

Graphical Models

Aarti Singh

Slides Courtesy: Carlos Guestrin

Machine Learning 10-701/15-781

Nov 10, 2010



MACHINE LEARNING DEPARTMENT



Recitation

- HMMs & Graphical Models
- Strongly recommended!!
- Place: NSH 1507 (Note)
- Time: 5-6 pm

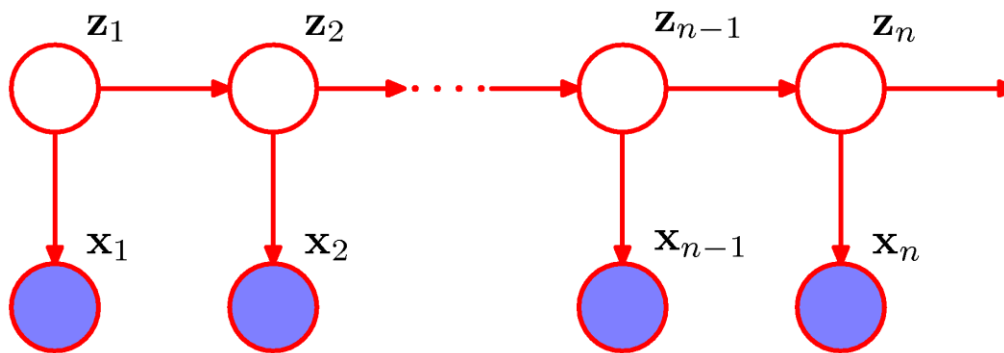


Min

iid to dependent data

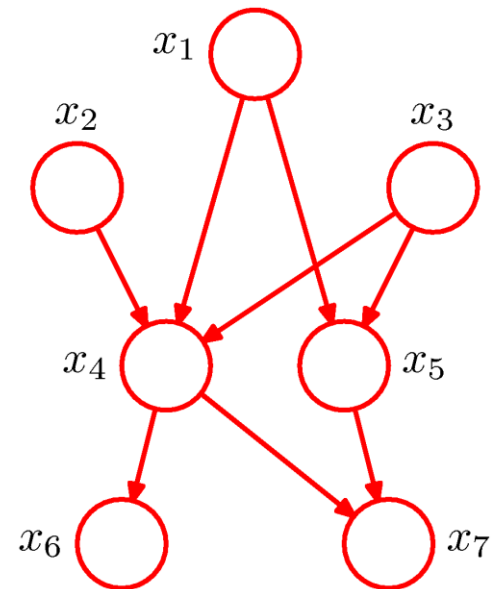
HMM

- sequential dependence



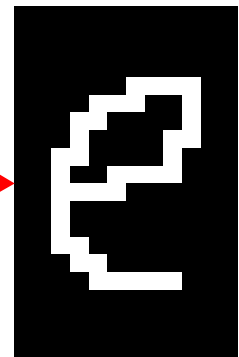
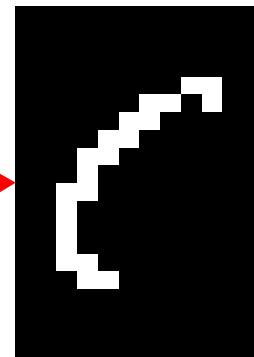
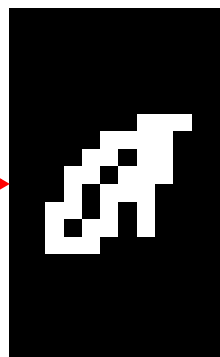
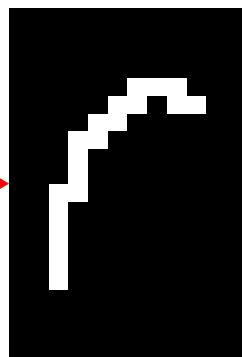
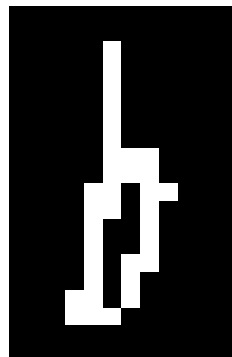
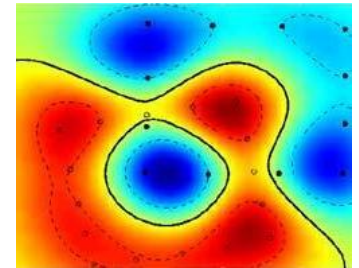
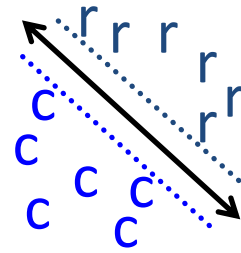
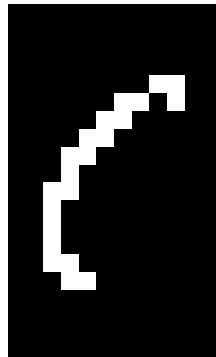
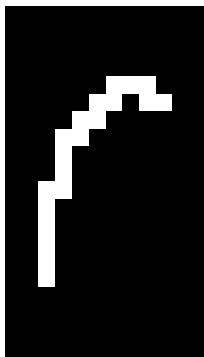
Graphical Models

- general dependence



Applications

- Character recognition, e.g., kernel SVMs



Applications

- Webpage Classification



Sports
Science
News

HOME • PUBLICATIONS • TALKS • PREL • BOOKS • BIOGRAPHY • OUTREACH

Pattern Recognition and Machine Learning

This leading textbook provides a comprehensive introduction to the fields of pattern recognition and machine learning. It is aimed at self-motivated undergraduates or first-year PhD students, as well as researchers and practitioners. No previous knowledge of pattern recognition or machine learning concepts is assumed. This is the first machine learning textbook to include a comprehensive coverage of recent developments such as probabilistic graphical models and deterministic inference methods, and to emphasize a modern Bayesian perspective. It is suitable for courses on machine learning, statistics, computer science, signal processing, computer vision, data mining, and bioinformatics. This hardcover book has 738 pages in full colour, and there are 431 graded exercises (with solutions available below). Extensive support is provided for course instructors.

