorization and (equivalently) the independence assertions of a joint distribution P.
- The nodes of the DAG are the random variables in our domain, and there is one CPD defined per node.

The Student Model

d^0	d^1
0.60	0.40

i^0	i^1
0.70	0.30



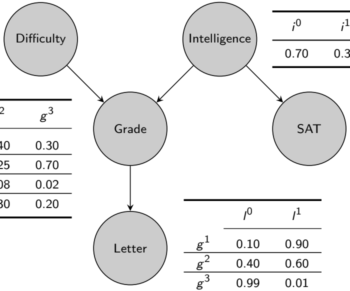
	g^1	g^2	g^3
i^0, d^0	0.30	0.40	0.30
i^0, d^1	0.05	0.25	0.70
i^1, d^0	0.90	0.08	0.02
i^1, d^1	0.50	0.30	0.20

	s^0	s^1
i^0	0.95	0.05
i^1	0.20	0.80

	l^0	l^1
g^1	0.10	0.90
g^2	0.40	0.60
g^3	0.99	0.01

- In the student model, $Val(D) = \{easy, hard\}$, $Val(I) = \{low, high\}$, $Val(G) = \{A, B, C\}$, $Val(S) = \{low, high\}$, and $Val(L) = \{weak, strong\}$.
- The joint distribution is $P(i, d, g, s, l) = P(i)P(d)P(g \mid i, d)P(s \mid i)P(l \mid g)$.

The Student Model



d^0	d^1
0.60	0.40

i^0	i^1
0.70	0.30

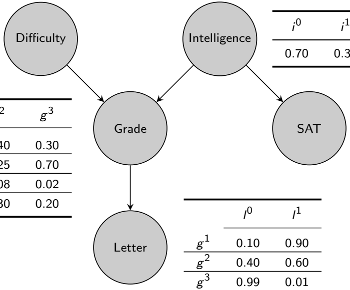
	g^1	g^2	g^3
i^0, d^0	0.30	0.40	0.30
i^0, d^1	0.05	0.25	0.70
i^1, d^0	0.90	0.08	0.02
i^1, d^1	0.50	0.30	0.20

	s^0	s^1
i^0	0.95	0.05
i^1	0.20	0.80

	l^0	l^1
g^1	0.10	0.90
g^2	0.40	0.60
g^3	0.99	0.01

- Suppose that initially, we only have a student's recommendation letter (which is weak) and transcript (indicating that he received a 'C' in the course). How does finding that he received a high SAT score affect our beliefs about his intelligence?

The Student Model



d^0	d^1
0.60	0.40

i^0	i^1
0.70	0.30

	g^1	g^2	g^3
i^0, d^0	0.30	0.40	0.30
i^0, d^1	0.05	0.25	0.70
i^1, d^0	0.90	0.08	0.02
i^1, d^1	0.50	0.30	0.20

	s^0	s^1
i^0	0.95	0.05
i^1	0.20	0.80

	l^0	l^1
g^1	0.10	0.90
g^2	0.40	0.60
g^3	0.99	0.01

- What is the largest set W such that
  - $(L \perp W \mid G)$?
  - $(S \perp W \mid I)$?
  - $(G \perp W \mid \pi(G))$, where $\pi(G)$ returns the set of parents of G?
  - $(D \perp W)$?
  - $(D \perp W \mid L)$?

Conclusions from the Student Model

- We view a Bayesian network as a generative model in which we sample the variables in topological order [2].
- The parents can be thought of as "generating" their children — that is, once we know about the values of the parents, the child is "shielded" from probabilistic influence by non-descendants.
- Consequently, we can think of a Bayesian network as representing a set of conditional independence assertions.

I-Maps

- A Bayesian network structure G encodes the following set of conditional independence assertions $I_\ell(G)$, called the *local independencies*:

$$\forall X_i \in Val(X) : (X_i \perp \text{NonDescendants}_{X_i} \mid \pi(X_i)).$$

- We define $I(P)$ to be the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in P. Obviously $I_\ell(G) \subseteq I(P)$.

I-Maps

- Let K be any graph object associated with a set of independencies $I(K)$. We say that K is an I-map for a set of independencies I if $I(K) \subseteq I$.
- That is, all the conditional independence assertions that hold in $I(K)$ also hold in I.

I-Maps and Factorization

- Given that G is an I-map for P, can we simplify the representation of P [3]?
- Applying $I_\ell(G)$ to each factor in the naive decomposition proves that if G is an I-map for P, then P factorizes according to G. That is,

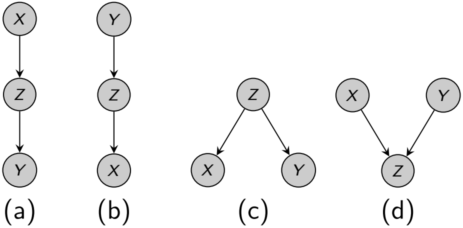
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i)).$$

- The converse is also true.
- The compact factorization can result in an exponential reduction in the number of parameters that need to be specified!

