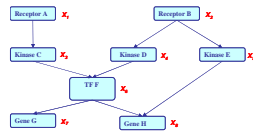


Graphical Models (1)

Representation

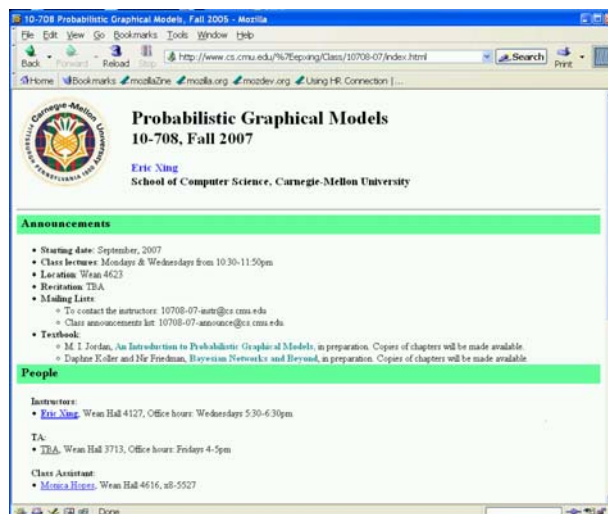
Eric Xing

Carnegie Mellon University
May 31, 2007



Eric Xing,
A lecture series at the Institute of Theoretical Computer
Science, Tsinghua University, May 31-June 7, 2007

1



10-708 Probabilistic Graphical Models, Fall 2007 - Mozilla

http://www.cs.cmu.edu/~Eepang/Class/10708-07/index.html

Probabilistic Graphical Models 10-708, Fall 2007

Eric Xing
School of Computer Science, Carnegie Mellon University

Announcements

- Starting date: September, 2007
- Class lectures: Mondays & Wednesdays from 10:30-11:50pm
- Location: Wean 4623
- Revision: TBA
- Mailing List:
 - To contact the instructor: 10708-07-aut@cs.cmu.edu
 - Class announcements list: 10708-07-announce@cs.cmu.edu
- Textbook:
 - M. I. Jordan, An Introduction to Probabilistic Graphical Models, in preparation. Copies of chapters will be made available.
 - Daphne Koller and Nir Friedman, Bayesian Networks and Beyond, in preparation. Copies of chapters will be made available.

People

Instructors:

- Eric Xing: Wean Hall 4127, Office hours: Wednesdays 5:30-6:30pm

TA:

- TBA: Wean Hall 3713, Office hours: Fridays 4-5pm

Class Assistant:

- Mona Hoss: Wean Hall 4616, s5-5527

Eric Xing

2

What is this?

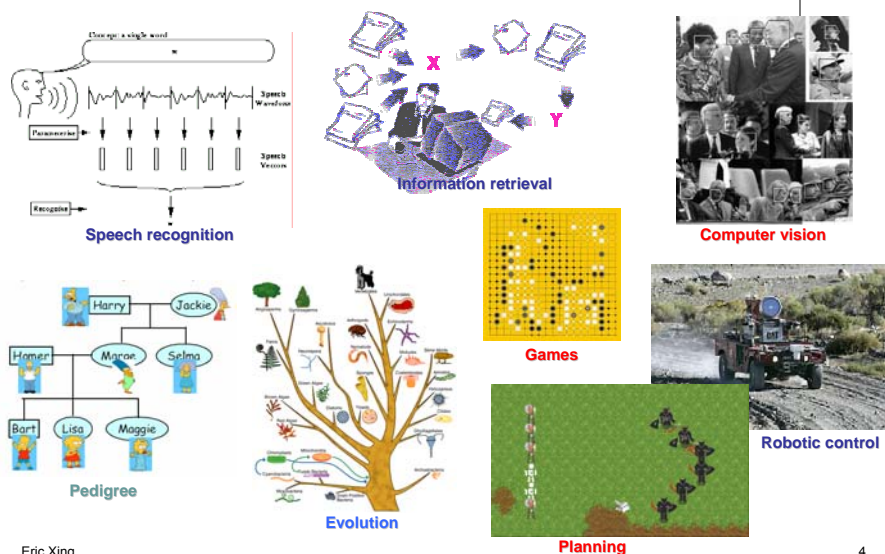


- Classical AI and ML research ignored this phenomena
- The Problem (an example):
 - you want to catch a flight at 10:00am from Beijing to Pittsburgh, can I make it if I leave at 7am and take a Taxi at the east gate of Tsinghua?
 - partial observability (road state, other drivers' plans, etc.)
 - noisy sensors (radio traffic reports)
 - uncertainty in action outcomes (flat tire, etc.)
 - immense complexity of modeling and predicting traffic
- Reasoning under **uncertainty**!

Eric Xing

3

A universal task ...



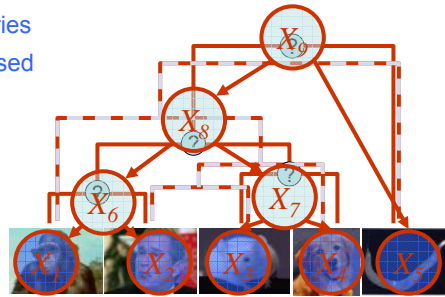
Eric Xing

4

The Fundamental Questions



- Representation
 - How to capture/model uncertainties in possible worlds?
 - How to encode our domain knowledge/assumptions/constraints?
- Inference
 - How do I answer questions/queries according to my model and/or based given data?
e.g.: $P(X_i | \mathcal{D})$
- Learning
 - What model is "right" for my data?
e.g.: $\mathcal{M} = \arg \max_{\mathcal{M} \in \mathcal{M}} F(\mathcal{D}; \mathcal{M})$



Eric Xing

5

Graphical Models



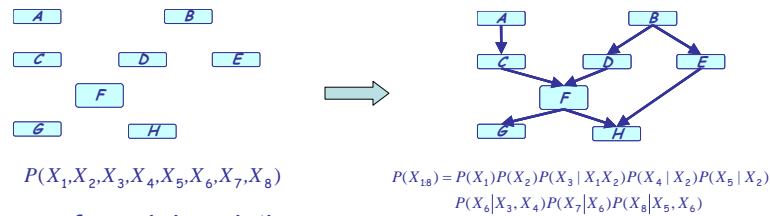
- Graphical models are a marriage between graph theory and probability theory
- One of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats,...) in the last two decades...
- Some advantages of the graphical model point of view
 - Inference and learning are treated together
 - Supervised and unsupervised learning are merged seamlessly
 - Missing data handled nicely
 - A focus on conditional independence and computational issues
 - Interpretability (if desired)
- Are having significant impact in science, engineering and beyond!

