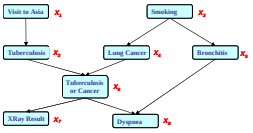
Machine Learning

10-701/15-781, Fall 2006

Graphical Models

Eric Xing



Lecture 12, October 24, 2006

Reading: Chap. 8, C.B book

What is a graphical model?

--- example from medical diagnostics

- A possible world for a patient with lung problem:

Visit to Asia X_1

Smoking X_2

Tuberculosis X_3

Lung Cancer X_4

Bronchitis X_5

Tuberculosis or Cancer X_6

XRay Result X_7

Dyspnea X_8

Recap of Basic Prob. Concepts

- Representation: what is the joint probability dist. on multiple variables?

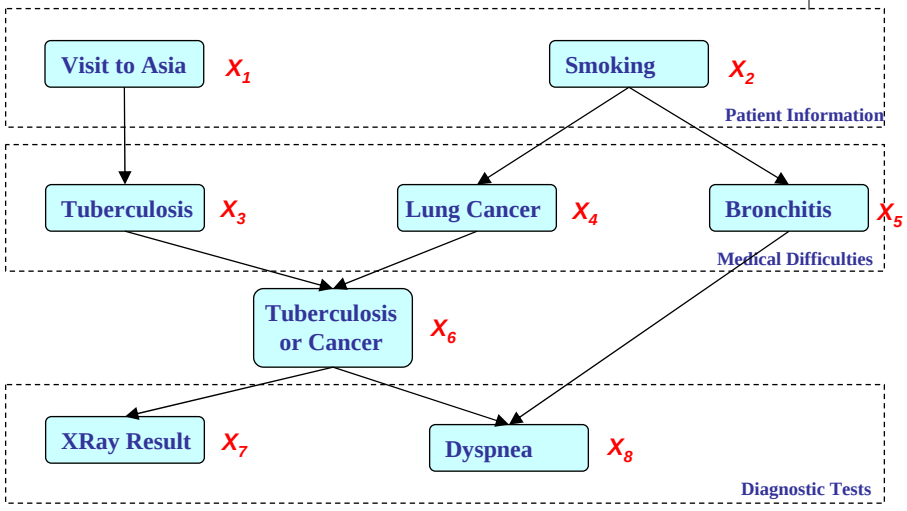
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

  - How many state configurations in total? --- 2^8
  - Are they all needed to be represented?
  - Do we get any scientific/medical insight?

- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

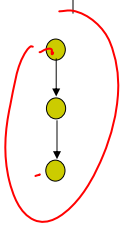
- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?

Dependencies among variables



Probabilistic Graphical Models

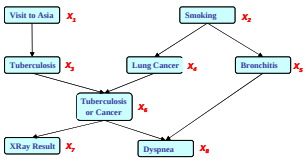
- Represent dependency structure with a graph
  - Node <-> random variable
  - Edges encode dependencies
    - Absence of edge -> conditional independence
  - Directed and undirected versions



- Why is this useful?
  - A language for communication
  - A language for computation
  - A language for development
- Origins:
  - Wright 1920's
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's

Probabilistic Graphical Models, con'd

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)$$
$$P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

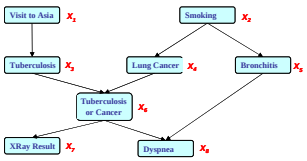
- Why we may favor a PGM?
  - Representation cost: how many probability statements are needed?

    2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8!
  - Algorithms for systematic and efficient inference/learning computation
    - Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics
  - Incorporation of domain knowledge and causal (logical) structures

Two types of GMs

- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):
- Undirected edges simply give (physical or symmetric) correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

Bayesian Network: Factorization Theorem



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

- **Theorem:**

  Given a DAG, The most general form of the probability distribution that is consistent with the graph factors according to "node given its parents":

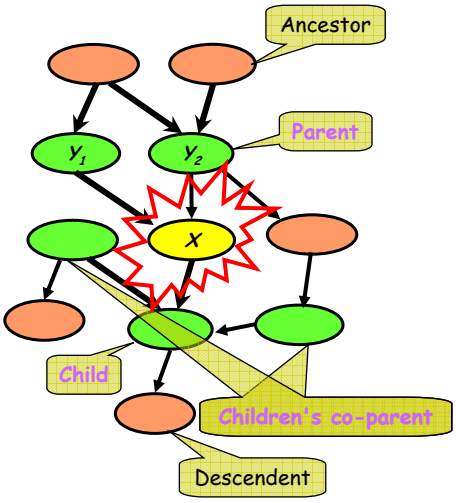
$$P(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{X}_{\pi_i})$$

  where $\mathbf{X}_{\pi_i}$ is the set of parents of x_i, d is the number of nodes (variables) in the graph.