To view this book go to [downloads](#).

Downloads

- Contents list and sample chapter (Chapter 8: Graphical Models) in PDF format.
- Synopsis of the book.
- Solutions manual for the [www exercises](#) in PDF format (version: 8 September, 2009).
- Data sets.
- Complete set of figures in JPEG, PNG, PDF and EPS formats.
- A PDF file of errata. There are three versions of this. To determine which one to download, look at the bottom of the page opposite the dedication paragraph in your copy of the book. If it says "corrected ... 2009" then download [version 3](#). If it says "corrected ... 2007" then download [version 2](#). Otherwise download [version 1](#).
- The book has been translated into Japanese in two volumes. [Volume 1](#) contains chapters 1-5 plus the appendices, while [Volume 2](#) contains chapters 6-14. Support for the

PEML is Available from

- Amazon
- Amazon UK
- Amazon India

Get • Firefox • Highlight • High use • Reached end of page, continued from top

Machine Learning

10-701/15-781, Fall 2010

[Aarti Singh](#)

Home	People	Lectures	Recitations	Homeworks	Project	Previous material	Table of contents
------	--------	----------	-------------	-----------	---------	-------------------	-------------------

Lectures: Date and Time: Monday and Wednesday, 10:30 - 11:50 am
Location: 7500 Wean Hall

Recitation: Date and Time: Thursday, 5:00 - 6:00 pm
Location: NSH 1507 (normal location), NSH 1507 (Sep 16, 23, 30; Oct 7; Nov 11; Dec 2, 9), GBC 6115 (Nov 4)

Announcements:

- Midterm [exam](#), [solution](#) and [score distribution](#) are now available.
- No recitation tomorrow (Oct 21)
- Midterm exam: Oct 20, Wednesday (In Class), open books, open notes, no internet access
- The [blackboard](#) is now available and we encourage students to use it for discussions related to the course, homework, clarifications etc.
- First day of class is Sept 16. See you in class!
- First day of recitation is Sept 16.
- If you are on the waiting list, [email the instructor](#), and you will be allowed to enroll if there is space and you meet the pre-requisites.

ML MACHINE LEARNING DEPARTMENT

Carnegie Mellon
School of Computer Science

PROSPECTIVE STUDENTS | CURRENT STUDENTS | RESEARCH | PEOPLE | SEMINARS | NEWS | HONORS & AWARDS | CONTACT US

A computer that teaches itself. That's the idea behind MLL, the Machine Learning Laboratory. MLL Professor and Department Head, Tom Mitchell, leads the team of Carnegie Mellon scientists creating the one of a kind computer system that's learning how to read the web. [Read more](#)

The Never Ending Language Learning (NELL) project, headed by Tom Mitchell, chief of the Machine Learning Department, is featured in the Oct. 5 issue of the New York Times. The story by Steve Lohr is part of the newspaper's series on artificial intelligence. "Smarter Than You Think," NELL has been running 24/7 since the beginning of the year, constantly reading Web pages and extracting more than 300,000 facts to date, as it

Aarti Singh

Assistant Professor

[Machine Learning Department](#)
[Carnegie Mellon University](#)

Office: 5207 Gates-Hillman Center
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Ph: (412) 268-4266 (O)

Email: narsingh@ATYum.edu



RESEARCH:

The nature of information processing has evolved dramatically over the years as we try to explore increasingly complex and diverse systems ranging from the Internet and wireless networks, to the human brain and genome. I am interested in developing techniques at the intersection of statistical machine learning and signal processing that can adaptively learn and exploit the low-dimensional information structure inherent in high-dimensional systems for efficient inference. The primary thrust of my research is on bridging the gap between theoretically optimal and practically useful methods with applications that include wireless and sensor networks, Internet data analysis and bioinformatics.

Current Projects:

- Resource-constrained data collection and fusion for identifying weak distributed patterns in networks (Sponsored by AFOSR)
- Stable, Robust and Active methods for nonparametric Clustering

Applications

- Speech recognition
- Diagnosis of diseases
- Study Human genome
- Robot mapping
- Modeling fMRI data
- Fault diagnosis
- Modeling sensor network data
- Modeling protein-protein interactions
- Weather prediction
- Computer vision
- Statistical physics
- Many, many more ...

Graphical Models

- Key Idea:
 - Conditional independence assumptions useful
 - but Naïve Bayes is extreme!
 - Graphical models express sets of conditional independence assumptions via graph structure
 - Graph structure plus associated parameters define joint probability distribution over set of variables/nodes
- Two types of graphical models:
 - Directed graphs (aka Bayesian Networks)
 - Undirected graphs (aka Markov Random Fields)

Topics in Graphical Models

- Representation
 - Which joint probability distributions does a graphical model represent?
- Inference
 - How to answer questions about the joint probability distribution?
 - Marginal distribution of a node variable
 - Most likely assignment of node variables
- Learning
 - How to learn the parameters and structure of a graphical model?

Conditional Independence

- X is **conditionally independent** of Y given Z:
probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

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

- Also to:

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

Directed - Bayesian Networks

- Representation

- Which joint probability distributions does a graphical model represent?