Eric Xing

6

What is a Graphical Model?

- The informal blurb:
 - It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with structured semantics



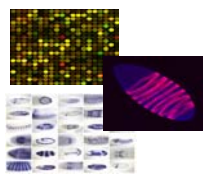
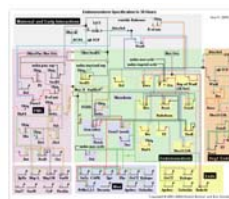
- A more formal description:
 - It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

Eric Xing

7

Statistical Inference

probabilistic
generative
model

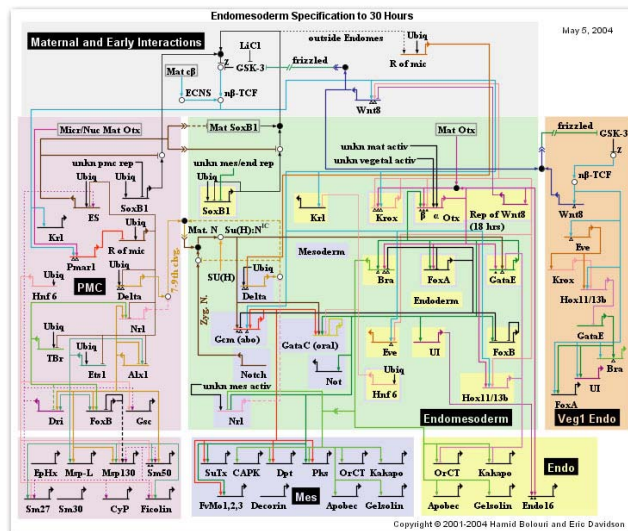


gene expression profiles

Eric Xing

8

Statistical Inference

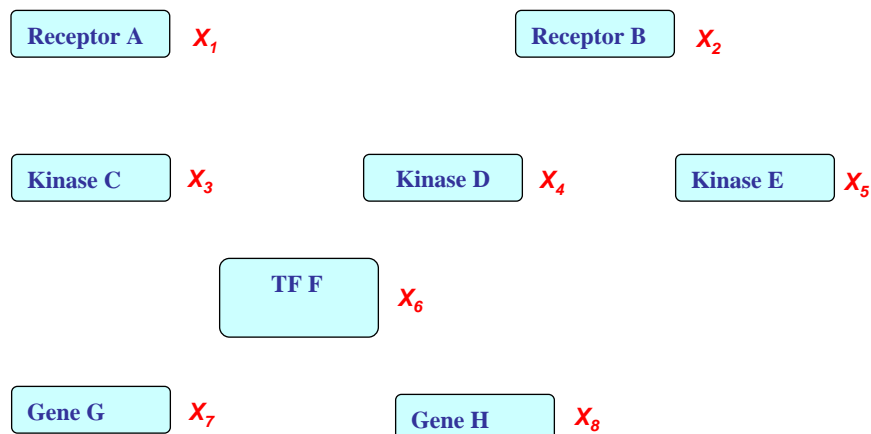


Eric Xing

9

Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:



Eric Xing

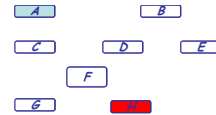
10

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?
 - Computing $p(H|A)$ would require summing over all 2^6 configurations of the unobserved variables

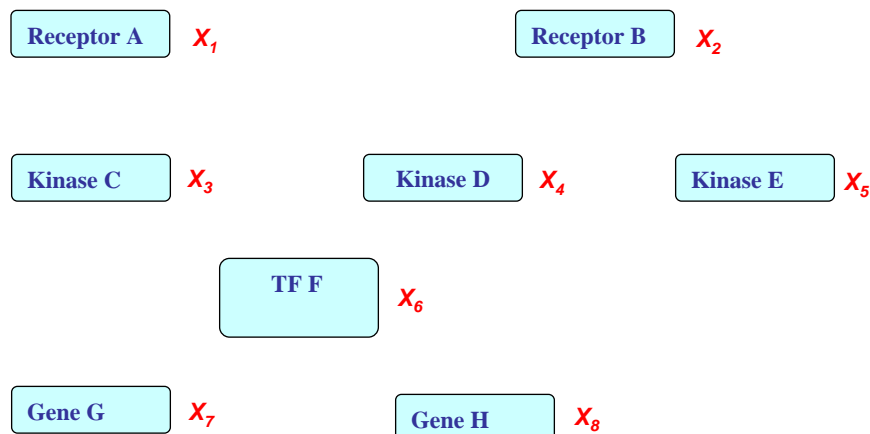
Eric Xing

11

What is a Graphical Model?

--- example from a signal transduction pathway

- A possible world for cellular signal transduction:

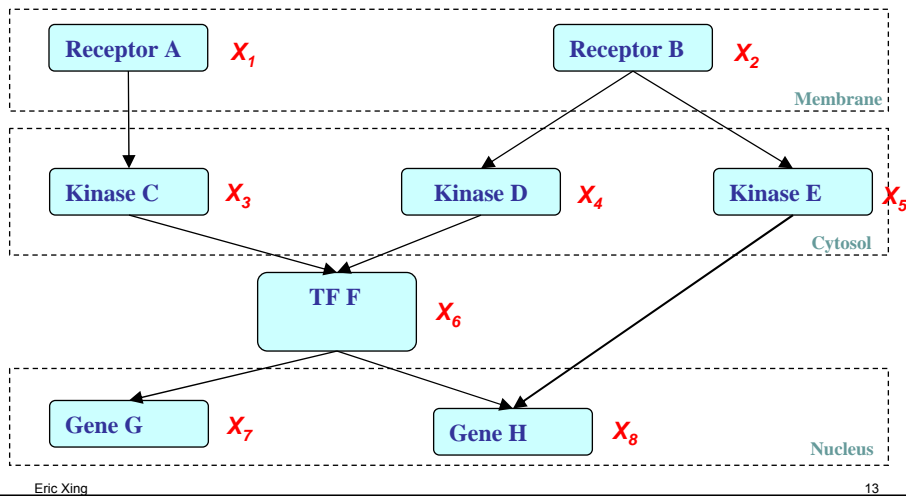


Eric Xing

12

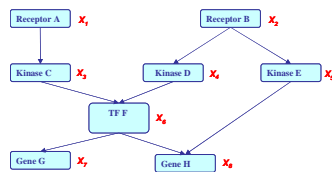
GM: Structure Simplifies Representation

- Dependencies among variables



Probabilistic Graphical Models

- 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.,

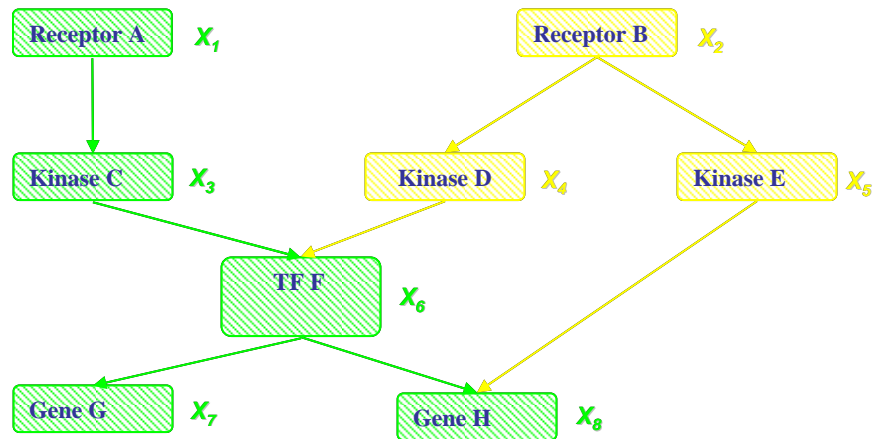


$$\begin{aligned}
 &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
 &= P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) \\
 &\quad P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6)
 \end{aligned}$$

Stay tune for what are these independencies!

- Why we may favor a PGM?
 - Incorporation of domain knowledge and causal (logical) structures
 $2+2+4+4+4+8+4+8=36$, an 8-fold reduction from 2^8 in representation cost !

GM: Data Integration

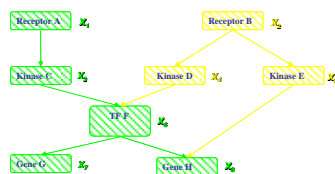


Eric Xing

15

Probabilistic Graphical Models

- 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.,



$$\begin{aligned}
 &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
 &= P(X_2) P(X_4/X_2) P(X_5/X_2) P(X_1) P(X_3/X_1) \\
 &\quad P(X_6/X_3, X_4) P(X_7/X_6) P(X_8/X_5, X_6)
 \end{aligned}$$

- Why we may favor a PGM?
 - Incorporation of domain knowledge and causal (logical) structures
 $2+2+4+4+4+8+4+8=36$, an 8-fold reduction from 2^8 in representation cost !
 - Modular combination of heterogeneous parts – data fusion

Eric Xing

16

Rational Statistical Inference

The Bayes Theorem:

$$p(h | d) = \frac{p(d | h) p(h)}{\sum_{h' \in H} p(d | h') p(h')}$$

Posterior probability $p(h | d)$

Likelihood $p(d | h)$

Prior probability $p(h)$

Sum over space of hypotheses $\sum_{h' \in H}$

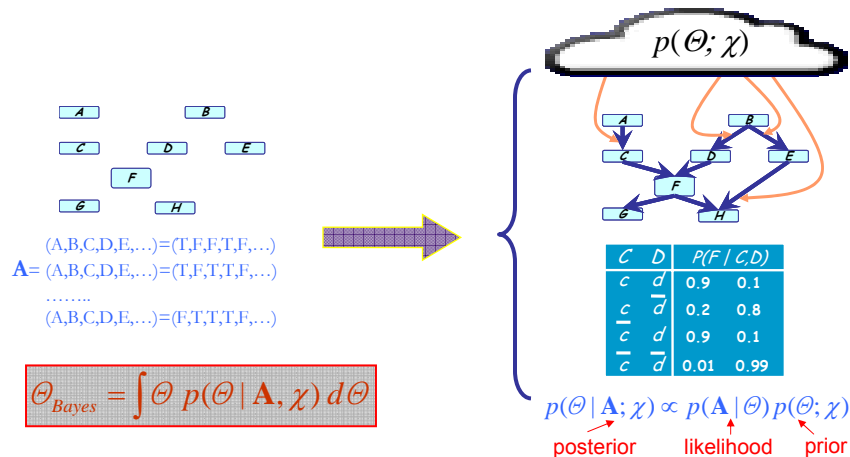
- This allows us to capture uncertainty about the model in a principled way
- But how can we specify and represent a complicated model?
 - Typically the number of genes need to be modeled are in the order of thousands!

Eric Xing

17

GM: MLE and Bayesian Learning

- Probabilistic statements of θ is conditioned on the values of the observed variables \mathbf{A}_{obs} and prior $p(\cdot | \chi)$

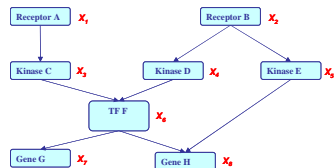


Eric Xing

18

Probabilistic Graphical Models

- 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?
 - Incorporation of domain knowledge and causal (logical) structures
 $2+2+4+4+4+8+4+8=36$, an 8-fold reduction from 2^8 in representation cost !
 - Modular combination of heterogeneous parts – data fusion
 - Bayesian Philosophy
 - Knowledge meets data

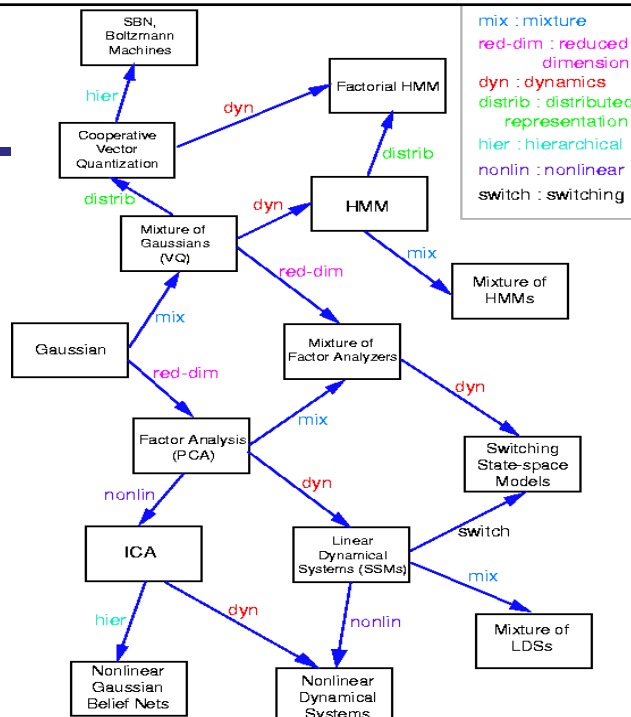


Eric Xing

19

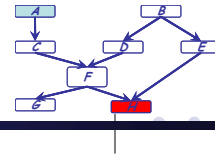
An (incomplete) genealogy of graphical models