I-Maps and Factorization

- Can we go the other way around and recover the graph G given the factorization of the distribution P?
- We will first take a detour and explore more deeply how probabilistic influence flows across a graph.

Path Blockage [3]



The four possible edge trail combinations from X to Y via Z: (a) An indirect causal effect; (b) An indirect evidential effect; (c) A common cause; (d) A common effect.

- Edge trails (a)–(c) are active if and only if Z is not observed.
- Edge trail (d) is active if and only if either Z or one of Z's descendants is observed. Intuitively, this can be understood as a consequence of intercausual reasoning.

Path Blockage

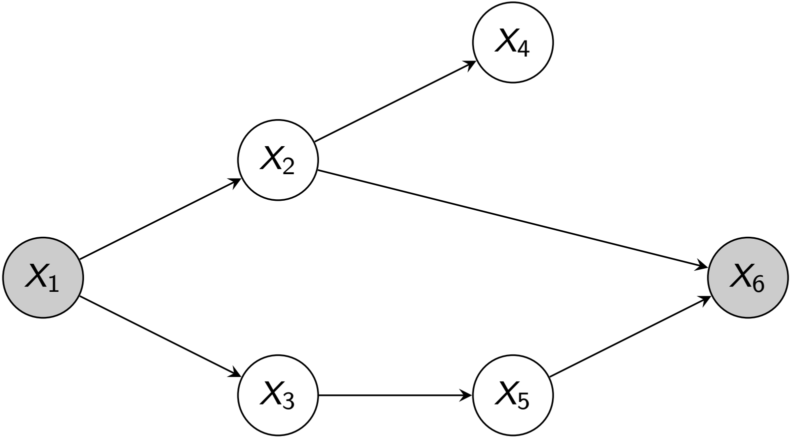
- Now we can extend our finding about three-node edge-trails to trails of arbitrary length.
- The trail $X_1 \leftrightharpoons \cdots \leftrightharpoons X_n$ is active given a subset of observed variables Z if
    - Whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants are in Z.
    - No other node along the trail is in Z.

D-Separation

- We say that X and Y are d-separated ("directed-separated") given Z, denoted $\text{d-sep}_G(X; Y \mid Z)$, if there is no active trail between any node $X \in X$ and $Y \in Y$ given Z.
- We use $I(G)$ to denote the set of independencies (aka *global Markov independencies*) that correspond to d-separation:

$$I(G) = \{(X \perp Y \mid Z) : \text{d-sep}_G(X; Y \mid Z)\}.$$

D-Separation [2]



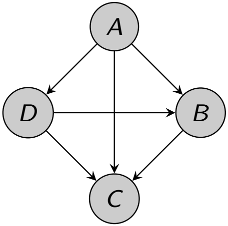
Independence and D-Separation

- We now return to make the connection between the factorization of P and the construction of G.
- For any distribution G that factorizes over G, we have that we have that if X and Y are not d-separated given Z in G, then X and Y are dependent in all distributions P that factorize over G.
- Otherwise, G would contain some independence assertion that is not reflected in P, a contradiction.

Independence and D-Separation

- For almost all distributions P that factorize over G, that is, for all distributions except for a set of measure zero in the space of CPD parameterizations, we have that $I(P) = I(G)$.
- D-separation reduces statistical independencies in graphs (hard) to connectivity in graphs (easy) [2].

Minimal I-Maps [3]



- By itself, the concept of an I-map is not sufficient for us to get very far.
- The graph depicted above (a *complete* DAG) is an I-map for *any* distribution P.

Minimal I-Maps [3]

- A graph K is a minimal I-map for a set of independence assertions I if K is an I-map for I, and the removal of even a single edge from K renders it not an I-map.
- Taking $I = I(P)$ or $I = I(K')$, we can talk about K as being a minimal I-map for a distribution or another graph.

Algorithm for Constructing a Minimal I-Map

Require: An ordering $X_1, \ldots, X_n$ of random variables in $\mathcal{X}$
Require: A set of independencies I

Algorithm for Constructing a Minimal I-Map

Require: An ordering $X_1, \ldots, X_n$ of random variables in $\mathcal{X}$
Require: A set of independencies I
Set G to an empty graph over $\mathcal{X}$

Algorithm for Constructing a Minimal I-Map

Require: An ordering $X_1, \dots, X_n$ of random variables in $\mathcal{X}$
Require: A set of independencies I
Set G to an empty graph over $\mathcal{X}$
for $i = 1, \dots, n$ **do**
U is the current candidate for parents of X_i
$U \leftarrow \{X_1, \dots, X_{i-1}\}$