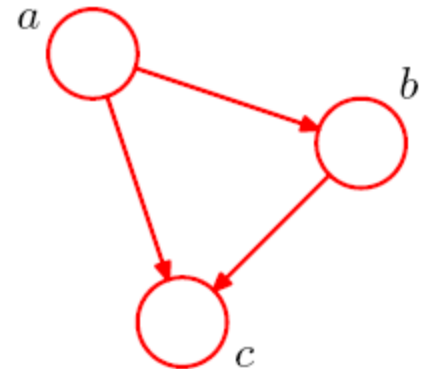
For any arbitrary distribution,

Chain rule:

$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

More generally:

$$p(\mathbf{X}) = \prod_{i=1}^n p(X_i | X_{i-1}, \dots, X_1)$$



Fully connected
directed graph
between X_1, \dots, X_n

Directed - Bayesian Networks

- Representation

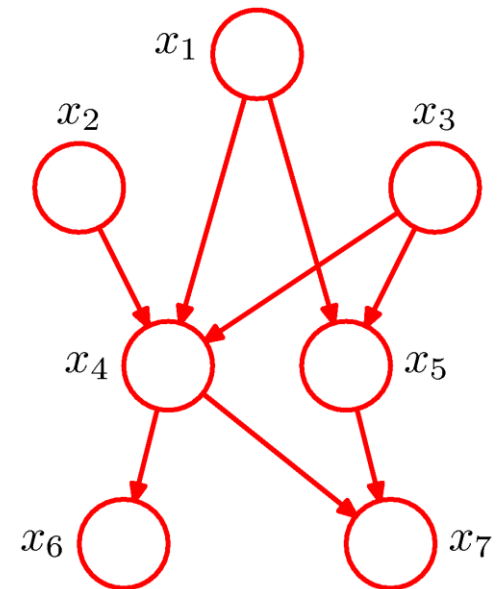
- Which joint probability distributions does a graphical model represent?

Absence of edges in a graphical model conveys useful information.

$$p(x_1, x_2, \dots, x_6) =$$

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$

$$p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



Directed - Bayesian Networks

- Representation

- Which joint probability distributions does a graphical model represent?

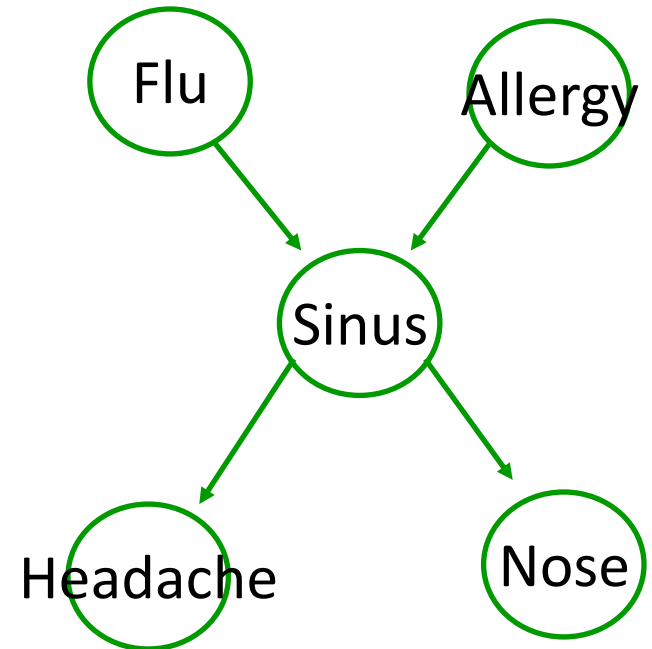
BN is a directed acyclic graph (DAG) that provides a compact representation for joint distribution

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents)

Bayesian Networks Example

- Suppose we know the following:
 - The flu causes sinus inflammation
 - Allergies cause sinus inflammation
 - Sinus inflammation causes a runny nose
 - Sinus inflammation causes headaches

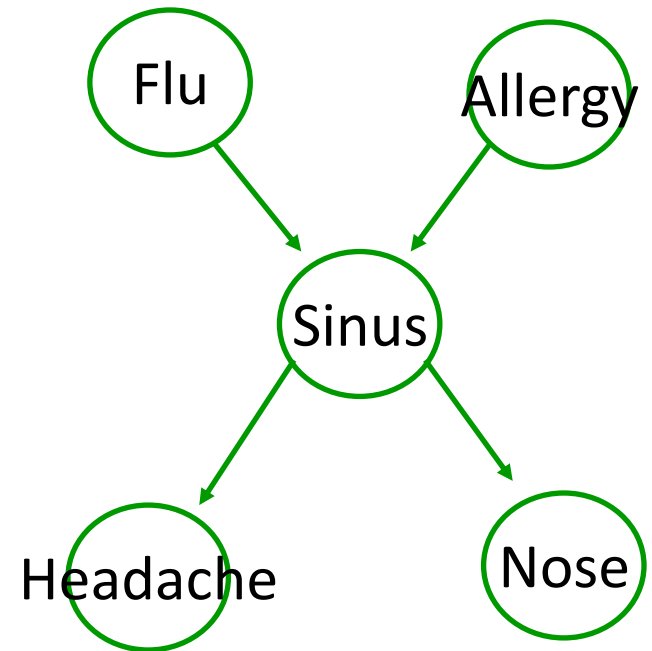


- Causal Network
- Local Markov Assumption: If you have no sinus infection, then flu has no influence on headache (flu causes headache but only through sinus)

Markov independence assumption

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents)

	parents	non-desc	assumption
S	F,A	-	-
H	S	F,A,N	$H \perp \{F,A,N\} S$
N	S	F,A,H	$N \perp \{F,A,H\} S$
F	-	A	$F \perp A$
A	-	F	$A \perp F$



