Probabilistic Graphical Models and Their Applications

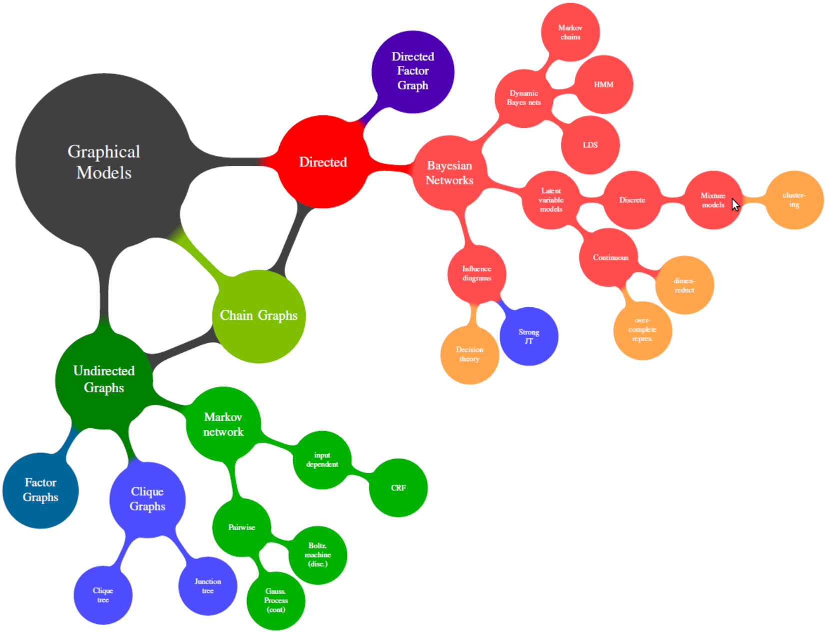
Gerard Pons-Moll and Paul Swoboda

Max Planck Institute for Informatics

slides adapted from Peter Gehler

October 24, 2018

max planck institut
informatik


Graphical Models
Directed
Directed Factor Graph
Bayesian Networks
Dynamic Bayes nets
Markov chains
HMM
LDS
Latent variable models
Discrete
Mixture models
cluster-ing
Continuous
dimen-reduct
over-complete repres.
Influence diagrams
Decision theory
Strong JT
Chain Graphs
Undirected Graphs
Markov network
input dependent
CRF
Pairwise
Boltz. machine (disc.)
Gauss. Process (cont)
Factor Graphs
Clique Graphs
Clique tree
Junction tree

Today's Topics

- Directed Graphical Models
  - Belief Networks or Bayesian Networks
- Some Graph Terminology
- Undirected Graphical Models
  - Markov Networks or Markov Random Fields

Reading Material:

- D. Barber, *Bayesian Reasoning and Machine Learning*, Sections: 3.1, 3.2, 3.3, 4.1, 4.2
- C. Bishop, *Pattern Recognition and Machine Learning*, Chapter 8.1, 8.2, 8.3

Some Notation for Random Variables

Modeling Your Knowledge

- *Events* (random variables) - notation: (X, Y, Z)
  - e.g. it rained, the street is wet, you are older than 23
  - may affect each other
  - may be (conditionally) independent
- We will use graphs to encode this information
  - event is a vertex
  - "dependence is an edge"
- This leads to a "graphical model" that captures and expresses relations among variables
  - Think of graphical models as a modeling language
- Algorithms for learning and inference in these graph based representations exists

Probability Variables – Notation

- Random variables X, Y, and Z

Chain Rule

$$p(X,Y) = p(X|Y)p(Y)$$

$$\begin{aligned} p(X,Y,Z) &= p(X|Y,Z)p(Y,Z) \\ &= p(X|Y,Z)p(Y|Z)p(Z) \end{aligned}$$