(Picture by Zoubin Ghahramani and Sam Roweis)



Eric Xing

Probabilistic Inference



- Computing statistical queries regarding the network, e.g.:
 - Is node X independent on node Y given nodes Z,W ?
 - What is the probability of X=true if (Y=false and Z=true)?
 - What is the joint distribution of (X,Y) if Z=false?
 - What is the likelihood of some full assignment?
 - What is the most likely assignment of values to all or a subset the nodes of the network?
- General purpose algorithms exist to fully automate such computation
 - Computational cost depends on the topology of the network
 - Exact inference:
 - The junction tree algorithm
 - Approximate inference;
 - Loopy belief propagation, variational inference, Monte Carlo sampling

Eric Xing

21

A few myths about graphical models



- They require a localist semantics for the nodes ✓
- They require a causal semantics for the edges ✗
- They are necessarily Bayesian ✗
- They are intractable ✗

Eric Xing

22

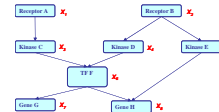
Two types of GMs

- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

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

$$= P(X_1) P(X_2) P(X_3/X_1) P(X_4/X_2) P(X_5/X_2)$$

$$P(X_6/X_3, X_4) P(X_7/X_6) P(X_8/X_5, X_6)$$

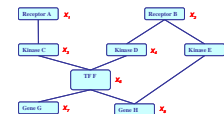


- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

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

$$= \frac{1}{Z} \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2)$$

$$+E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$

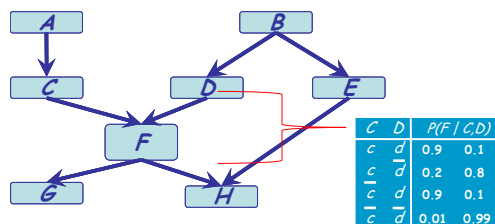


Eric Xing

23

Specification of a directed GM

- There are two components to any GM:
 - the qualitative specification
 - the quantitative specification



Eric Xing

24

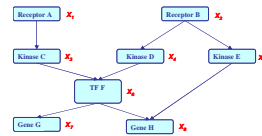
Bayesian Network: Factorization Theorem

- 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”:

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

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



$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ &= P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) \\ &\quad P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6) \end{aligned}$$

Eric Xing

25

Qualitative Specification

- Where does the qualitative specification come from?

- Prior knowledge of causal relationships
- Prior knowledge of modular relationships
- Assessment from experts
- Learning from data
- We simply link a certain architecture (e.g. a layered graph)
- ...

Eric Xing

26

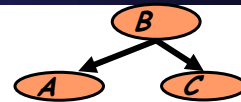
Local Structures & Independencies



- 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"

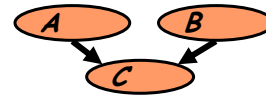


- V-structure

- 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 B will decrease"



- The language is compact, the concepts are rich!

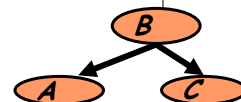
Eric Xing

27

A simple justification



$$A \perp\!\!\!\perp C \mid B$$



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

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

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

$$= P(A|B) P(C|B)$$

Eric Xing

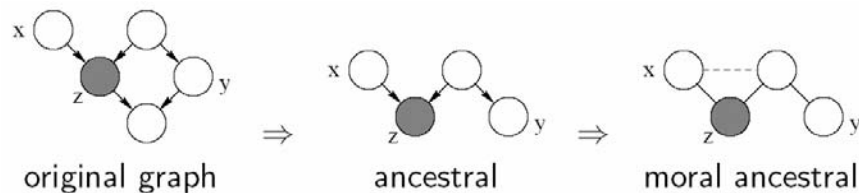
28

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:

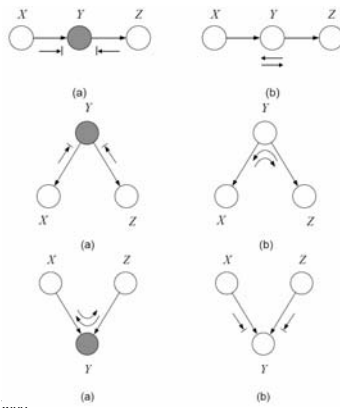


Eric Xing

29

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 below (and plus some boundary conditions):



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

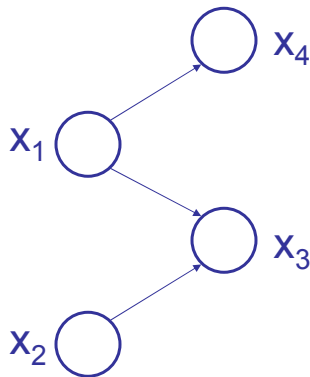
$$I(G) = \{X \perp Z | Y : \text{dsep}_G(X; Z | Y)\}$$

- D-separation is sound and complete

Eric Xing

30

Example:



- Complete the I(G) of this graph:

$$x_1 \perp\!\!\!\perp x_2$$

$$x_2 \perp\!\!\!\perp x_4$$

$$x_2 \perp\!\!\!\perp x_4 \mid \{x_1, x_3\}$$

$$x_2 \perp\!\!\!\perp x_4 \mid x_1$$

$$x_3 \perp\!\!\!\perp x_4 \mid x_1$$

$$x_4 \perp\!\!\!\perp \{x_2, x_3\} \mid x_1$$

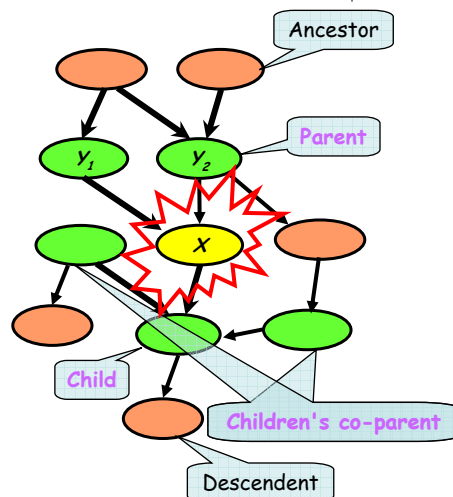
Eric Xing

31

Summary: Conditional Independence Semantics in an BN

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



Eric Xing

32

Toward quantitative specification of probability distribution



- Separation properties in the graph imply independence properties about the associated variables
- **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 ,