Markov independence assumption

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents)

Joint distribution:

$$P(F, A, S, H, N)$$

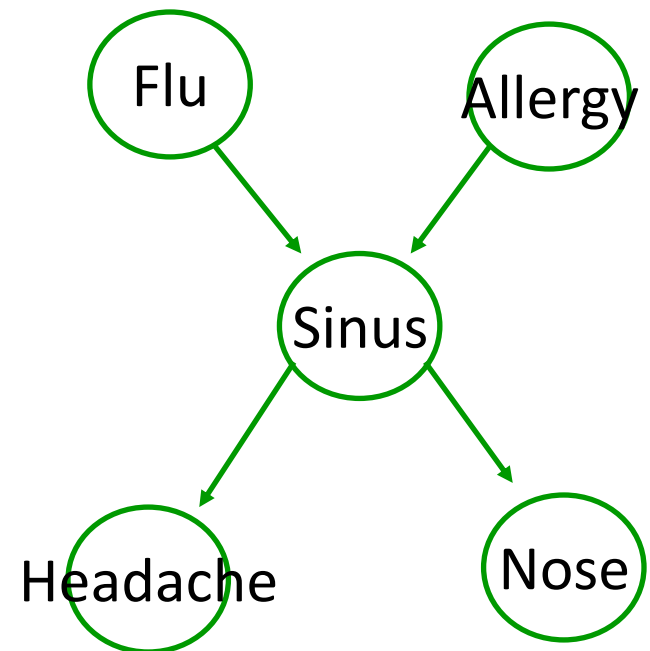
$$= P(F) P(F|A) P(S|F,A) P(H|S,F,A) P(N|S,F,A,H)$$

Chain rule

$$= P(F) P(A) P(S|F,A) P(H|S) P(N|S)$$

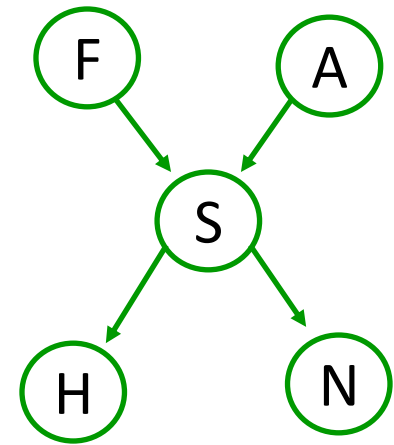
Markov Assumption

$$F \perp A, \quad H \perp \{F,A\} | S, \quad N \perp \{F,A,H\} | S$$



How many parameters in a BN?

- Discrete variables X_1, \dots, X_n
- Directed Acyclic Graph (DAG)
 - Defines parents of X_i , \mathbf{Pa}_{X_i}
- CPTs (Conditional Probability Tables)
 - $P(X_i | \mathbf{Pa}_{X_i})$



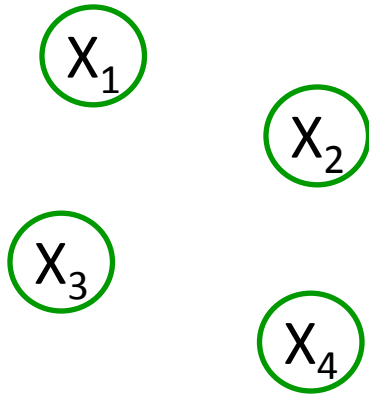
E.g. $X_i = S$, $\mathbf{Pa}_{X_i} = \{F, A\}$

