Lecture 6: February 2, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Review

- Assignment #2 (a) posted; 2 (b) to be posted next week; both due Feb 9.
- Previous lecture:
  - Context specific independences
    - Tree CPDs, Multiplexer nodes
    - Deterministic Nodes
  - Noisy-OR
  - Generalized linear models
    - Apply to continuous valued parents of discrete nodes
  - Continuous variables (very briefly)
  - Conditional BNs
- Today's objective
  - Undirected Graphs

Undirected Models (Markov Networks)

- Like Bayesian Networks but edges do not have directions (no parent/child or cause/effect type of relationships)
- Instead of CPDs, we specify "factors" (details to come soon)
- In some problem domains, there are no obvious causal relationships between the variables,
  - e.g. pixels in an image, words in a sentence, spread of communicable disease, weather measurements in a grid...
- Topics in Undirected Graphs
  - Basic Representation
  - Independences and Factorization
  - Log-Linear Models
  - CRFs
  - Ch. 4 has 29 definitions, 14 theorems, 10 propositions…; we will cover most informally

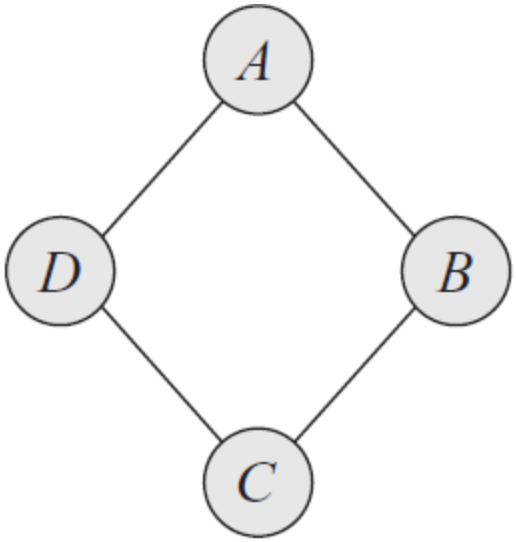
Misconception Example

- Pairs of students in a class meet to discuss home work. Pairs are: (A, B), (B,C), (C,D) and (D,A); however (A, C) and (B,D) don't get along with each other so don't communicate directly.

- Professor misspeaks in class. Each student subsequently may figure out the mistake (may be by reading the book or just by logic). Each student communicates his/her understanding of the possible misconception in the study group.

- We may want to compute the probability of misconception for each student after their meetings

Misconception Model



Interactions for this problem are best represented as an undirected model. Note that no directed model captures the independences correctly.

We will show that in the undirected model:

$(A \perp C \mid B,D)$; $(B \perp D \mid A,C)$

Factors

- Factors for a set of random variables, define "affinities" between them; they are like marginal probability tables but not necessarily normalized.

$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100
(a)			(b)			(c)			(d)		

- φ_1 indicates that A and B agree much more often than not, more weight when they are both right than both wrong. When they disagree, more weight when A is right. However, C and D disagree most of the time (φ_3) etc.

Let $\mathbf{D}$ *be a set of random variables. We define a* factor ϕ *to be a function from* $Val(\mathbf{D})$ *to* $\mathbb{R}$. *A factor is* nonnegative *if all its entries are nonnegative. The set of variables* $\mathbf{D}$ *is called the* scope *of the factor and denoted* $Scope[\phi]$. ■

Factor Product

- Let **X**, **Y** and **Z** be three disjoint sets of variables and let φ_1 (**X**,**Y**) and φ_2 (**Y**,**Z**) be two factors, then factor product is another factor ψ (**X**, **Y**, **Z**)= φ_1 (**X**, **Y**) . φ_2 (**Y**, **Z**).