Algorithm for Constructing a Minimal I-Map

Require: An ordering $X_1, \ldots, X_n$ of random variables in $\mathcal{X}$
Require: A set of independencies I
  Set G to an empty graph over $\mathcal{X}$
  for $i = 1, \ldots, n$ **do**
    U is the current candidate for parents of X_i
    $U \leftarrow \{X_1, \ldots, X_{i-1}\}$
    Find the minimal set U satisfying
    $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
    for $U' \subseteq \{X_1, \ldots, X_{i-1}\}$ **do**
      if $U' \subset U$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in I$ **then**
        $U \leftarrow U'$

Algorithm for Constructing a Minimal I-Map

Require: An ordering $X_1, \ldots, X_n$ of random variables in $\mathcal{X}$
Require: A set of independencies I

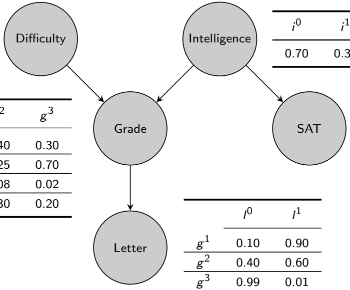
Set G to an empty graph over $\mathcal{X}$
for $i = 1, \ldots, n$ **do**
  U is the current candidate for parents of X_i
  $U \leftarrow \{X_1, \ldots, X_{i-1}\}$
  Find the minimal set U satisfying
  $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
  for $U' \subseteq \{X_1, \ldots, X_{i-1}\}$ **do**
    if $U' \subset U$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in I$ **then**
      $U \leftarrow U'$
  Now set U to be the parents of X_i
  for $X_j \in U$ **do**
    Add X_j—X_i to G

Algorithm for Constructing a Minimal I-Map

Require: An ordering $X_1, \ldots, X_n$ of random variables in $\mathcal{X}$
Require: A set of independencies I

Set G to an empty graph over $\mathcal{X}$
for $i = 1, \ldots, n$ **do**
  U is the current candidate for parents of X_i
  $U \leftarrow \{X_1, \ldots, X_{i-1}\}$
  Find the minimal set U satisfying
  $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
  for $U' \subseteq \{X_1, \ldots, X_{i-1}\}$ **do**
    if $U' \subset U$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in I$ **then**
      $U \leftarrow U'$
  Now set U to be the parents of X_i
  for $X_i \in U$ **do**
    Add X_j—X_i to G
  return G

Constructing a Minimal I-Map



d^0	d^1
0.60	0.40

i^0	i^1
0.70	0.30

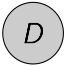
	g^1	g^2	g^3
i^0, d^0	0.30	0.40	0.30
i^0, d^1	0.05	0.25	0.70
i^1, d^0	0.90	0.08	0.02
i^1, d^1	0.50	0.30	0.20

	s^0	s^1
i^0	0.95	0.05
i^1	0.20	0.80

	l^0	l^1
g^1	0.10	0.90
g^2	0.40	0.60
g^3	0.99	0.01

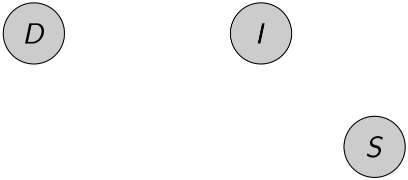
- We now apply the algorithm to the variables in the student model, listed in topological order: D, I, S, G, L.