Bayes' Theorem

$$p(X|Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(Y|X)p(X)}{p(Y)}$$

Probability Variables – Notation

- Two random variables X and Y

Independence

X and Y are independent if

$$p(X, Y) = p(X)p(Y)$$

- Provided $p(X) \neq 0, p(Y) \neq 0$ this is equivalent with

$$p(X \mid Y) = p(X) \Leftrightarrow p(Y \mid X) = p(Y) \tag{1}$$

Probability Variables – Notation

- Sets of random variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$

Conditional independence

$\mathcal{X}$ and $\mathcal{Y}$ are independent provided we know the state of $\mathcal{Z}$ if $p(\mathcal{X}, \mathcal{Y} \mid \mathcal{Z}) = p(\mathcal{X} \mid \mathcal{Z})p(\mathcal{Y} \mid \mathcal{Z})$ for all states of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$.
They are conditionally independent given $\mathcal{Z}$

- For conditional independence we write

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} \tag{2}$$

- And thus we write for (unconditional) independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \emptyset \text{ or shorter } \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \tag{3}$$

Probability Variables – Notation

- Similarly we write

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} \mid \mathcal{Z} \tag{4}$$

  for conditionally dependent sets of random variables

- and

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} \mid \emptyset \ \text{ or shorter } \ \mathcal{X} \top\!\!\!\top \mathcal{Y} \tag{5}$$

  for unconditionally dependent random variables

Dependent or Not?

- a is independent of b ($a \perp\!\!\!\perp b$)
- b is independent of c ($b \perp\!\!\!\perp c$)
- c and a are ... ?

- Consider this distribution

$$p(a, b, c) = p(b)p(a, c) \tag{6}$$

- $a \perp\!\!\!\perp b$ and $b \perp\!\!\!\perp c$ because:

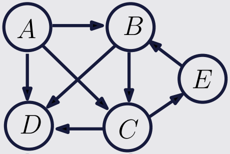
$$\begin{aligned} p(a, b) &= p(b) \sum_c p(a, c) = p(b)p(a) & (7) \\ p(c, b) &= p(b) \sum_a p(a, c) = p(b)p(c) & (8) \end{aligned}$$

- So a and c may or may not be independent

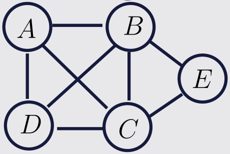
Graph Definitions

- A graph consists of *vertices* and *edges*

Graph



A directed graph – directed edges.
Bayesian Networks
(or Belief Networks)



An undirected graph – undirected edges.
Markov Random Fields
(or Markov Networks)

Belief Networks or Bayesian Networks (BN)

An Example

- Mr. Holmes leaves his house
  - He sees that the lawn in front of his house is wet
  - This can have two reasons: he left the sprinkler turned on or it rained during the night.
  - Without any further information the probability of both events increases
- Now he also observes that his neighbour's lawn is wet
  - This lowers the probability that he left his sprinkler on. This event is *explained away*

Example Continued

- Let's formalize:
- There are several random variables
  - $R \in \{0,1\}$, $R = 1$ means it has been raining
  - $S \in \{0,1\}$, $S = 1$ means sprinkler was left on
  - $N \in \{0,1\}$, $N = 1$ means neighbour's lawn is wet
  - $H \in \{0,1\}$, $H = 1$ means Holmes' lawn is wet
- How many states to be specified?

$$p(R,S,N,H) = \underbrace{p(H \mid R,S,N)}_{2^3=8} \underbrace{p(N \mid R,S)}_{2^2=4} \underbrace{p(R \mid S)}_{2} \underbrace{p(S)}_{1}$$

- $8 + 4 + 2 + 1 = 15$ numbers needed to specify all probabilities
- In general $2^n - 1$ for binary states only

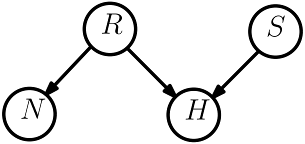
Example – Conditional Independence

- As a modeler of this problem we have prior knowledge of causal dependencies
- **H**olmes' grass, **N**eighbour's grass, **R**ain, **S**prinkler
- $p(H \mid R, S, N) = p(H \mid R, S)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(R \mid S) = p(R)$
- In effect our model becomes

$$p(R, S, N, H) = \underbrace{p(H \mid R, S)}_{4}\underbrace{p(N \mid R)}_{2}\underbrace{p(R)}_{1}\underbrace{p(S)}_{1}$$