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	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

Misconception Example

- Joint distribution is defined by the product of factors
  - Does not follow from chain rule as in BNs
  - Formal analysis comes a bit later
- In the misconception example:

  $P(a,b,c,d) = (1/Z)\ \varphi_1(a,b) \cdot \varphi_2(b,c) \cdot \varphi_3(c,d) \cdot \varphi_4(d.a)$ where

  $Z = \sum_{a,b,c,d} \varphi_1(a,b) \cdot \varphi_2(b,c) \cdot \varphi_3(c,d) \cdot \varphi_4(d.a)$

  - Z is a normalizing constant, called the *partition function* (note exponential number of summations in computing Z)
- Table shows values for all combinations
  - $\varphi_1(a^1, b^1) \cdot \varphi_2(b^1, c^0) \cdot \varphi_3(c^0, d^1) \cdot \varphi_4(d^1 .a^1) = 10 \cdot 1 \cdot 100 \cdot 100 = 100{,}000$
- Z is computed by summing all entries (=7,201,840)
- Normalized numbers shown in the last column
- Marginalize over A, C and D to get $P(b^1) = .268$, $P(b^1 | c^0) = .06$

Misconception Example: Joint Distribution

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300,000	0.04
a^0	b^0	c^0	d^1	300,000	0.04
a^0	b^0	c^1	d^0	300,000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5,000,000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1,000,000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100,000	0.014
a^1	b^1	c^1	d^0	100,000	0.014
a^1	b^1	c^1	d^1	100,000	0.014

Gibbs Distribution

A distribution P_Φ is a Gibbs distribution *parameterized by a set of factors* $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ *if it is defined as follows:*

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

where

$$\tilde{P}_\Phi(X_1, \ldots, X_n) = \phi_1(\boldsymbol{D}_1) \times \phi_2(\boldsymbol{D}_2) \times \cdots \times \phi_m(\boldsymbol{D}_m)$$

is an unnormalized measure and

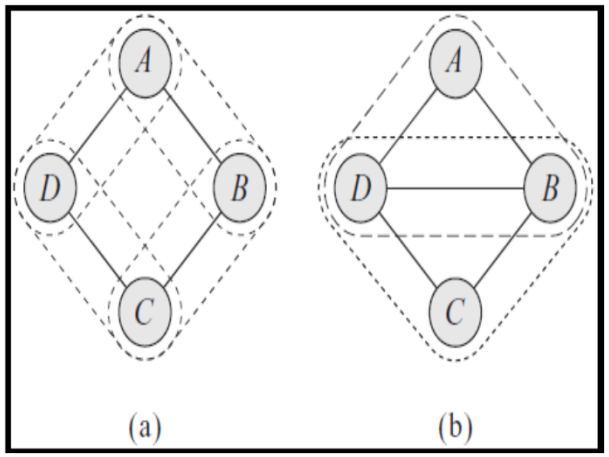
$$Z = \sum_{X_1, \ldots, X_n} \tilde{P}_\Phi(X_1, \ldots, X_n)$$

is a normalizing constant called the partition function. ■

Note $\mathbf{D}_i$ are sets of variables X_i .

Factorization and Clique Potentials

- Each factor in a Markov network (MN) must be defined over a *complete* subgraph (*i.e.* a clique) of H
- Factors that parameterize a MN are also called *clique potentials*.
- We can use any subset of cliques. Potentials over maximal cliques suffice but may hide some structure



In (a) cliques are {A,B}, {B,C}, {C,D} and {A,D}

In (b) cliques are {A,B,D} and {B,C,D}

Note: multiple ways of defining factors.
In (b), we can include {A,B},{B,D},{A} and {B} in addition to or in place of {A,B,D}

Factors

- Note that a factor is not the same as a marginal distribution: shown below are marginals (derived from fig 4.2) and the given factor for (a,b); even the rank order of terms is not the same.
  - Marginals include the influence of other variables (C and D)

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

- How about the case of a MN with just two nodes?
  - Factor corresponds to unnormalized joint distribution

Reduced Markov Networks

- Given evidence variables **U** have assignment **u**. (also called context); in each factor, remove entries inconsistent with the evidence.
- In the example, evidence consists of $C=c^1$ so we drop entries where $C=c^2$.