Constructing a Minimal I-Map

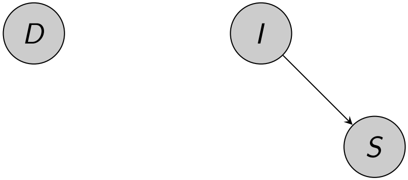


Constructing a Minimal I-Map

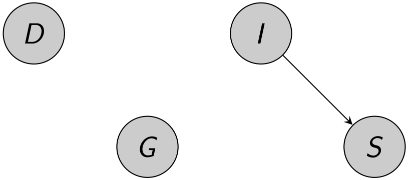
Constructing a Minimal I-Map



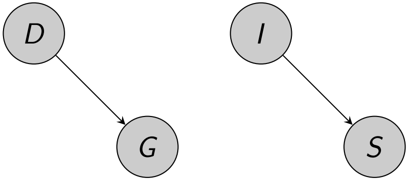
Constructing a Minimal I-Map



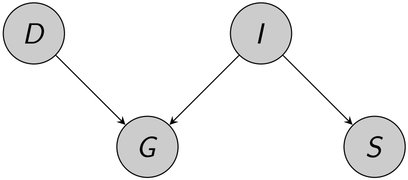
Constructing a Minimal I-Map



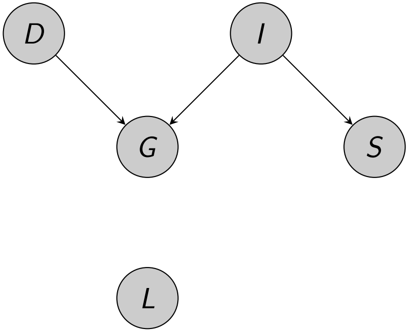
Constructing a Minimal I-Map



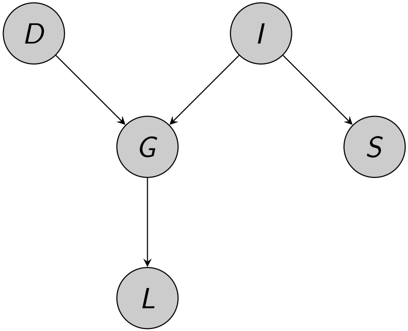
Constructing a Minimal I-Map



Constructing a Minimal I-Map



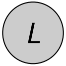
Constructing a Minimal I-Map



Constructing a Minimal I-Map

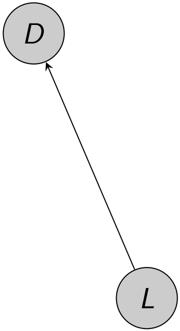
- If we run the algorithm with an ordering that is topological for G, then the algorithm returns G.
- This is because the set of parents that are considered for each X_i is precisely $\pi(X_i)$.
- Now, we consider a less natural ordering: L, D, S, I, G.