- How many states? 8

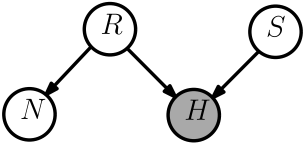
This Example as a Belief Network



$$p(R, S, N, H) = p(H \mid R, S)p(N \mid R)p(R)p(S)$$

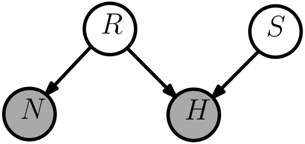
- This is called a directed graphical model or belief network

This example as a Belief Network



- This is called a directed graphical model or belief network
- Observed variables are drawn shaded
  - observing the wet grass

This example as a Belief Network



- This is called a directed graphical model or belief network
- Observed variables are drawn shaded
  - observing the wet grass
  - observing the neighbours wet grass

Example – Inference

- The most pressing question is: was the sprinkler on?
  - in other words what is $p(S = 1 \mid H = 1)$?
- First we need to specify the eight states (conditional probability table = CPT)

$$
\begin{gathered}
p(R = 1) = 0.2, \qquad p(S = 1) = 0.1 \\
p(N = 1 \mid R = 1) = 1, \qquad p(N = 1 \mid R = 0) = 0.2 \\
p(H = 1 \mid R = 1, S) = 1, \qquad p(H = 1 \mid R = 0, S = 1) = 0.9 \\
p(H = 1 \mid R = 0, S = 0) = 0
\end{gathered}
$$

- $p(S = 1 \mid H = 1) = \ldots = 0.3382$
- $p(S = 1 \mid H = 1, N = 1) = \ldots = 0.1604$ (explained away)

Belief Networks

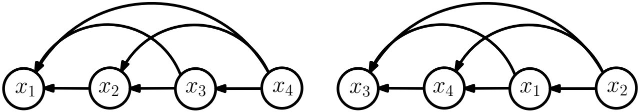
Belief network

A belief network is a distribution of the form

$$p(x_1, \dots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)), \tag{9}$$

where $pa(x)$ denotes the parental variables of x

Different Factorizations



- Two factorizations of four variables:

$$
\begin{aligned}
p(x_1, x_2, x_3, x_4) &= p(x_1 \mid x_2, x_3, x_4) p(x_2 \mid x_3, x_4) p(x_3 \mid x_4) p(x_4) \\
p(x_1, x_2, x_3, x_4) &= p(x_3 \mid x_1, x_2, x_4) p(x_4 \mid x_1, x_2) p(x_1 \mid x_2) p(x_2)
\end{aligned}
$$

- Any distribution can be written in such a cascade form as a belief network (using chain rule)
- With independence assumptions the factorization often becomes simpler

Belief Networks

- Structure of the DAG corresponds to a set of conditional independence assumptions
  - which parents are sufficient (are the causes) to specify the CPT
  - for completeness we need to specify all $p(x \mid pa(x))$
- This does **not** mean non-parental variables have no influence:

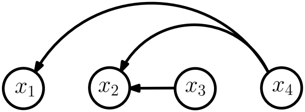
$$p(x_1 \mid x_2)p(x_2 \mid x_3)p(x_3) \tag{10}$$

with DAG $x_1 \leftarrow x_2 \leftarrow x_3$ does **not** imply (Exercise)

$$p(x_2 \mid x_1, x_3) = p(x_2 \mid x_3) \tag{11}$$

Conditional Independence

- Important task:
  - given graph, read of conditional independence statements
- Question:
  - are x_1 and x_2 conditionally independent given x_4 $(x_1 \perp\!\!\!\perp x_2 \mid x_4)$?
  - and what about $x_1 \perp\!\!\!\perp x_2 \mid x_3$?



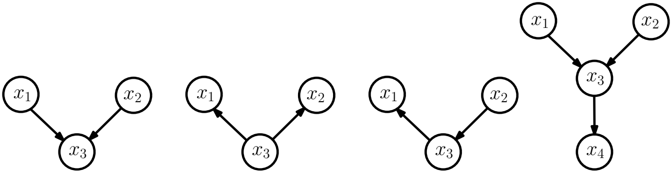
- how to automate?