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

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

Eric Xing

33

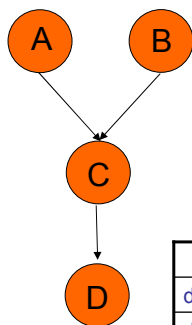
Conditional probability tables (CPTs)



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	0.7	0.5

Eric Xing

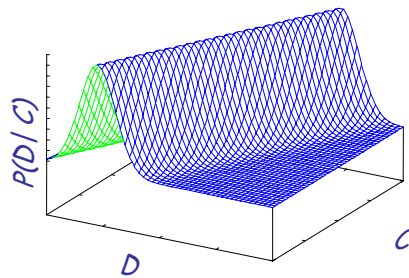
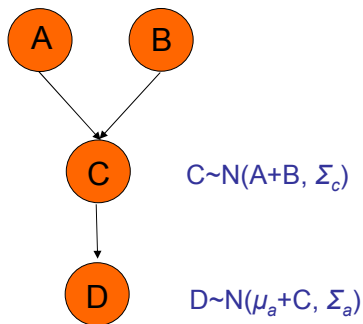
34

Conditional probability density func. (CPDs)



$$A \sim N(\mu_a, \Sigma_a) \quad B \sim N(\mu_b, \Sigma_b)$$

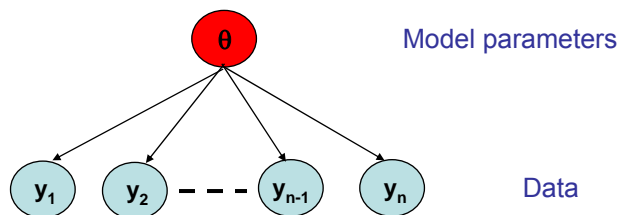
$$P(a,b,c,d) = P(a)P(b)P(c|a,b)P(d|c)$$



Eric Xing

35

Conditionally Independent Observations



Eric Xing

36

“Plate” Notation

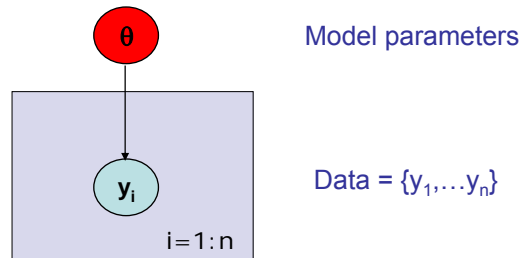


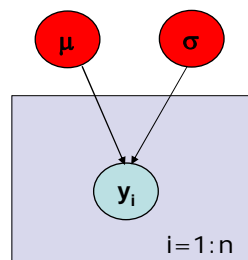
Plate = rectangle in graphical model

variables within a plate are replicated
in a conditionally independent manner

Eric Xing

37

Example: Gaussian Model



Generative model:

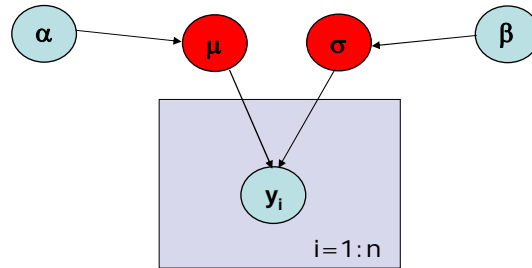
$$\begin{aligned} p(y_1, \dots, y_n \mid \mu, \sigma) &= \prod_i p(y_i \mid \mu, \sigma) \\ &= p(\text{data} \mid \text{parameters}) \\ &= p(D \mid \theta) \\ \text{where } \theta &= \{\mu, \sigma\} \end{aligned}$$

- Likelihood = $p(\text{data} \mid \text{parameters})$
= $p(D \mid \theta)$
= $L(\theta)$
- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
 - Often easier to work with $\log L(\theta)$

Eric Xing

38

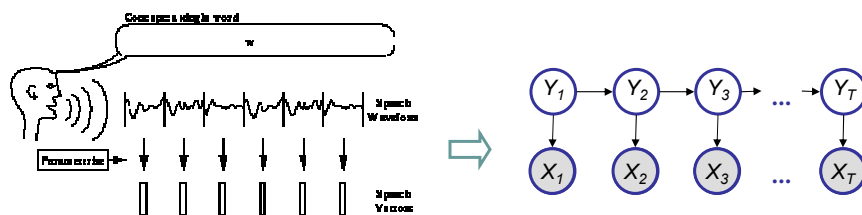
Example: Bayesian Gaussian Model



Note: priors and parameters are assumed independent here

Example

- Speech recognition

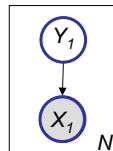
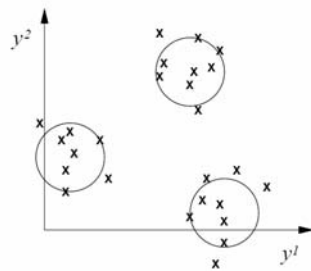


Hidden Markov Model

Hidden Markov Model: from static to dynamic mixture models

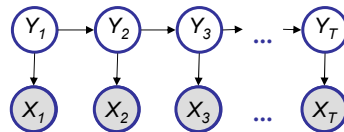
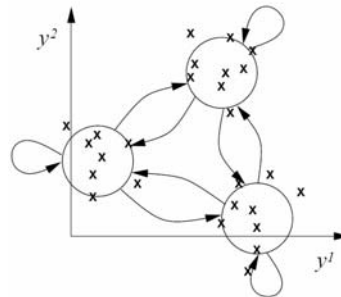