Constructing a Minimal I-Map

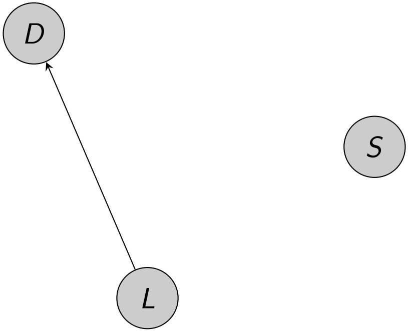


Constructing a Minimal I-Map

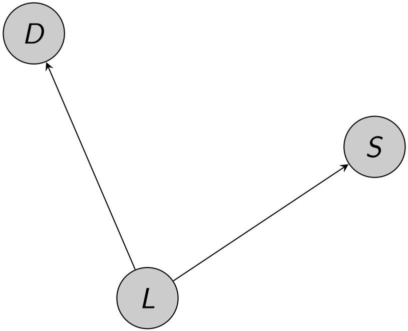
Constructing a Minimal I-Map



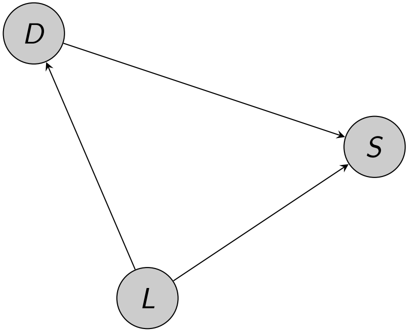
Constructing a Minimal I-Map



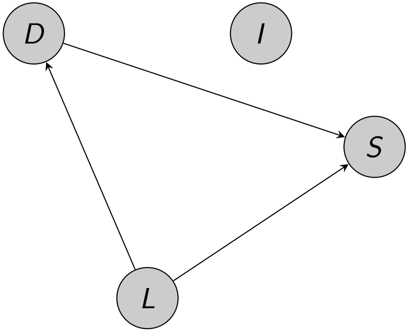
Constructing a Minimal I-Map



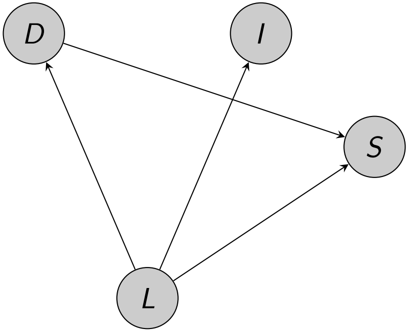
Constructing a Minimal I-Map



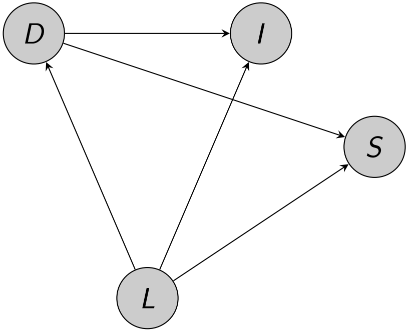
Constructing a Minimal I-Map



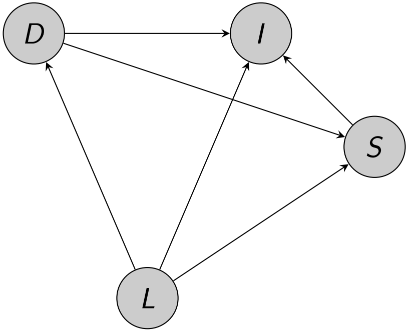
Constructing a Minimal I-Map



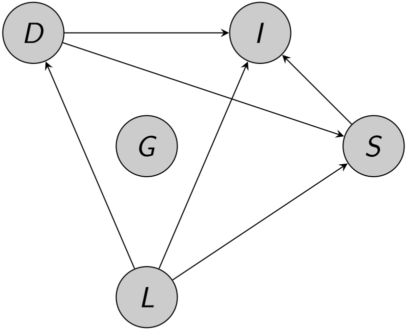
Constructing a Minimal I-Map



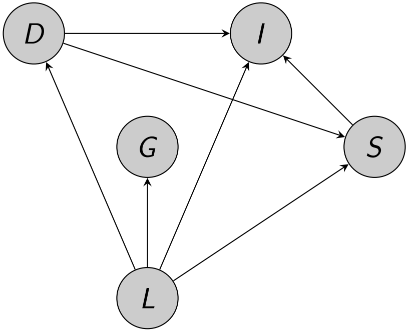
Constructing a Minimal I-Map



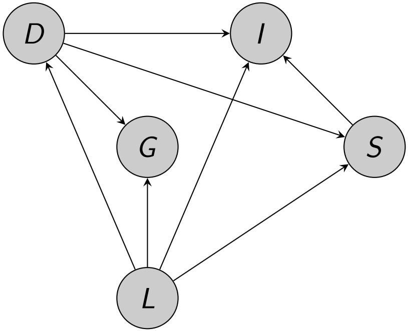
Constructing a Minimal I-Map



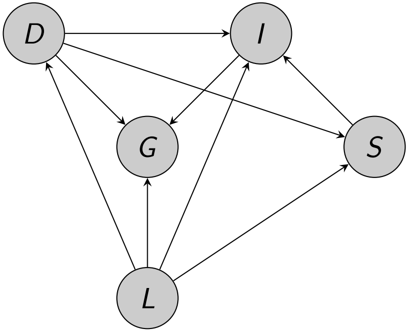
Constructing a Minimal I-Map



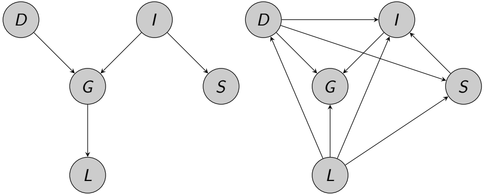
Constructing a Minimal I-Map



Constructing a Minimal I-Map



Construction of Minimal I-Maps



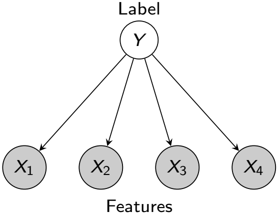
(a) (b)

- Both graphs (a) and (b) are valid I-maps for G, but have completely different edge configurations.
- Ironically, we cannot "read off" the independence assertions from a minimal I-map. Even minimal I-maps fail to capture some or all of the independencies that hold in the distribution.

Constructing Minimal I-Maps

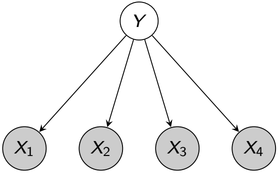
- A more restrictive definition called the perfect map (P-map) captures the independencies in a given distribution P.
- We say that a graph K is a P-map for P if $I(K) = I(P)$.
- Through an involved process, it is possible to construct a PDAG (partially directed acyclic graph) from a distribution P that encodes all P-maps in the I-equivalence class of P.
- Unfortunately, not all distributions have P-maps.

Email Classification [2]



- We now shift our focus to a few applications of Bayesian networks.
- To generate an email, recall that we can sample the variables in the Bayesian network in topological order.
  - First, we sample $y \sim P(Y)$ to decide whether or not the email is spam.
  - Then, $\forall i \in [1, n]$ sample $x_i \sim P(X_i \mid Y = y)$.

Email Classification [2]



- To determine whether an email is spam given the features X_i, we need to compute

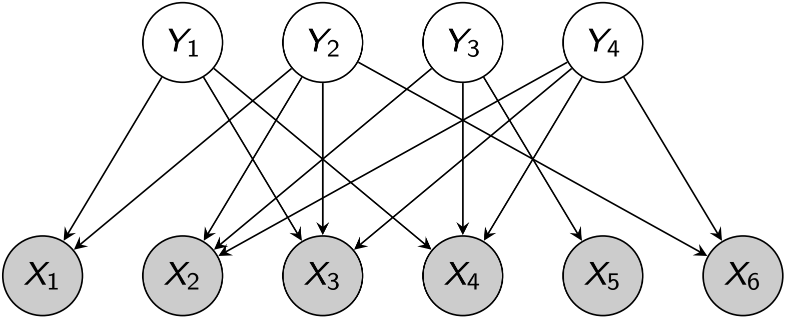
$$\begin{aligned} P(Y \mid X_i) &= \frac{P(Y, X_i)}{P(X_i)} \\ &= \frac{P(Y)\prod_{i=1}^{n} P(Y \mid X_i)}{\sum_{y \in Y} P(Y, X_i)} \\ &= \frac{P(Y)\prod_{i=1}^{n} P(Y \mid X_i)}{\sum_{y \in Y} P(Y)\prod_{i=1}^{n} P(Y \mid X_i)}. \end{aligned}$$

Email Classification

- We can view every email in our corpus as a bag of words and let the features X_i be functions of their counts (e.g. "Nigeria", "bank").
- The local probability models (in our case $P(Y)$ and $P(X_i \mid Y)\,\forall X_i$) are now easily computed, and we have designed a primitive email spam-detection system.
- What is the name of this classifier?