Bayesian Network: Conditional Independence Semantics

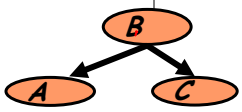
Structure: ***DAG***

- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**
- Local conditional distributions (**CPD**) and the **DAG** completely determine the **joint** dist.
- Give **causality** relationships, and facilitate a **generative** process



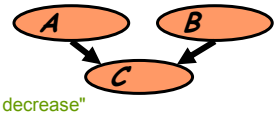
Local Structures & Independencies

- Common parent
  - Fixing B decouples A and C
    "given the level of gene B, the levels of A and C are independent"



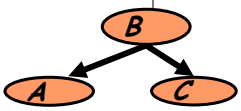
- Cascade
  - Knowing B decouples A and C
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"

- V-structure
  - Knowing C couples A and B
    because A can "explain away" B w.r.t. C
    "If A correlates to C, then chance for B to also correlate to C will decrease"



- The language is compact, the concepts are rich!

A simple justification



$A \perp C \mid B$

$P(A, B, C) = P(B)\,P(A \mid B)\,P(C \mid B)$

$P(A, C \mid B) = \frac{P(A, C, B)}{P(B)} = \frac{P(B)\,P(A \mid B)\,P(C \mid B)}{P(B)}$

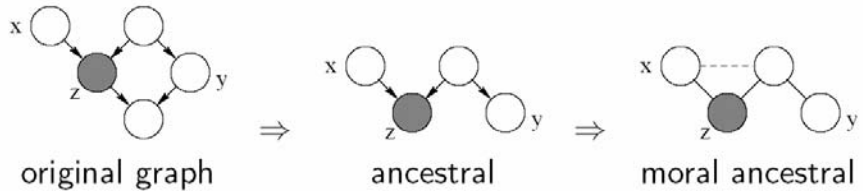
$= P(A \mid B)\,P(C \mid B)$

Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

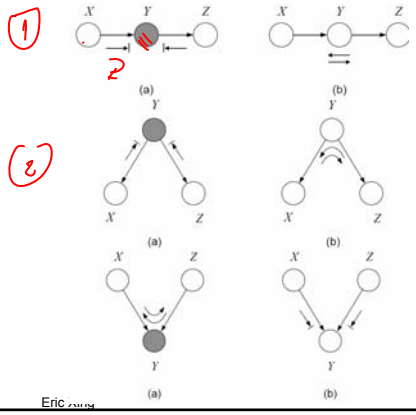
  Definition: variables x and y are *D-separated* (conditionally independent) given z if they are separated in the *moralized* ancestral graph

- Example:



Global Markov properties of DAGs

- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):

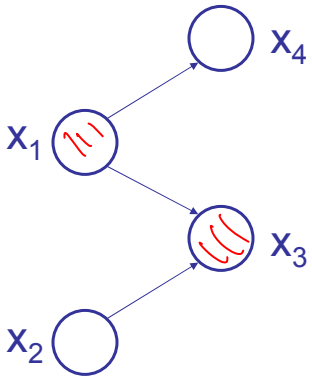


- Defn: $\mathcal{I}(\mathcal{G})$=all independence properties that correspond to d-separation:

$$\mathrm{I}(G) = \{X \perp Z | Y : \mathrm{dsep}_G(X; Z | Y)\}$$

- D-separation is sound and complete

Example:



- Complete the I(G) of this graph:

$X_1 \perp X_2$

$X_2 \perp X_4$

$X_2 \perp X_4 \mid \{X_1, X_3\}$

$X_2 \perp X_4 \mid X_1$

$X_3 \perp X_4 \mid X_1$

$X_4 \perp \{X_1, X_2\} \mid X_3$

Towards quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- **The Equivalence Theorem**