Static mixture



Eric Xing

Dynamic mixture

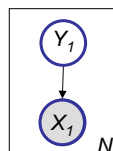
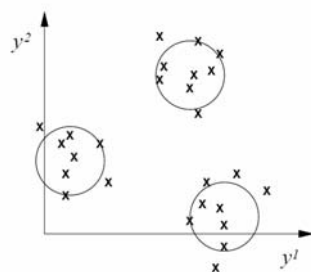


41

Hidden Markov Model: from static to dynamic mixture models

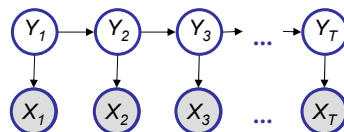
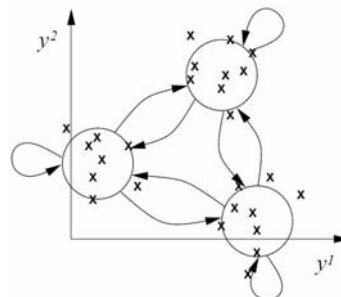


Static mixture



Eric Xing

Dynamic mixture



The underlying source:

Speech signal,
dice,

The sequence:
Phonemes,
sequence of rolls,

42

The Dishonest Casino

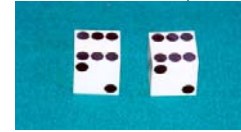
A casino has two dice:

- Fair die
 $P(1) = P(2) = P(3) = P(5) = P(6) = 1/6$
- Loaded die
 $P(1) = P(2) = P(3) = P(5) = 1/10$
 $P(6) = 1/2$

Casino player switches back-&-forth between fair and loaded die once every 20 turns

Game:

1. You bet \$1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins \$2

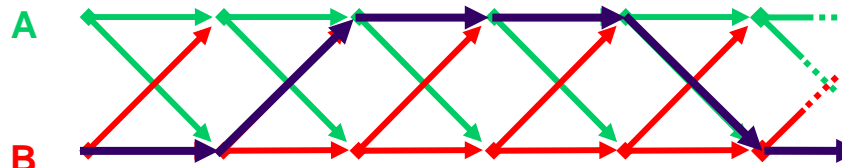


Eric Xing

43

A stochastic generative model

- Observed sequence:



- Hidden sequence (a parse or segmentation):

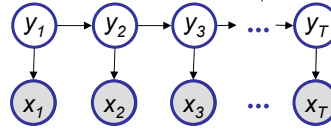


Eric Xing

44

Definition (of HMM)

- **Observation space**
 Alphabetic set: $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$
 Euclidean space: \mathbb{R}^d
- **Index set of hidden states**
 $\mathcal{I} = \{1, 2, \dots, M\}$
- **Transition probabilities** between any two states
 $p(y_t^j = 1 | y_{t-1}^i = 1) = a_{i,j}$,
 or $p(y_t | y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \dots, a_{i,M}), \forall i \in \mathcal{I}$.
- **Start probabilities**
 $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_M)$.
- **Emission probabilities** associated with each state
 $p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \dots, b_{i,K}), \forall i \in \mathcal{I}$.
 or in general:
 $p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in \mathcal{I}$.



Eric Xing

45

Puzzles regarding the dishonest casino

GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

QUESTION

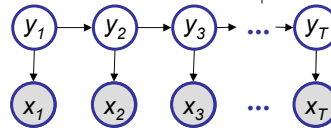
- How likely is this sequence, given our model of how the casino works?
 - This is the **EVALUATION** problem in HMMs
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
 - This is the **DECODING** question in HMMs
- How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?
 - This is the **LEARNING** question in HMMs

Eric Xing

46

Probability of a parse

- Given a sequence $\mathbf{x} = x_1, \dots, x_T$ and a parse $\mathbf{y} = y_1, \dots, y_T$,
- To find how likely is the parse: (given our HMM and the sequence)



$$\begin{aligned}
 p(\mathbf{x}, \mathbf{y}) &= p(x_1, \dots, x_T, y_1, \dots, y_T) && \text{(Joint probability)} \\
 &= p(y_1) p(x_1 | y_1) p(y_2 | y_1) p(x_2 | y_2) \dots p(y_T | y_{T-1}) p(x_T | y_T) \\
 &= p(y_1) p(y_2 | y_1) \dots p(y_T | y_{T-1}) \times p(x_1 | y_1) p(x_2 | y_2) \dots p(x_T | y_T) \\
 &= p(y_1, \dots, y_T) p(x_1, \dots, x_T | y_1, \dots, y_T)
 \end{aligned}$$

$$\begin{aligned}
 \text{Let } \pi_{y_1} &\stackrel{\text{def}}{=} \prod_{i=1}^M [\pi_i]^{y_1^i}, \quad a_{y_t, y_{t-1}} \stackrel{\text{def}}{=} \prod_{i,j=1}^M [a_{ij}]^{y_t^i y_{t-1}^j}, \quad \text{and } b_{y_t, x_t} \stackrel{\text{def}}{=} \prod_{i=1}^M \prod_{k=1}^K [b_{ik}]^{y_t^i x_t^k}, \\
 &= \pi_{y_1} a_{y_1, y_0} \dots a_{y_{T-1}, y_{T-2}} b_{y_1, x_1} \dots b_{y_T, x_T}
 \end{aligned}$$

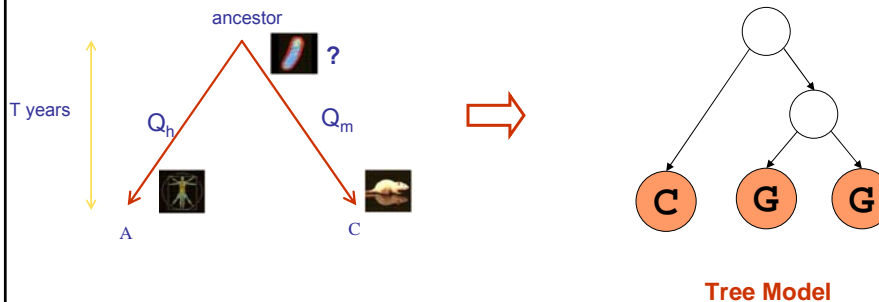
- Marginal probability: $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = \sum_{y_1} \sum_{y_2} \dots \sum_{y_T} \pi_{y_1} \prod_{t=2}^T a_{y_{t-1}, y_t} \prod_{t=1}^T p(x_t | y_t)$
- Posterior probability: $p(\mathbf{y} | \mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / p(\mathbf{x})$