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

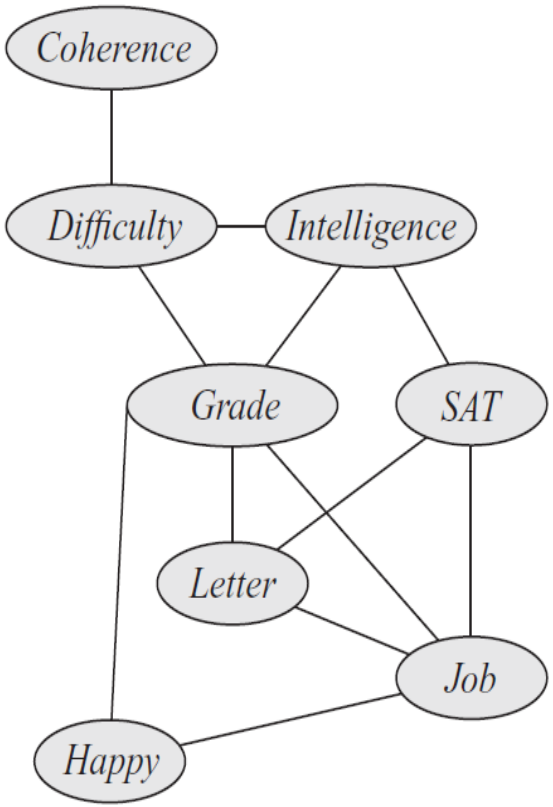
a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Actually, c^1 column s/b omitted

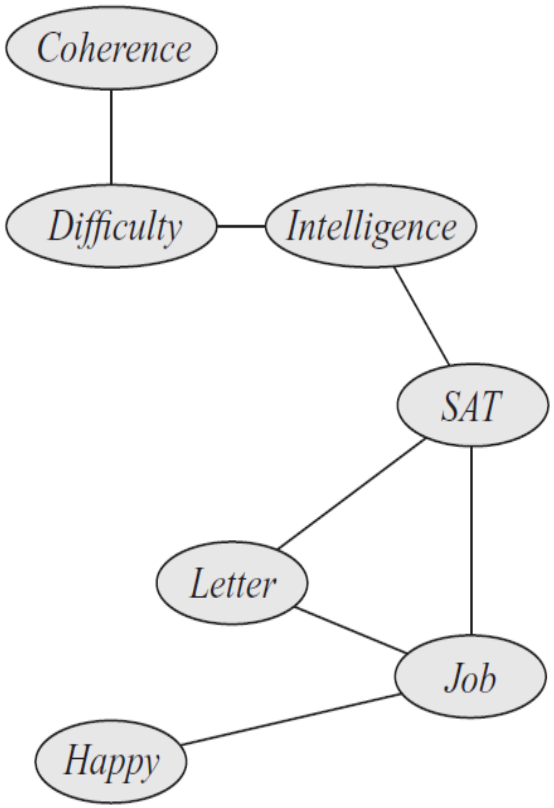
Reduced Distributions (informal)

- Drop terms containing evidence, reduce factor scope
- Reduced Gibbs distribution is parameterized by the reduced factors
- The reduced Gibbs distribution represents conditional probability of non-evidence variables given the evidence variables
- Reduced MN by dropping evidence variables (and edges connecting them) and using the reduced factors
- More precise definitions and theorems in the book
  - Defs 4.5, 4.6, 4.7
  - Propositions 4.1, 4.2

Reduced Network: Example

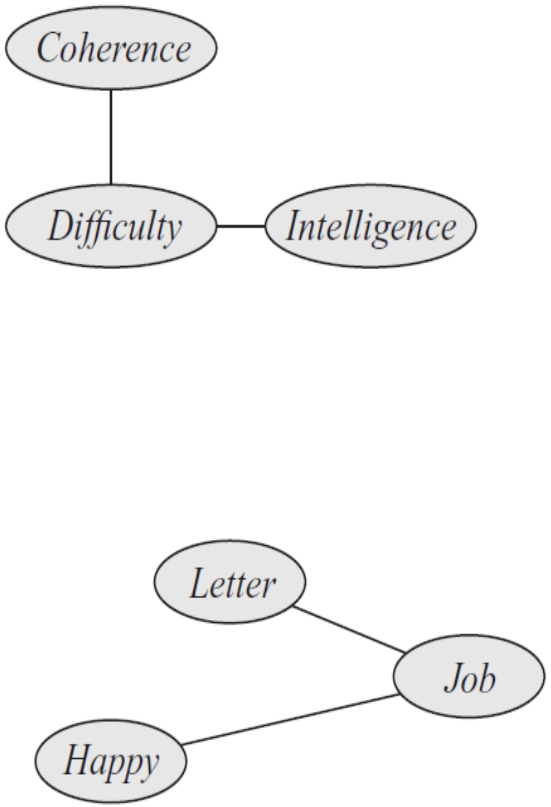


(a)