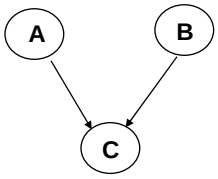
  For a graph G,

  Let $\mathcal{D}_1$ denote the family of all distributions that satisfy I(G),

  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to G,

  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

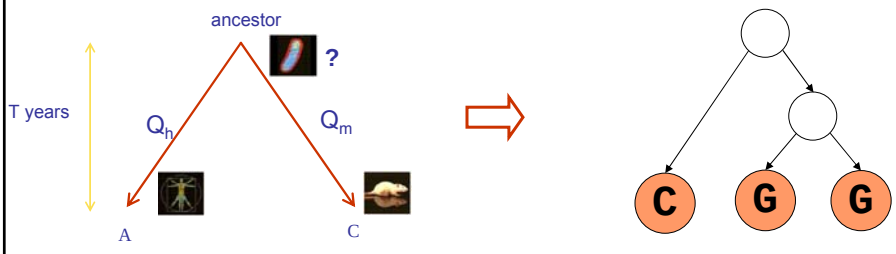
Example



$\longleftrightarrow$ p(A,B,C) =

Example, con'd

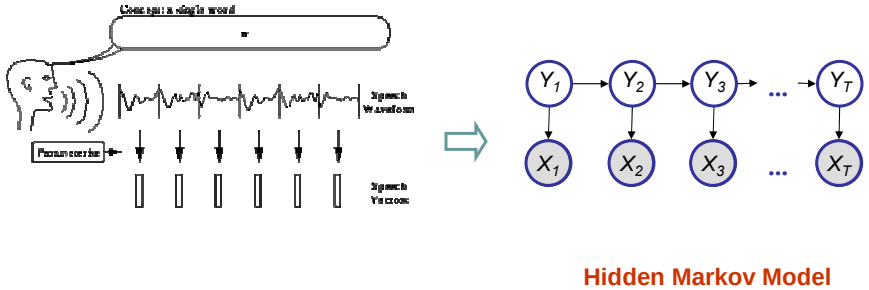
- Evolution



Tree Model

Example, con'd

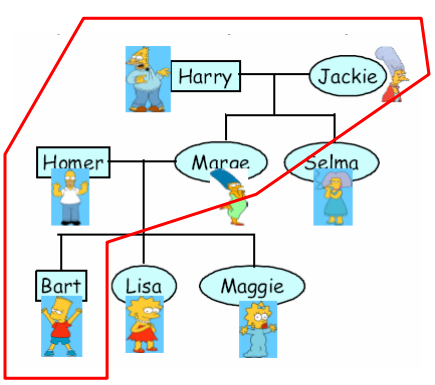
- Speech recognition

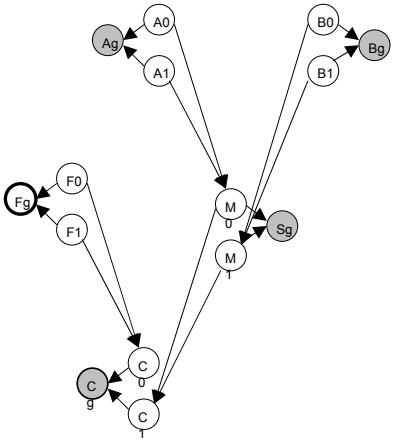


Hidden Markov Model

Example, con'd

- Genetic Pedigree



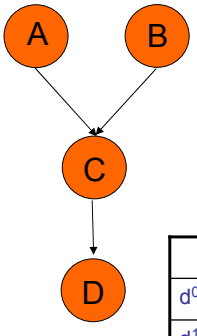


Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)



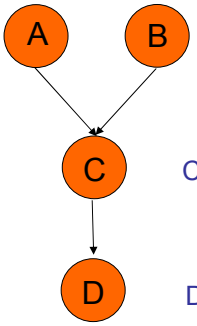
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Conditional probability density func. (CPDs)

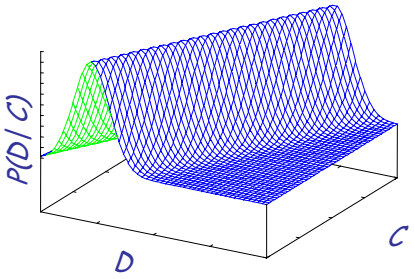
A~N(μ_a, Σ_a) B~N(μ_b, Σ_b)

P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)

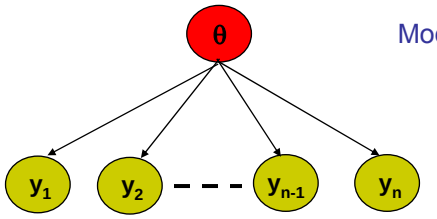


C~N(A+B, Σ_c)

D~N(μ_a+C, Σ_a)



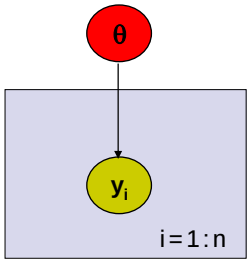
Conditionally Independent Observations



Model parameters

Data

"Plate" Notation



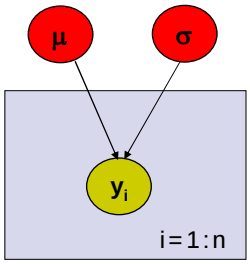
Model parameters

Data = $\{y_1,\ldots y_n\}$

Plate = rectangle in graphical model

variables within a plate are replicated
in a conditionally independent manner

Example: Gaussian Model



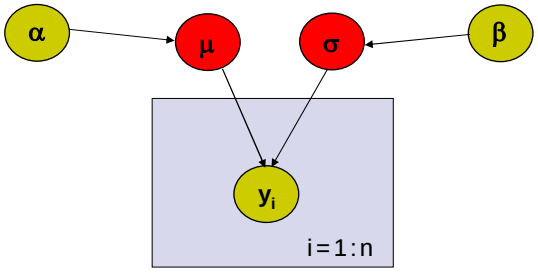
Generative model:

$p(y_1,\ldots y_n \mid \mu, \sigma) = \prod_i p(y_i \mid \mu, \sigma)$
$= p(\text{data} \mid \text{parameters})$
$= p(D \mid \theta)$

where $\theta = \{\mu, \sigma\}$

- Likelihood = p(data | parameters)
  = p(D | θ)
  = L (θ)
- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with log L (θ)

Example: Bayesian Gaussian Model



Note: priors and parameters are assumed independent here

Why graphical models

- **Probability theory** provides the **glue** whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.
- The **graph theoretic** side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- **Many of the classical multivariate probabilistic systems** studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics **are special cases of the general graphical model formalism**
  - examples include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models.
- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.

--- M. Jordan