Collisions

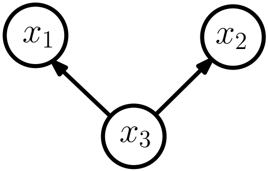
Collision

Given a path from node x to y, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

- Let's check these for colliders:



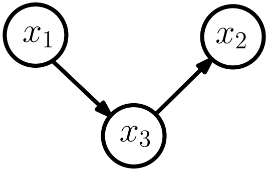
Collider and Conditional Independence



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_1 \mid x_3)p(x_2 \mid x_3)p(x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1 \mid x_3)
\end{aligned}
$$

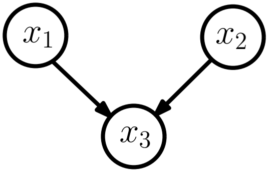
Collider and Conditional Independence



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_3 \mid x_1)p(x_1)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1, x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1 \mid x_3)
\end{aligned}
$$

Collider and Conditional Independence



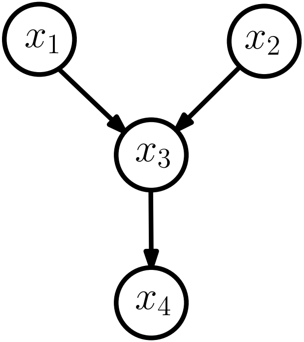
- x_3 a collider ? yes
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? no! (explaining away)

$$\begin{aligned} p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\ &= p(x_1)p(x_2)\underbrace{p(x_3 \mid x_1, x_2)/p(x_3)}_{\neq 1 \text{ in general}} \end{aligned}$$

- $x_1 \perp\!\!\!\perp x_2$? yes

$$p(x_1, x_2) = \sum_{x_3} p(x_3 \mid x_1, x_2)p(x_1)p(x_2) = p(x_1)p(x_2)$$

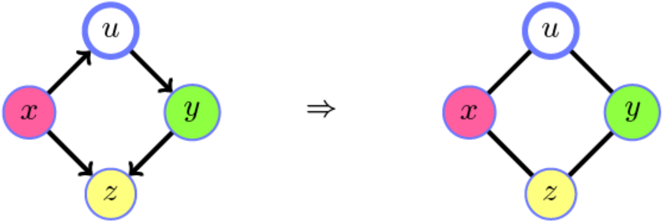
Collider and Conditional Independence



- x_3 a collider ? yes ($x_1 \to x_2$), no ($x_1 \to x_4$)
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_4$? maybe

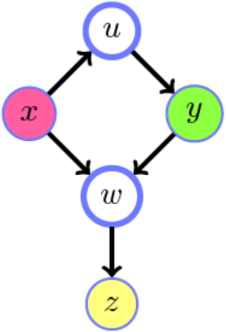
- x_1 and x_2 are "graphically" dependent on x_4
  - There are distributions with this DAG with $x_1 \perp\!\!\!\perp x_2 \mid x_4$ and those with $x_1 \top\!\!\!\top x_2 \mid x_4$
- BN good for representing independence but not good for representing dependence!

Graphical Manipulations to Check for Independence

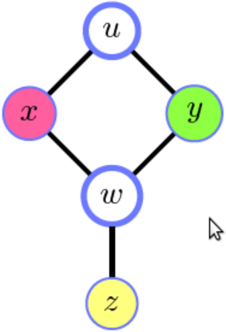


- Question: $x \perp\!\!\!\perp y | z$?
- White nodes are not in the conditioning set
- if z is collider, keep undirected links between neighbours

Graphical Manipulations to Check for Independence



$\Rightarrow$

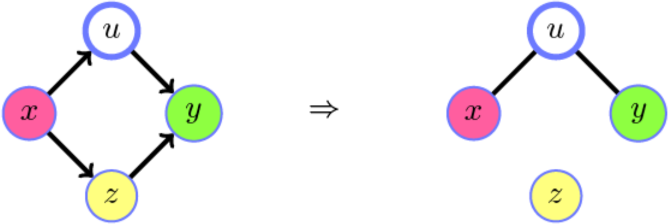


- if z is descendant of a collider (here w), keep links

Graphical Manipulations to Check for Independence

- if a collider is not in the conditioning set (here u): cut the links
- this path is blocked