(b)

Remove Grade



(c)

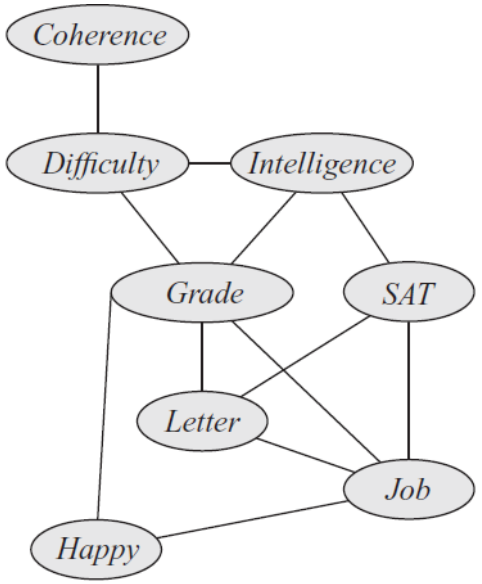
Remove SAT

Local Independencies

- If X and Y are not neighbors, they are conditionally independent given all other node values: $I_p(H)$
  - e.g. D ⊥ J |C,I,G,S,L,H
- Markov Blanket: all neighbors
  - e.g. MB(G) = {D,,H, L, J}
  - Node C.I. of all given MB
  - $I_l(H)$
- Can be shown that:

  if P|= $I_l(H)$ then P|= $I_p(H)$

  (*i.e.* $I_p(H)$ is strictly weaker than $I_l(H)$)



Global Independencies

- Active Trail: No node in the trail is in **Z** (**Z** is the set of conditioning variables)
  - Note: no special cases, such as a v-structure
- Separation

  Sets **X** and **Y** are separated by **Z** if all trails from **X** to **Y** are blocked by **Z** , notation sep_H (**X**; **Y** | **Z**)

  In this case (**X** ⊥ **Y** | **Z**)
- All such independences constitute I(H)
- For the misconception example: (A⊥C| B,D) and (B ⊥D|A, C)

Soundness and Completeness

- Soundness
  - If P factorizes according to H, H is an I-map of P (Thm 4.1)
  - If H is an I-map of P, does P factorize over H?
    - Yes, if P is *strictly positive*, *i.e.* no values in the joint distribution are exactly zero (no impossible combinations).
    - Thm 4.2 (Hammersley-Clifford Thm; history)
  - A positive P factorizes over H *iff* H is an I-map of P
- Completeness (separation discovers all independences)
  - As for BN, the completeness property holds for *almost all* distributions P that factorize over H so I(P) = I(H)
- Main difference with BN theorems is requirement for P to be positive
- Can be shown that P |= I_l(H); P |= I*p*(H) ; P |= I(H) are equivalent if P is ***strictly positive***; otherwise Ip(H) is weaker then I_l(H) which is weaker then I(H)

Distributions to Graphs

- Construct graph given distribution independencies
- Use pairwise independence properties or local (MB) properties.
- Pairwise property
  - If X and Y are *not* conditionally independent, given values of all the other nodes, then there must exist edge X – Y in H.
- MB property
  - Find the minimal set **U** which has the MB property, i.e. X is conditionally independent of all other nodes (excluding those in X and **U**) given values in set **U**.
  - Create edges between X and all members of **U**.
- For positive distributions, both lead to the unique minimal I-map (Thms 4.5 and 4.6)