Eric Xing

47

Example, con'd

- Evolution

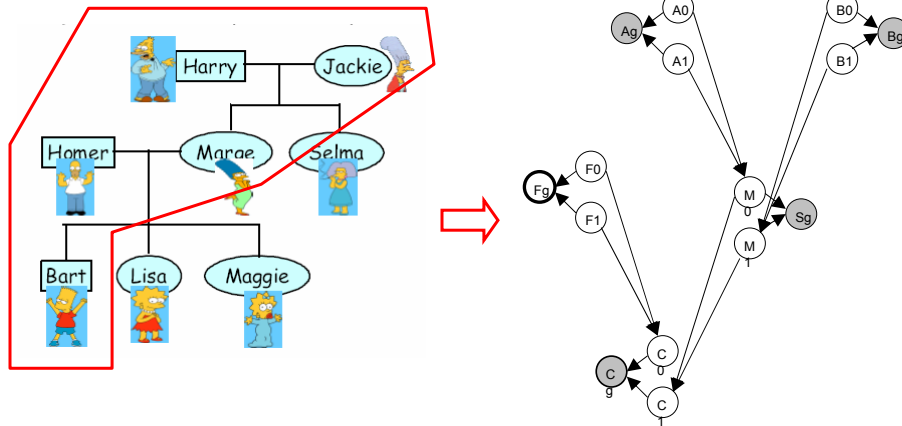


Eric Xing

48

Example, con'd

- Genetic Pedigree



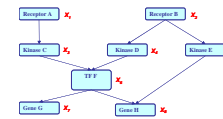
Eric Xing

49

Two types of GMs

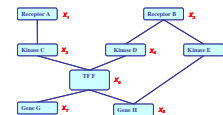
- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

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



- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ = 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) \\ + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$

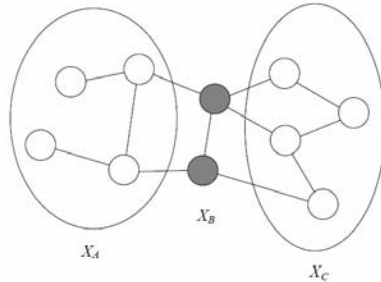


Eric Xing

50

Semantics of Undirected Graphs

- Let H be an undirected graph:



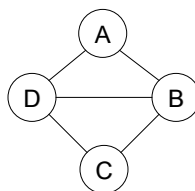
- B **separates** A and C if every path from a node in A to a node in C passes through a node in B : $\text{sep}_H(A; C|B)$
- A probability distribution satisfies the **global Markov property** if for any disjoint A, B, C , such that B separates A and C , A is independent of C given B : $I(H) = \{(A \perp C|B) : \text{sep}_H(A; C|B)\}$

Eric Xing

51

Cliques

- For $G=\{V, E\}$, a complete subgraph (clique) is a subgraph $G'=\{V' \subseteq V, E' \subseteq E\}$ such that nodes in V' are fully interconnected
- A (maximal) clique is a complete subgraph s.t. any **superset** $V'' \supset V'$ is not complete.
- A sub-clique is a not-necessarily-maximal clique.



- Example:
 - max-cliques = $\{A, B, D\}, \{B, C, D\}$,
 - sub-cliques = $\{A, B\}, \{C, D\}, \dots \rightarrow$ all edges and singletons

Eric Xing

52

Quantitative Specification

- Defn: an **undirected graphical model** represents a distribution $P(X_1, \dots, X_n)$ defined by an undirected graph H , and a set of positive **potential functions** ψ_c associated with cliques of H , s.t.

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

where Z is known as the partition function:

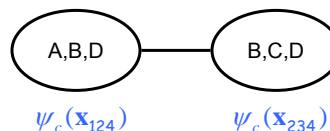
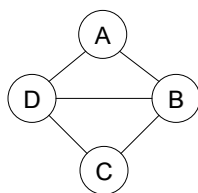
$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

- Also known as **Markov Random Fields**, **Markov networks** ...
- The **potential function** can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.

Eric Xing

53

Example UGM – using max cliques



$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

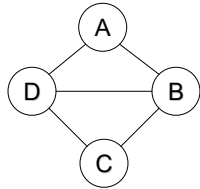
$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

- For discrete nodes, we can represent $P(X_{1:n})$ as two 3D tables instead of one 4D table

Eric Xing

54

Example UGM – using subcliques

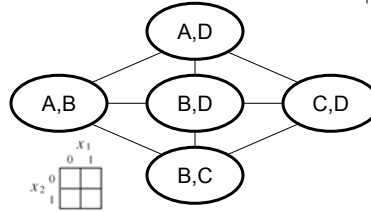


$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

$$= \frac{1}{Z} \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34})$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

- For discrete nodes, we can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table



Eric Xing

55

Hammersley-Clifford Theorem

- If arbitrary potentials are utilized in the following product formula for probabilities,

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

then the family of probability distributions obtained is exactly that set which respects the qualitative specification (the conditional independence relations) described earlier

Eric Xing

56

Interpretation of Clique Potentials



- The model implies $X \perp Z | Y$. This independence statement implies (by definition) that the joint must factorize as:

$$p(x, y, z) = p(y)p(x|y)p(z|y)$$

- We can write this as: $p(x, y, z) = p(x, y)p(z|y)$, but $p(x, y, z) = p(x|y)p(z, y)$
 - cannot** have all potentials be **marginals**
 - cannot** have all potentials be **conditionals**
- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

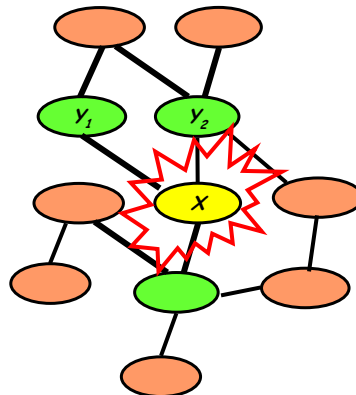
Eric Xing

57

Summary: Conditional Independence Semantics in an MRF

Structure: an **undirected graph**

- Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**
- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** dist.
- Give **correlations** between variables, but no explicit way to generate samples



Eric Xing

58

Exponential Form



- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential $\psi_c(\mathbf{x}_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(\mathbf{x}_c)$:

$$\psi_c(\mathbf{x}_c) = \exp\{-\phi_c(\mathbf{x}_c)\}$$

For convenience, we will call $\phi_c(\mathbf{x}_c)$ a potential when no confusion arises from the context.