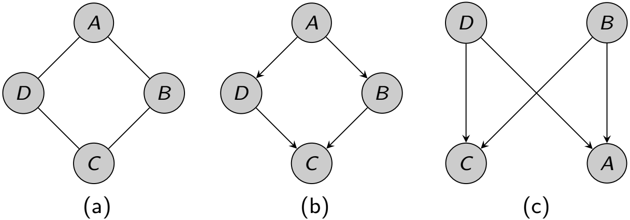
Naive Bayes [2]

Diseases (e.g. "pneumonia", "flu", "common cold")



Findings (e.g. "cough", "fever", "fast breathing")

Motivation



A professor misspeaks in class, causing four students to have a misconception. This model predicts whether or not each of them continues to carry the misconception after the students meet in pairs (Alice and Bob, Bob and Charles, Charles and Debbie, and Debbie and Alice).

- We want to encode $(A \perp C \mid D, B)$ and $(B \perp D \mid A, C)$. Why do (b) and (c) fail to capture these independencies?
- The fact that the a CPD involving a variable in a Bayesian network can only be over itself and its parents poses limitations.

What is a Markov Network?

- Similarly to Bayesian networks, a Markov network, or Markov random field (MRF) is an undirected graphical model with one node per variable.
- Unlike Bayesian networks, the non-negative potential functions (or *factors*) are associated with cliques of variables.

$$P(X_1, \ldots, X_n) = \frac{1}{Z}\tilde{P}(X_1, \ldots, X_n),$$

where the normalizing constant (or *partition function*) Z is defined as

$$Z = \sum_{X_1, \ldots, X_n} \tilde{P}(X_1, \ldots, X_n).$$

- A distribution of this kind is called a *Gibbs distribution*.

What is a Markov Network?

- Unlike Bayesian networks, the factors do not have to be normalized, so $\tilde{P}$ does not have to be normalized.
- We define the the *factor product* over some $\phi_1(X, Y)$ and $\phi_2(Y, Z)$, $\phi_1 \times \phi_2$, by

$$\psi(X, Y, Z) = \phi_1(X, Y) \cdot \phi_2(Y, Z).$$

- The notation $\phi_1(X, Y) \cdot \phi_2(Y, Z)$ is deceptively concise. For discrete distributions, $\psi(X, Y, Z)$ is a tensor of order $|X| + |Y| + |Z|$, each of whose elements

$$\psi(x, y, z) = \phi_1(x, y) \cdot \phi_2(y, z)$$

corresponds to a unique assignment $x \in X, y \in Y$, and $z \in Z$.

Factor Product

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

a^1	b^1	c^1	$0.50 \cdot 0.50 = 0.25$
a^1	b^1	c^2	$0.50 \cdot 0.70 = 0.35$
a^1	b^2	c^1	$0.80 \cdot 0.10 = 0.08$
a^1	b^2	c^2	$0.80 \cdot 0.20 = 0.16$
a^2	b^1	c^1	$0.10 \cdot 0.50 = 0.05$
a^2	b^1	c^2	$0.10 \cdot 0.70 = 0.07$
a^2	b^2	c^1	$0.00 \cdot 0.10 = 0.00$
a^2	b^2	c^2	$0.00 \cdot 0.20 = 0.10$
a^3	b^1	c^1	$0.30 \cdot 0.50 = 0.15$
a^3	b^1	c^2	$0.30 \cdot 0.70 = 0.21$
a^3	b^2	c^1	$0.90 \cdot 0.10 = 0.09$
a^3	b^2	c^2	$0.90 \cdot 0.20 = 0.18$

$$\phi(A, B) \times \phi(B, C) = \phi(A, B, C)$$

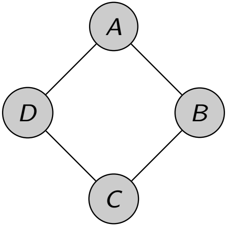
What is a Markov Network?

- Note that computing the factor product over two discrete distributions can quickly become computationally expensive as the number of variables involved increases.
- Now we can define $\tilde{P}$ given a set of factors $\mathbf{\Phi} = \{\phi_1(D_1), \ldots, \phi_m(D_m)\}$ as

$$\tilde{P}(X_1, \ldots, X_n) = \phi_1(D_1) \times \cdots \times \phi_m(D_m),$$

  where each D_i is a clique in the Markov network.
- Evaluating an *n*-ary factor product pairwise will result in the generation of intermediate temporaries. If performance is a concern, we can avoid doing this by implementing a variadic factor product function.