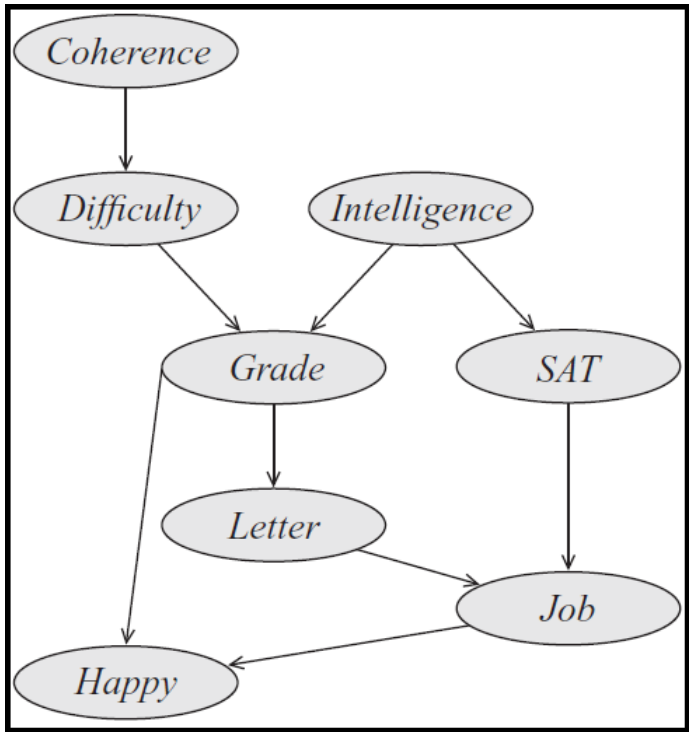
BN ⇔ MN

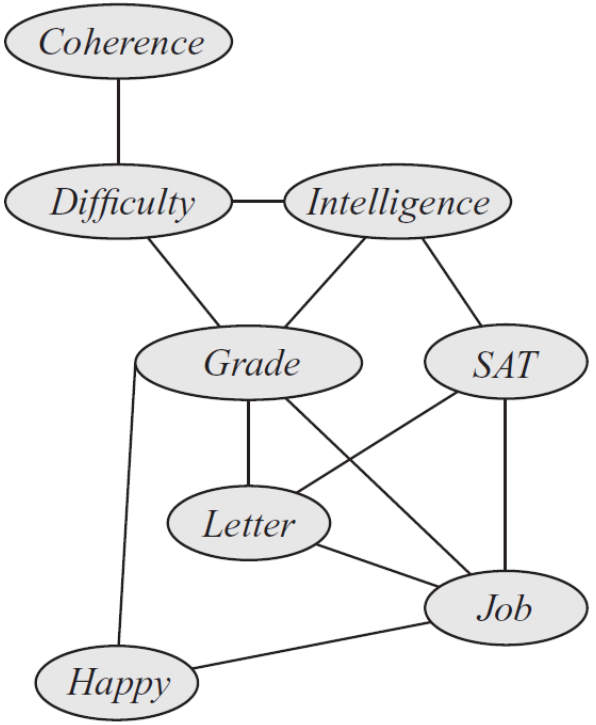
- First consider a simple 2 node network X →Y
  - In MN, we have X –Y;
    - to get factors, use CPD and prior for X
    - Multiply the 2 into a single factor or leave as two factors
  - Overparameterization- Canonical parameters, over all cliques, including single variables (skip details, sec 4.4.2)
- Consider X →Z ←Y (v-structure, X and Y are parents, Z is child node)
- In MN, edges X – Z and Y-Z are obvious but this alone implies that X ⊥ Y | Z which does not hold
  - We need to also put a link between X and Y (moralize)
  - This destroys (X ⊥ Y) independence
- Also implies that a MN P-map may not exist for all distributions

BN ⇔ MN: Factors

- Suppose we start with a BN (graph G) and want to convert to MN (graph H)
- First, assume G has no v-structures
- Then all edges in H can be same as in G
- What about factors?
- Let CPD P (X_i| Pa_{Xi}) define the factor over {X_i, Pa_{Xi}}, resulting distribution corresponds to a Gibbs distribution with z=1
- Similar result in presence of evidence variables E= e (Prop 4.7)
- If v-structures exist, they must be moralized
  - Resulting factor contains 3 (or more) variables
- Example Fig. 4.6 (a) from Fig. 9.8

Example of Transforming BN to MN





Next Class

- Read sections 4.4, 4.5, 4.6.1 of the KF book