- This gives the joint a nice additive structure

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{-\sum_{c \in C} \phi_c(\mathbf{x}_c)\right\} = \frac{1}{Z} \exp\{-H(\mathbf{x})\}$$

where the sum in the exponent is called the "free energy":

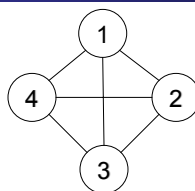
$$H(\mathbf{x}) = \sum_{c \in C} \phi_c(\mathbf{x}_c)$$

- In physics, this is called the "Boltzmann distribution".
- In statistics, this is called a log-linear model.

Eric Xing

59

Example: Boltzmann machines



- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for $x_i \in \{-1, +1\}$ or $x_i \in \{0, 1\}$) is called a Boltzmann machine

$$\begin{aligned} P(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \exp\left\{\sum_{ij} \phi_{ij}(x_i, x_j)\right\} \\ &= \frac{1}{Z} \exp\left\{\sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C\right\} \end{aligned}$$

- Hence the overall energy function has the form:

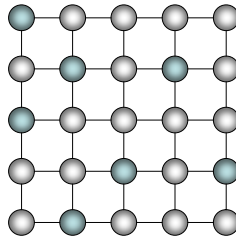
$$H(\mathbf{x}) = \sum_{ij} (x_i - \mu) \Theta_{ij} (x_j - \mu) = (\mathbf{x} - \mu)^T \Theta (\mathbf{x} - \mu)$$

Eric Xing

60

Example: Ising models

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.



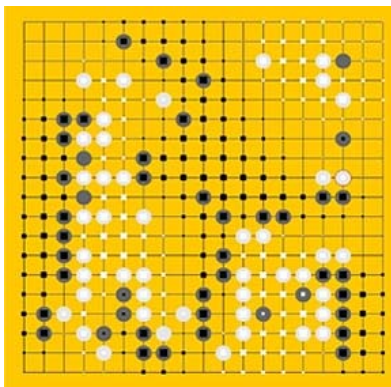
$$p(X) = \frac{1}{Z} \exp \left\{ \sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i \right\}$$

- Same as sparse Boltzmann machine, where $\theta_{ij} \neq 0$ iff i, j are neighbors.
 - e.g., nodes are pixels, potential function encourages nearby pixels to have similar intensities.
- Potts model:** multi-state Ising model.

Eric Xing

61

Application: Modeling Go

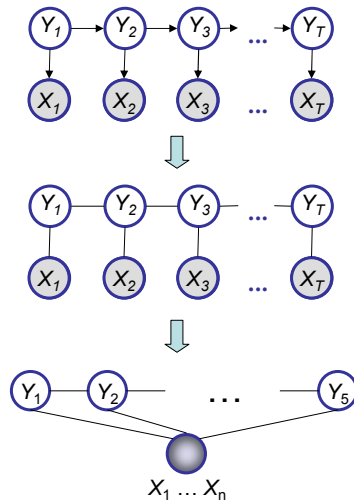


This is the middle position of a Go game.
Overlaid is the estimate for the probability of becoming black or white for every intersection.
Large squares mean the probability is higher.

Eric Xing

62

Example: Conditional Random Fields



- Discriminative

$$p_{\theta}(y | x) = \frac{1}{Z(\theta, x)} \exp \left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- Doesn't assume that features are independent
- When labeling X_i future observations are taken into account

Eric Xing

63

Conditional Models



- Conditional probability $P(\text{label sequence } y \mid \text{observation sequence } x)$ rather than joint probability $P(y, x)$
 - Specify the probability of possible label sequences given an observation sequence
- Allow arbitrary, non-independent features on the observation sequence X
- The probability of a transition between labels may depend on **past** and **future** observations
- Relax strong independence assumptions in generative models

Eric Xing

64

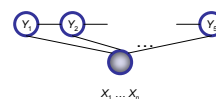
Conditional Distribution



- If the graph $G = (V, E)$ of \mathbf{Y} is a tree, the conditional distribution over the label sequence $\mathbf{Y} = \mathbf{y}$, given $\mathbf{X} = \mathbf{x}$, by fundamental theorem of random fields is:

$$p_{\theta}(\mathbf{y} | \mathbf{x}) \propto \exp \left(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x}) \right)$$

- \mathbf{x} is a data sequence
- \mathbf{y} is a label sequence
- v is a vertex from vertex set V = set of label random variables
- e is an edge from edge set E over V
- f_k and g_k are given and fixed. g_k is a Boolean vertex feature; f_k is a Boolean edge feature
- k is the number of features
- $\theta = (\lambda_1, \lambda_2, \dots, \lambda_n; \mu_1, \mu_2, \dots, \mu_n)$; λ_k and μ_k are parameters to be estimated
- $\mathbf{y}|_e$ is the set of components of \mathbf{y} defined by edge e
- $\mathbf{y}|_v$ is the set of components of \mathbf{y} defined by vertex v



Eric Xing

65

Conditional Distribution (cont'd)



- CRFs use the observation-dependent normalization $Z(\mathbf{x})$ for the conditional distributions:

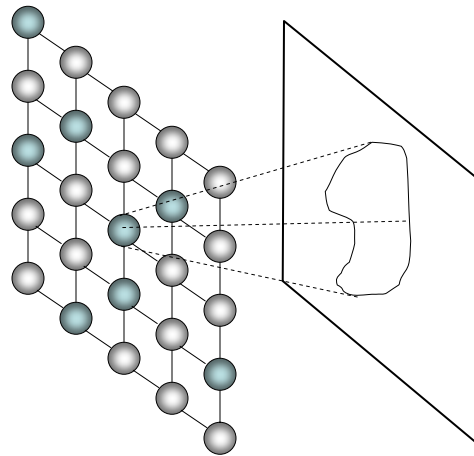
$$p_{\theta}(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x}) \right)$$

- $Z(\mathbf{x})$ is a normalization over the data sequence \mathbf{x}

Eric Xing

66

Conditional Random Fields



$$p_{\theta}(y|x) = \frac{1}{Z(\theta, x)} \exp\left\{\sum_c \theta_c f_c(x, y_c)\right\}$$

- Allow arbitrary dependencies on input
- Clique dependencies on labels
- Use approximate inference for general graphs

Eric Xing

67

Why graphical models



- 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

Eric Xing

68

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**
- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.