	F=f, A=f	F=t, A=f	F=f, A=t	F=t, A=t
S=t	0.9	0.8	0.7	0.3
S=f	0.1	0.2	0.3	0.7

n variables, K values, max d parents/node $O(nK \times K^d)$

Two (trivial) special cases

Fully disconnected graph



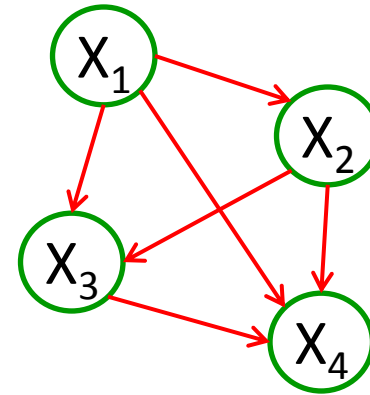
X_i

parents: ϕ

non-descendants: $X_1, \dots, X_{i-1},$
 X_{i+1}, \dots, X_n

$X_i \perp X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$

Fully connected graph



X_i

parents: X_1, \dots, X_{i-1}

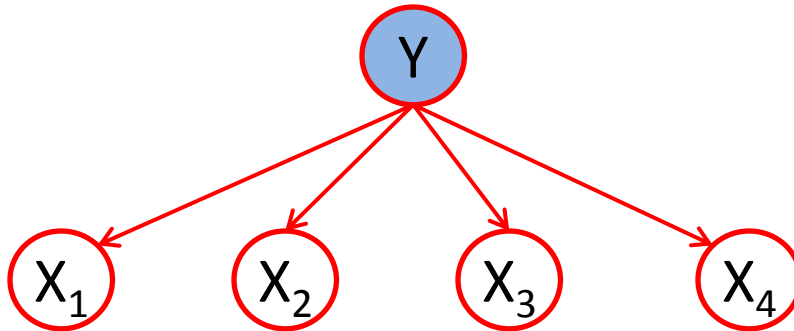
non-descendants: ϕ

No independence
assumption

Bayesian Networks Example

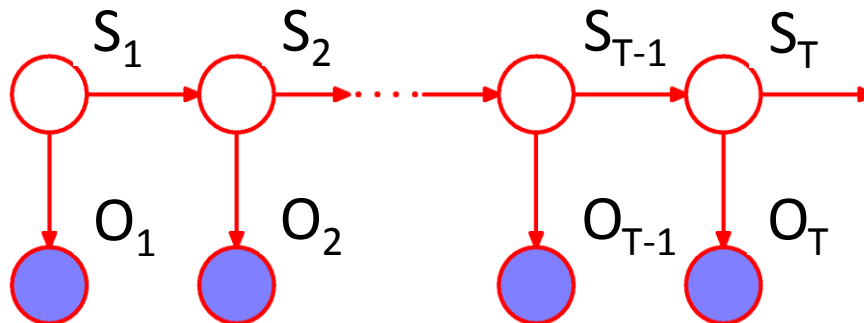
- Naïve Bayes

$$X_i \perp X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n | Y$$



$$P(X_1, \dots, X_n, Y) = P(Y)P(X_1 | Y) \dots P(X_n | Y)$$

- HMM



$$p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = p(S_1) \prod_{t=2}^T p(S_t | S_{t-1}) \prod_{t=1}^T p(O_t | S_t)$$

Explaining Away

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents)

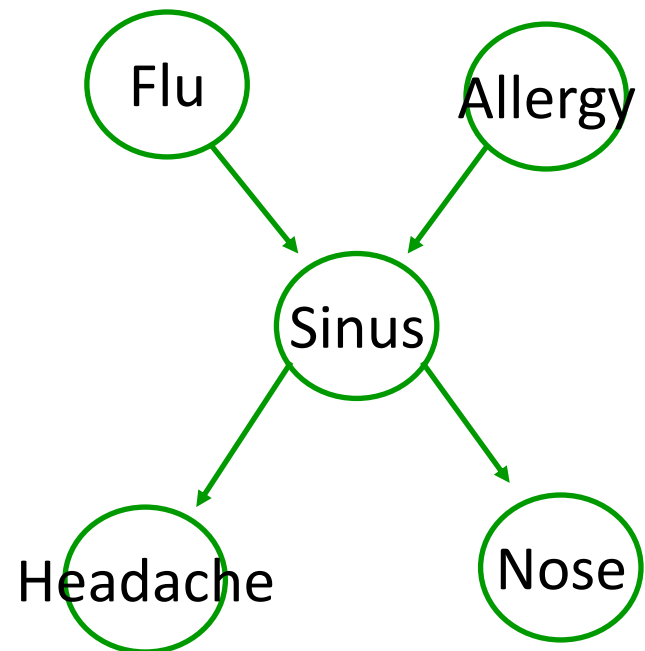
$$F \perp A \quad P(F|A=t) = P(F)$$

$$F \perp A|S? \quad \text{No!}$$
$$P(F|A=t, S=t) = P(F|S=t)?$$

$P(F=t|S=t)$ is high,
but $P(F=t|A=t, S=t)$ not as high
since $A = t$ explains away $S=t$

In fact, $P(F=t|A=t, S=t) < P(F=t|S=t)$

$$F \perp A|N? \quad \text{No!}$$

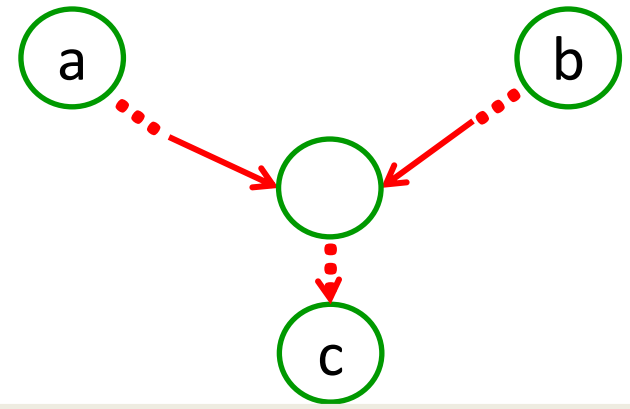
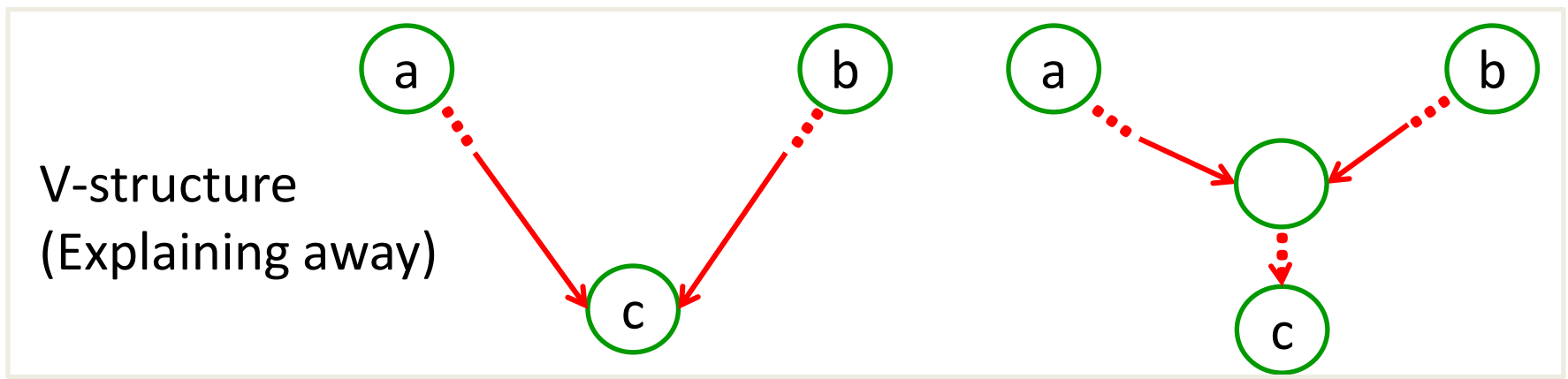
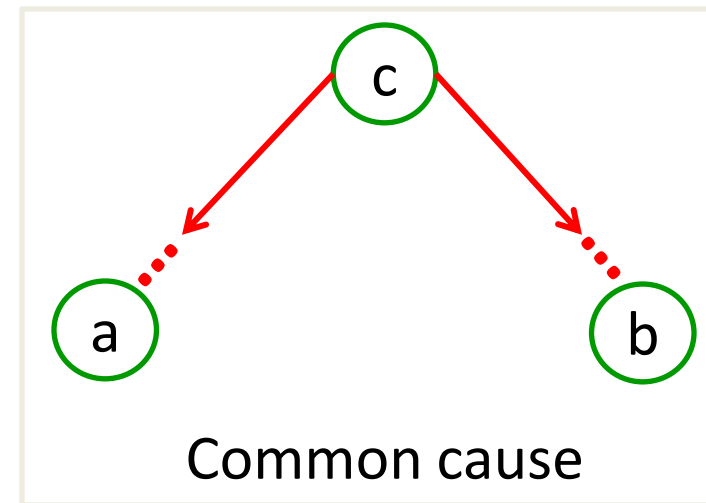
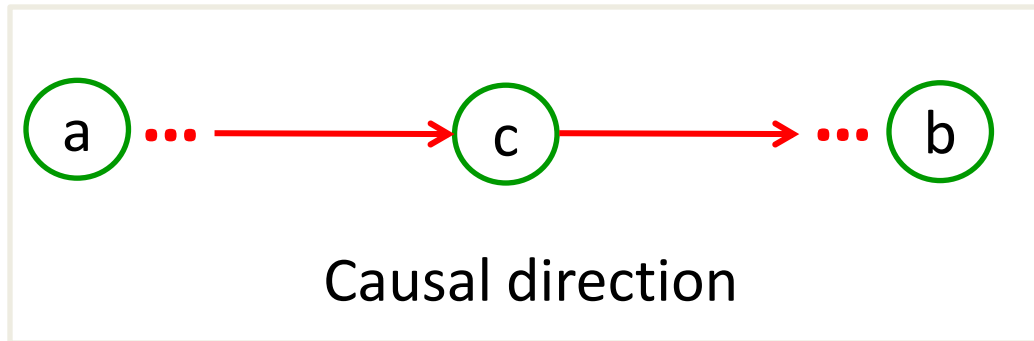