Graphical Manipulations to Check for Independence

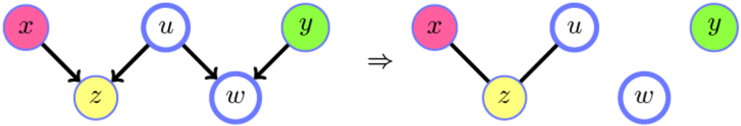


- if z is non-collider but in the conditioning set, cut the links
- this path is blocked

Graphical Manipulations to Check for Independence

- Result of the previous operations
- no path that could introduce dependence
- Hence $x \perp\!\!\!\perp y \mid z$ (both paths blocked)

Graphical Manipulations to Check for Independence



- Question: $x \perp\!\!\!\perp y \mid z$?
- yes

D-Separation

- Let's formalize:
- We have all tools to check for conditional independence $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ in any belief network

d separation

For every $x \in \mathcal{X}, y \in \mathcal{Y}$ check every path U between x and y.
A path is blocked if there is a node w on U such that either

1. w is a collider and neither w nor any descendant is in $\mathcal{Z}$
2. w is not a collider on U and w is in $\mathcal{Z}$

If all such paths are blocked then $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$

D-Connectedness

- And the opposite:

d-connected

$\mathcal{X}$ and $\mathcal{Y}$ are d-connected by $\mathcal{Z}$ if and only if they are not d-separated by $\mathcal{Z}$.

Markov Equivalence

Markov equivalence

Two graphs are Markov equivalent if they represent the same set of conditional independence statements. (holds for directed and undirected graphs)

skeleton

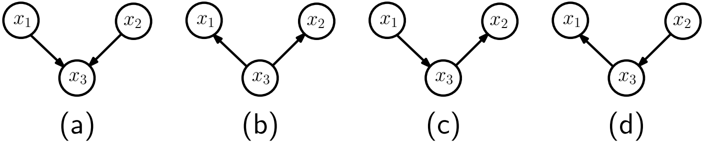
Graph resulting when removing all arrows of edges

immorality

Parents of a child with no connection

- Markov equivalent $\Leftrightarrow$ same skeleton and same set of immoralities

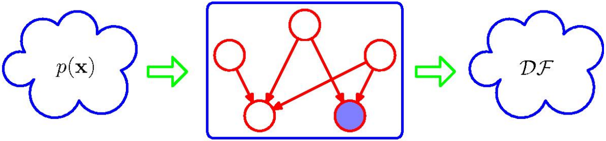
Three Variable Graphs Revisited



- All have the same skeleton
- (b,c,d) have no immoralities
- (a) has immorality (x_1, x_2) and is thus not equivalent

$$
\begin{aligned}
\text{(d)} : p(x_1|x_3)p(x_3|x_2)p(x_2) &= p(x_1|x_3)p(x_2, x_3) \\
&= p(x_1|x_3)p(x_3)p(x_2|x_3) \text{ equals to (b)} \\
&= p(x_1, x_3)p(x_2|x_3) \\
&= p(x_3|x_1)p(x_1)p(x_2|x_3) \text{ equals to (c)}
\end{aligned}
$$

Filter View of a Graphical Model



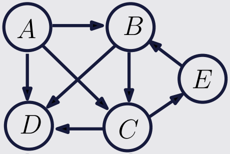
- Belief network (also undirected graph) implies a list of conditional independences
- Regard as filter:
  - only distributions that satisfy all conditional independences are allowed to pass
- One graph describes a whole family of probability distributions
- Extremes:
  - Fully connected, no constraints, all p pass
  - no connections, only product of marginals may pass

Graph Definitions

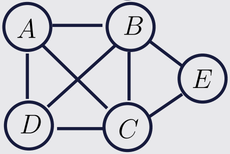
Graph Definitions

- A graph consists of *vertices* and *edges*

Graph



A directed graph – directed edges.
Bayesian Networks (or Belief Networks)