The Misconception Model [2]



- We introduce single-node potentials $\phi_A, \phi_B, \phi_C, \phi_D$ to represent probabilities that individuals correctly work out the misconceptions themselves.
- We introduce pairwise potentials $\phi_{AB}, \phi_{BC}, \phi_{CD}, \phi_{DA}$ to model whether partners agree after their meeting.

The Misconception Model [2]

- The joint distribution is

$$P(A,B,C,D) = \frac{1}{Z}\phi_A(A)\phi_B(B)\phi_C(C)\phi_D(D)$$
$$\phi_{AB}(A,B)\phi_{BC}(B,C)\phi_{CD}(C,D)\phi_{DA}(D,A).$$

- Entries in the potentials need not be normalized; we can only judge relative importance by normalizing (scaling by Z) before comparison.

Independence [2]

- Consider a Markov network A—B—C with the joint distribution
$$P(A, B, C) = \frac{1}{Z}\phi_{AB}(A, B)\phi_{BC}(B, C).$$
- We show that $P(a \mid b)$ can be computed using only $\phi_{AB}(a, b)$:
$$\begin{aligned} P(a \mid b) &= \frac{P(a, b)}{P(b)} \\ &= \frac{\frac{1}{Z}\sum_{c'} \phi_{AB}(a, b)\phi_{BC}(b, c')}{\frac{1}{Z}\sum_{a', c'} \phi_{AB}(a', b)\phi_{BC}(b, c')} \\ &= \frac{\phi_{AB}(a, b)\sum_{c'} \phi_{BC}(b, c')}{\sum_{a'} \phi_{AB}(a', b)\sum_{c'} \phi_{BC}(b, c')} \\ &= \frac{\phi_{AB}(a, b)}{\sum_{a'} \phi_{AB}(a', b)}. \end{aligned}$$

Independence [2]

- The probability of a variable conditioned on its neighbors depends *only* on the potentials involving that node.
- Observing even a single variable in a path between two variables impedes the flow of probabilistic influence, causing that path to become inactive.
- This means that a path $X_1 — \cdots — X_n$ is active given a subset of observed variables Z if and only if none of the X_i's is in Z.

I-Maps

- We say that a set of nodes Z separates X and Y in H, denoted $\mathrm{sep}_H(X; Y \mid Z)$, if there is no active path between any node $X \in X$ and $Y \in Y$ given Z. We define the global independencies associated with H to be:

$$I(H) = \{(X \perp Y \mid Z) : \mathrm{sep}_H(X; Y \mid Z)\}.$$

- If $\neg\mathrm{sep}_H(X; Y \mid Z)$, then X and Y are dependent given Z in some distribution P that factorizes over H.

Pairwise Independencies

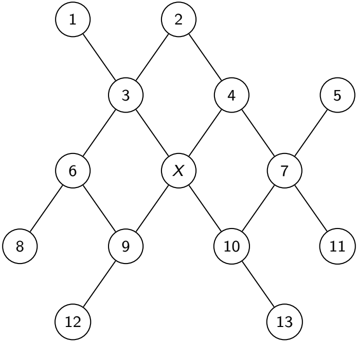
- When two variables are not directly connected, there *must* be some way of rendering them independent.
- We define the *pairwise independencies* associated with a Markov network H over a set of variables $\mathcal{X}$ to be

$$I_p(H) = \{(X \perp Y \mid \mathcal{X} - \{X, Y\}) : X—Y \notin H\}.$$

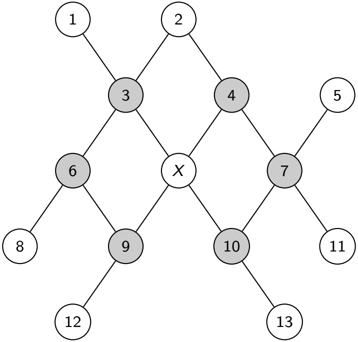
Markov Blanket

- What is the minimal set of variables U that need to be observed in a graph G over a set of variables $\mathcal{X}$ in order for a given node X to be independent of $\mathcal{X} - \{X\} - U$?
- This is called the *Markov blanket* of X in G, denoted $\mathrm{MB}_G(X)$.

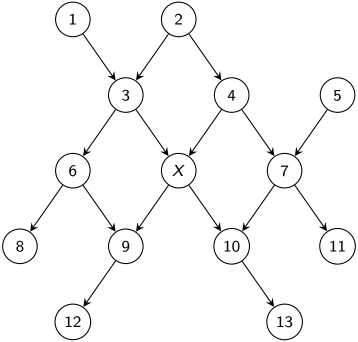
Markov Blanket: Undirected Graph



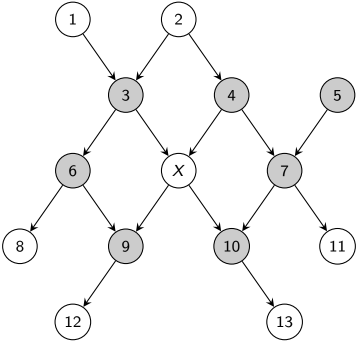
Markov Blanket: Undirected Graph



Markov Blanket: Directed Graph



Markov Blanket: Directed Graph



Equivalence of Local and Global Independencies

- We can now formulate a definition for $I_\ell(H)$ in terms of the Markov blankets of each $X \in H$:

$$I_\ell(H) = \{(X \perp \mathcal{X} - \{X\} - \mathrm{MB}_H(X) \mid \mathrm{MB}_H(X)) : X — Y \notin H\}.$$

- The following three statements are equivalent for a positive distribution P:
  1. $P \models I_\ell(H)$.
  2. $P \models I_p(H)$.
  3. $P \models I(H)$.