Independencies encoded in BN

- We said: All you need is the local Markov assumption
 - $(X_i \perp \text{NonDescendants}_{X_i} \mid \mathbf{Pa}_{X_i})$
- But then we talked about other (in)dependencies
 - e.g., explaining away
- What are the independencies encoded by a BN?
 - Only assumption is local Markov
 - But many others can be derived using the algebra of conditional independencies!!!

D-separation

- a is D-separated from b by $c \equiv a \perp b | c$
- Three important configurations

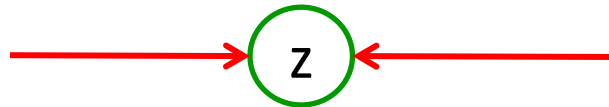


D-separation

- A, B, C – non-intersecting set of nodes
- A is D-separated from B by C $\equiv A \perp B | C$
if all paths between nodes in A & B are “blocked”
i.e. path contains a node z such that either



and z in C, OR



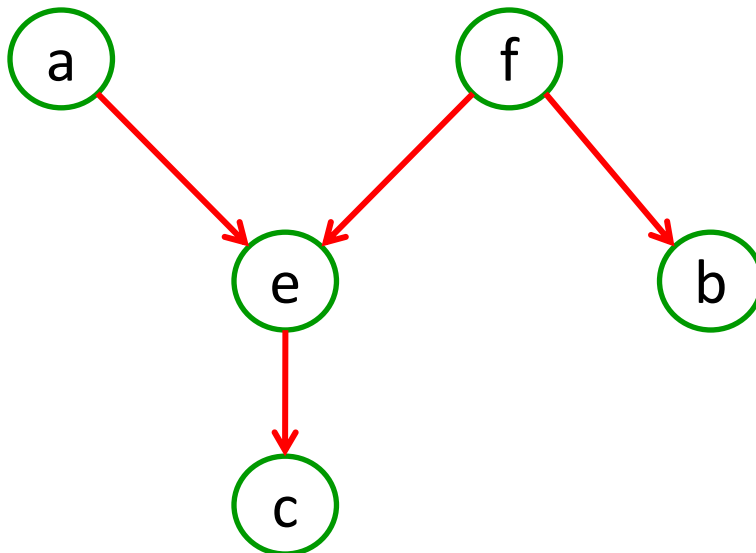
and neither z nor any of its descendants is in C.

D-separation Example

A is D-separated from B by C if every path between A and B contains a node z such that either



or → (z) ← And neither z nor its descendants are in C



$a \perp b \mid f ?$

Yes, Consider $z = f$ or $z = e$

$a \perp b \mid c ?$

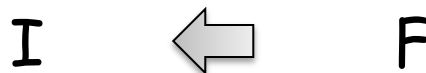
No, Consider $z = e$

Representation Theorem

- Set of distributions that factorize according to the graph - \mathcal{F}
- Set of distributions that respect conditional independencies implied by d-separation properties of graph – \mathcal{I}



Important because: **Given independencies of P can get BN structure G**



Important because: **Read independencies of P from BN structure G**

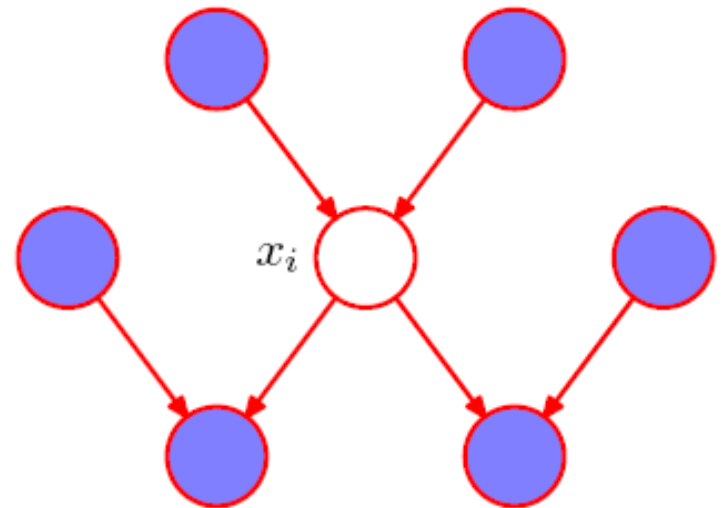
Markov Blanket

- Conditioning on the Markov Blanket, node i is independent of all other nodes.

$$p(\mathbf{X}_i | \mathbf{X}_{\{j \neq i\}}) = \frac{p(x_1, \dots, x_n)}{\sum_i p(x_1, \dots, x_n)} = \frac{\prod_k p(x_k | pa(x_k))}{\sum_i \prod_k p(x_k | pa(x_k))}$$

Only terms that remain are the ones which involve i

$$p(x_i | pa(x_i)) \quad p(x_k | pa(x_k) \ni i)$$



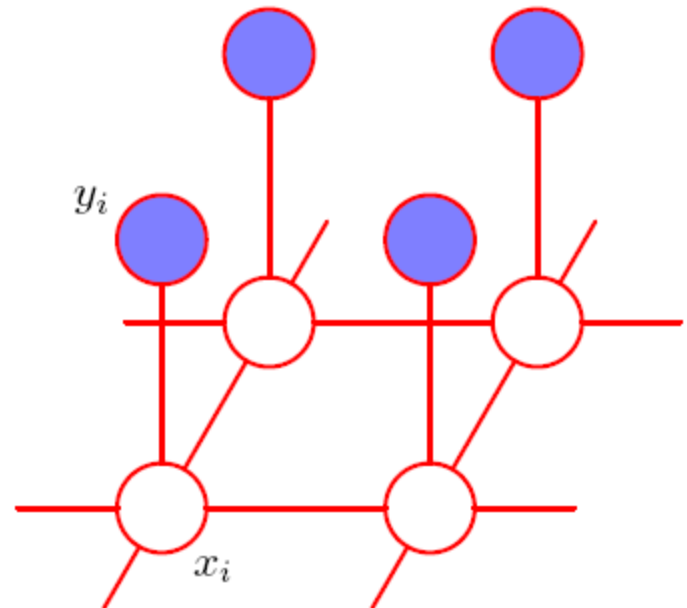
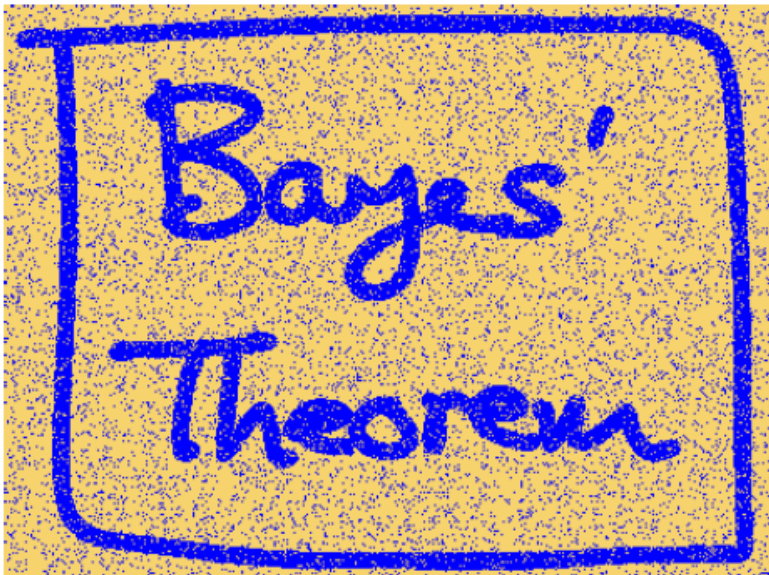
- Markov Blanket of node i - Set of parents, children and co-parents of node i

Undirected – Markov Random Fields

- Popular in statistical physics and computer vision communities
- Example – Image Denoising

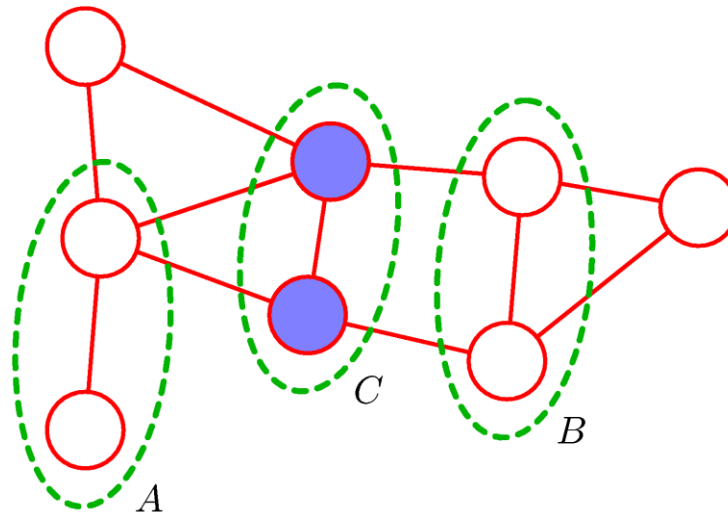
x_i – value at pixel i

y_i – observed noisy value



Conditional Independence properties

- No directed edges
- Conditional independence \equiv graph separation
- A, B, C – non-intersecting set of nodes
- $A \perp B | C$ if all paths between nodes in A & B are “blocked”
i.e. path contains a node z in C .



Factorization

- Joint distribution factorizes according to the graph

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

\mathcal{C} is the set of maximal cliques in the graph

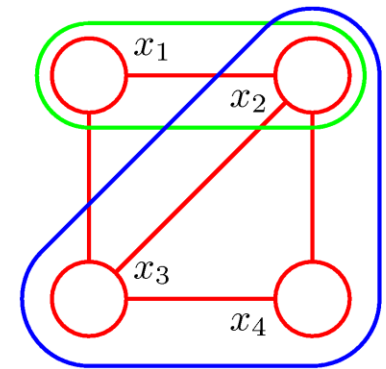
$\psi_C(x_C)$ is a potential function on the clique x_C

└─ Arbitrary positive function

normalization factor

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

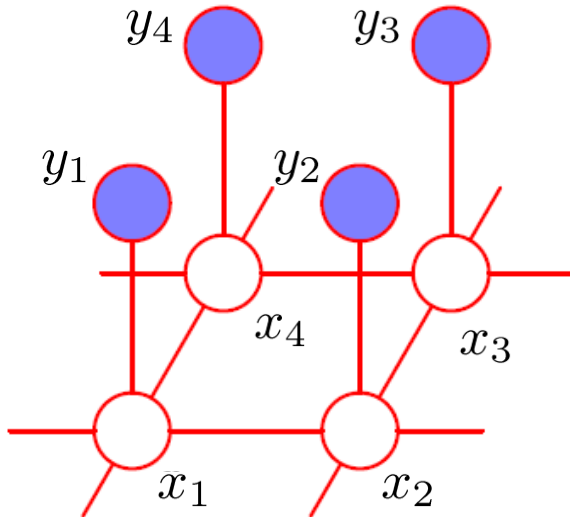
typically NP-hard to compute



Clique, $x_C = \{x_1, x_2\}$

Maximal clique
 $x_C = \{x_2, x_3, x_4\}$

MRF Example



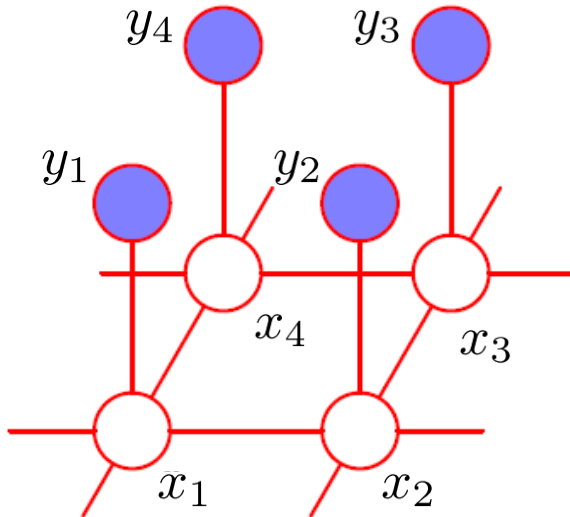
$$P(x, y) \propto \Psi(x_1, x_2) \Psi(x_1, x_3) \Psi(x_2, x_4) \Psi(x_3, x_4) \prod_{i=1}^4 \Psi(x_i, y_i)$$

Often $\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$

└→ Energy of the clique (e.g. lower if variables in clique take similar values)

$$p(\mathbf{x}) = \prod_{C \in \mathcal{C}} \exp\{-E(\mathbf{x}_C)\} = \exp\left\{-\sum_{C \in \mathcal{C}} E(\mathbf{x}_C)\right\}$$

MRF Example



Ising model:

cliques are edges $x_C = \{x_i, x_j\}$
binary variables $x_i \in \{-1, 1\}$

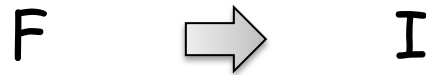
$$\psi_C(\mathbf{x}_C) = \exp\{\underbrace{\beta x_i x_j}_{\substack{1 \text{ if } x_i = x_j \\ -1 \text{ if } x_i \neq x_j}}\}$$

$$p(\mathbf{x}) = \prod_{(i,j) \in E} \exp\{\beta x_i x_j\} = \exp\left\{ \sum_{(i,j) \in E} \beta x_i x_j \right\}$$

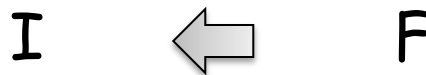
Probability of assignment is higher if neighbors x_i and x_j are same

Hammersley-Clifford Theorem

- Set of distributions that factorize according to the graph - \mathcal{F}
- Set of distributions that respect conditional independencies implied by graph-separation – \mathcal{I}



Important because: **Given independencies of P can get MRF structure G**



Important because: **Read independencies of P from MRF structure G**

What you should know...

- Graphical Models: Directed Bayesian networks, Undirected Markov Random Fields
 - A compact **representation** for large probability distributions
 - Not an algorithm
- Representation of a BN, MRF
 - Variables
 - Graph
 - CPTs
- Why BNs and MRFs are useful
- D-separation (conditional independence) & factorization