An undirected graph – undirected edges.
Markov Random Fields

Graph Definitions

Path, Ancestor, Descendant

- A path $A \to B$ is a sequence of vertices

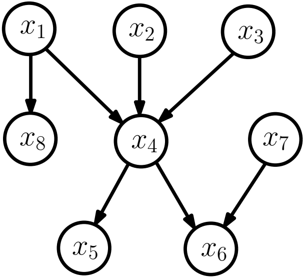
$$A_0 = A, A_1, \ldots, A_{N-1}, A_N = B \tag{12}$$

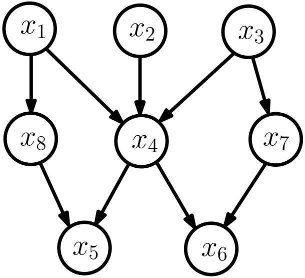
  with (A_n, A_{n+1}) an edge in the graph.
- In directed graphs, the vertices A such that $A \to B$ and $B \not\to A$ are the ancestors of B.
- Vertices B such that $A \to B$ and $B \not\to A$ are the descendants of A.

Graph Definitions

Directed Acyclic Graph (DAG)

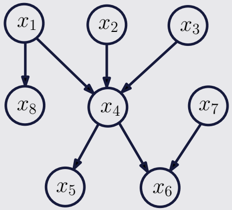
A DAG is a graph G with directed edges between the vertices such that by following a directed path of vertices no path will revisit a vertex.





Graph Definitions

The Family



The parents of x_4 are $pa(x_4) = \{x_1, x_2, x_3\}$. The children of x_4 are $ch(x_4) = \{x_5, x_6\}$.
The family of x_4 are the node itself, its parents and children.
The Markov blanket is the node, its parents, the children and the parents of the children. In this case $x_1, \ldots, x_7$

- Why DAGs? Structure prevents circular (cyclic) reasoning

Graph Definitions

Neighbour

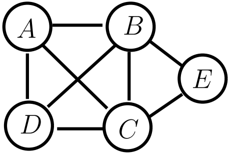
In an undirected graph a neighbour of x are all vertices that share an edge with x.

Clique

Given an undirected graph a clique is a subset of fully connected vertices. All members of the clique are neighbours, there is no larger clique that contains the clique.

Graph Definitions

- Example of cliques



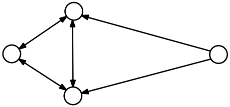
- Two cliques (A, B, C, D) and (B, C, E)
- (A, B, C) are no (maximal) clique (sometimes called a *cliquo*)

- Why cliques?
- In *modelling* they describe variables that all depend on each other.
- In *inference* they describe sets of variables with no simpler structure to describe their relationships

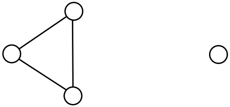
Graph Definitions

Connected Graph

A graph is connected if there is a path between any two vertices. Otherwise there are connected components.



connected graph

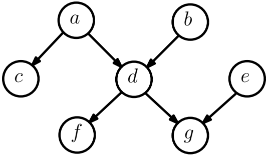


graph with
two connected components

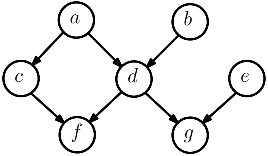
Graph Definitions

Singly- and Multiply Connected

A graph is singly-connected if for any vertex a and b there exists not more than one path between them. Otherwise it is multiply-connected. Another name for a singly-connected graph is a tree. A multiply connected graph is also called loopy.



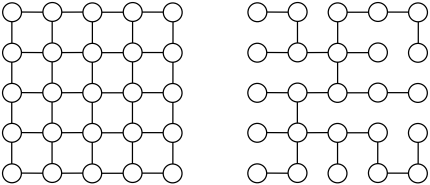
singly-connected



multiply-connected

Spanning Tree

A spanning tree of an undirected graph G is a singly-connected subset of the existing edges such that the resulting singly-connected graph covers all vertices of G. A maximum (weight) spanning tree is a spanning tree such that the sum of all weights on the edges is larger than for any other spanning tree of G.



- There might be more than one maximum spanning tree.