Distributions to Graphs

- Given a distribution P, can we go the other way around and construct a graph H such that H is a minimal I-map for P?
- We can satisfy $I(H)$ by visiting each node $X \in H$ and ensuring that the $I_\ell(H)$ is satisfied by observing all $Y \in \mathrm{MB}_H(X)$.

Distributions to Graphs

- By construction, H satisfies $I_\ell(H)$ and thus is an I-map for H.
- Assume that H is not a minimal I-map. Removing an edge in H results in some assertion $(X \perp Y \mid \mathcal{X} - \{X, Y\})$ not being reflected in P, violating $I_p(H)$.
- Hence, H is a minimal I-map.

Hammersley-Clifford Theorem

- Let P be a positive distribution over $\mathcal{X}$, and H a Markov network graph over $\mathcal{X}$. If H is an I-map for P, then P is a Gibbs distribution over H.
- For positive distributions, *all four conditions* — factorization and the three types of Markov assumptions — are all equivalent.

Log-Linear Models

- Instead of encoding factors as complete tables over their scope, we can use an alternate parameterization that can make context-specific structure more apparent.
- We can rewrite a factor $\phi(D)$ as

$$\phi(D) = \exp(-\epsilon(D)),$$

where $\phi(D) = -\ln\phi(D)$.

- The logarithmic representation ensures that the distribution is positive.

Log-Linear Models

- A distribution P is a log-linear model over a Markov network H if it is associated with
    - a set of features $F = \{f_1(D_1), \ldots, f_k(D_k)\}$, where each D_i is a complete subgraph in H, and
    - a set of weights $w_1, \ldots, w_k$,

  such that

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \exp \left[- \sum_{i=1}^{k} w_i f_i(D_i) \right].$$

Log-Linear Models: The Ising Model

- The Ising model is a model for the energy of a particular system involving a system of interacting atoms.
- Each atom is associated with a binary variable $X_i = \{+1, -1\}$, whose value defines the direction of the atom's spin.
- The energy function associated with the edges is defined by

$$\epsilon_{ij}(x_i, x_j) = w_{i,j} x_i x_j.$$

- When two atoms X_i, X_j have the same spin, they make a contribution $w_{i,j}$; otherwise, they make a contribution $-w_{i,j}$.
- We also include single node potentials u_i that bias a particular atom to have one spin or another.

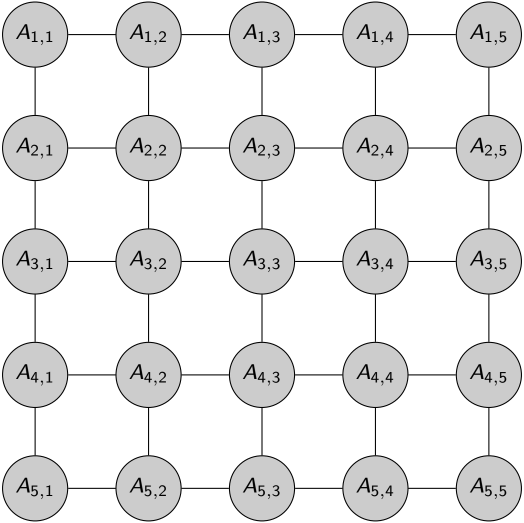
Log-Linear Models: The Ising Model

- The energy distribution is given by the definition of the log-linear model:

$$P(\xi) = \frac{1}{Z}\left(-\sum_{i<j} w_{i,j}x_i x_j - \sum_i u_i x_i\right).$$

- When $w_{i,j} > 0$ the model prefers to align the spin of the two atoms, and the interaction is called *ferromagnetic*.
- When $w_{i,j} < 0$, the interaction is called *antiferromagnetic*.
- When $w_{i,j} = 0$, the atoms are non-interacting.
- Also related is the Boltzmann machine, which has gained wide popularity in deep learning.

Log-Linear Models: The Ising Model



Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 1st Edition, 2009.

David Sontag, lecture slides on Probabilistic Graphical Models. Accessible at `http://cs.nyu.edu/~dsontag/courses/pgm12/`.

Nir Friedman, lecture slides on the theory of Bayesian networks. Accessible at `classes.soe.ucsc.edu/cmps290c/Spring04/paps/nir2.pdf`.