Markov Networks

Markov Networks

- So far, factorization with each factor a probability distribution
  - Normalization as a by-product
- Alternative:

$$p(a, b, c) = \frac{1}{Z}\phi(a, b)\phi(b, c) \tag{13}$$

- Here Z normalization constant or partition function

$$Z = \sum_{a,b,c} \phi(a, b)\phi(b, c) \tag{14}$$

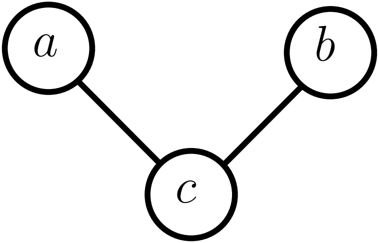
Definitions

Potential

A potential $\phi(x)$ is a non-negative function of the variable x. A joint potential $\phi(x_1, \ldots, x_D)$ is a non-negative function of the set of variables.

- Distribution (as in belief networks) is a special choice

Example



$$p(a,b,c) = \frac{1}{Z}\phi_{ac}(a,c)\phi_{bc}(b,c) \quad (15)$$

Markov Network

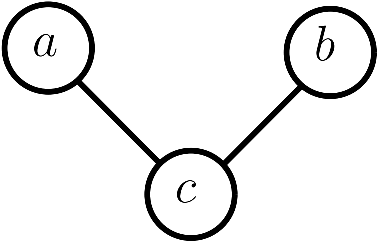
Markov Network

For a set of variables $\mathcal{X} = \{x_1, \dots, x_D\}$ a Markov network is defined as a product of potentials over the maximal cliques $\mathcal{X}_c$ of the graph $\mathcal{G}$

$$p(x_1, \dots, x_D) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c) \tag{16}$$

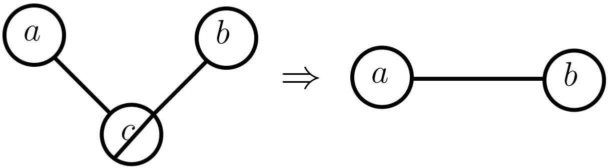
- Special case: cliques of size 2 – pairwise Markov network
- In case all potentials are strictly positive this is called a Gibbs distribution

Properties of Markov Networks



$$p(a, b, c) = \frac{1}{Z}\phi_{ac}(a, c)\phi_{bc}(b, c) \tag{17}$$

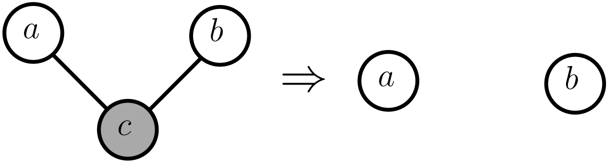
Properties of Markov Networks



- Marginalizing over c makes a and b "graphically" dependent

$$p(a,b) = \sum_c \frac{1}{Z}\phi_{ac}(a,c)\phi_{bc}(b,c) = \frac{1}{Z}\phi_{ab}(a,b) \tag{18}$$

Properties of Markov Networks



- Conditioning on c makes a and b independent

$$p(a, b \mid c) = p(a \mid c)p(b \mid c) \tag{19}$$

- This is opposite to the directed version $a \to c \leftarrow b$ where conditioning *introduced* dependency

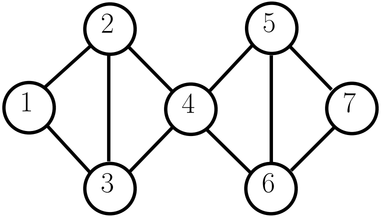
Local Markov Property

Local Markov Property

$$p(x \mid \mathcal{X} \setminus \{x\}) = p(x \mid ne(x)) \tag{20}$$

- Condition on neighbours independent on rest

Local Markov Property – Example



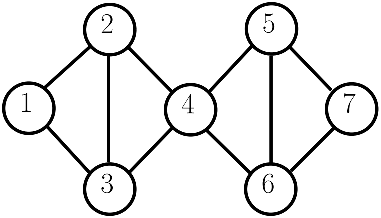
- $x_4 \perp\!\!\!\perp \{x_1, x_7\} \mid \{x_2, x_3, x_5, x_6\}$

Global Markov Property

Global Markov Property

For disjoint sets of variables $(\mathcal{A}, \mathcal{B}, \mathcal{S})$ where $\mathcal{S}$ separates $\mathcal{A}$ from $\mathcal{B}$, then $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}$

Local Markov Property – Example

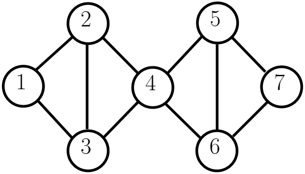


- $x_1 \perp\!\!\!\perp x_7 \mid \{x_4\}$
- and others

Hammersley-Clifford Theorem

- An undirected graph specifies a set of conditional independence statements
- Question: What is the most general factorization (form of the distribution) that satisfies these independences?
- In other words: given the graph, what is the implied factorization?

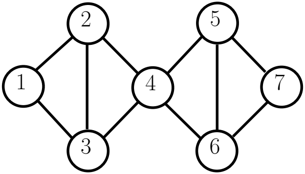
Finding the Factorization



- Eliminate variable one by one
- Let's start with x_1

$$p(x_1, \ldots, x_7) = p(x_1 \mid x_2, x_3) p(x_2, \ldots, x_7) \tag{21}$$

Finding the Factorization



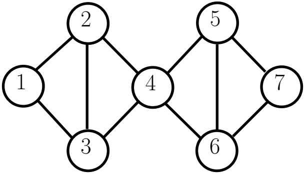
- Graph specifies:

$$\begin{aligned} & p(x_1, x_2, x_3 \mid x_4, \ldots, x_7) && = p(x_1, x_2, x_3 \mid x_4) \\ \Rightarrow \quad & p(x_2, x_3 \mid x_4, \ldots, x_7) && = p(x_2, x_3 \mid x_4) \end{aligned}$$

- Hence

$$p(x_1, \ldots, x_7) = p(x_1 \mid x_2, x_3) p(x_2, x_3 \mid x_4) p(x_4, x_5, x_6, x_7)$$

Finding the Factorization



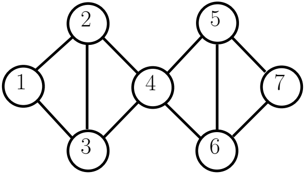
- We continue to find

$$
\begin{aligned}
p(x_1, \ldots, x_7) \quad = \quad & p(x_1 \mid x_2, x_3) p(x_2, x_3 \mid x_4) \\
& p(x_4 \mid x_5, x_6) p(x_5, x_6 \mid x_7) p(x_7)
\end{aligned}
$$

- A factorization into clique potentials (maximal cliques)

$$p(x_1, \ldots, x_7) = \frac{1}{Z} \phi(x_1, x_2, x_3) \phi(x_2, x_3, x_4) \phi(x_4, x_5, x_6) \phi(x_5, x_6, x_7)$$

Finding the Factorization



- Markov conditions of graph $G \Rightarrow$ factorization F into clique potentials
- And conversely: $F \Rightarrow G$

Hammersley-Clifford Theorem

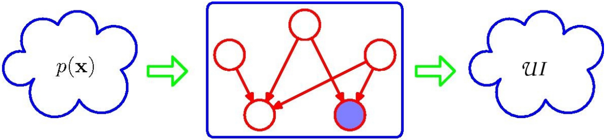
Hammersely-Clifford

This factorization property $G \Leftrightarrow F$ holds for any undirected graph provided that the potentials are positive

- Thus also loopy ones: $x_1 - x_2 - x_3 - x_4 - x_1$
- Theorem says, distribution is of the form

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_{12}(x_1, x_2)\phi_{23}(x_2, x_3)\phi_{34}(x_3, x_4)\phi_{41}(x_4, x_1)$$

Filter View



- Let $\mathcal{UI}$ denote the distributions that can pass
  - those that satisfy all conditional independence statements
- Let $\mathcal{UF}$ denote the distributions with factorization over cliques
- Hammersley-Clifford says : $\mathcal{UI} = \mathcal{UF}$

Next Time ...

- One graph to rule them all: