Probabilistic Graphical Models

Machine Learning

School of Artificial Intelligence, **Xidian University**

Bayesian Networks

Conditional Independence

Markov Random Fields

Inference in Graphical Models

Probabilistic Graphical Models

- A simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
- Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph.
- Complex computations can be expressed in terms of graphical manipulations.
  - required to perform inference and learning in sophisticated models,
  - underlying mathematical expressions are carried along implicitly.

Graph

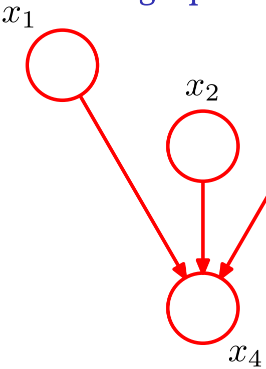
A graph comprises **nodes** (also called *vertices*) connected by **links** (also known as **edges** or *arcs*).

In a probabilistic graphical model

- each node represents a random variable (or group of random variables),
- and the links express probabilistic relationships between these variables.

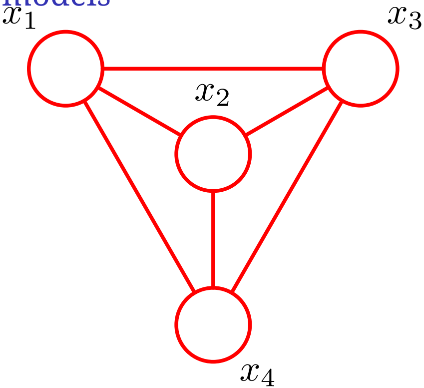
Directed and Undirected Graphical Models

Directed graphical models



Bayesian networks

Undirected graphical models



Markov random fields

Bayesian Networks

Bayesian Networks

- Consider an arbitrary joint distribution $p(a, b, c)$ over three variables a, b, and c.
- Applying the product rule of probability, we get the joint distribution as:

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

- Repeat the application of the product rule, we get:

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

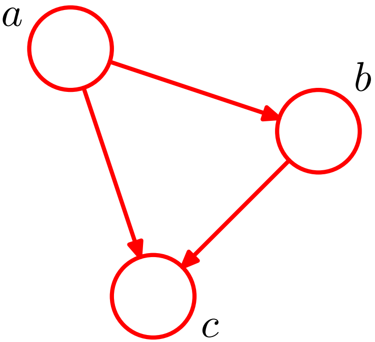
Bayesian Networks

- Consider an arbitrary joint distribution $p(a, b, c)$ over three variables a, b, and c.
- Applying the product rule of probability, we get the joint distribution as:

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

- Repeat the application of the product rule, we get:

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



- A specific graph can make probabilistic statements for a broad class of distributions.

Bayesian Networks

- If there is a link going from a node a to a node b, then we say that
  - node a is the **parent** of node b,
  - and we say that node b is the **child** of node a.
- Note that we have implicitly chosen a particular ordering, namely a, b, c.

Bayesian Networks

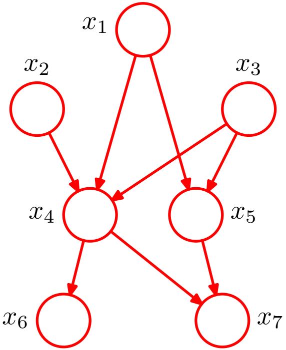
- Let us extend the previous example by considering the joint distribution over K variables given by $p(x_1, \dots, x_K)$.
- By repeated application of the product rule of probability, this joint distribution can be written as a product of conditional distributions,

$$p(x_1, \dots, x_K) = p(x_K | x_1, \dots, x_{K-1}) \cdots p(x_2 | x_1) p(x_1)$$

- This graph is **fully connected** because there is a link between every pair of nodes.

Bayesian Networks

- The **absence** of links in the graph that conveys interesting information about the properties of the class of distributions.
- In the right graph, there is no link from x_1 to x_2 or from x_3 to x_7.
- We can write the joint probability distribution in terms of the product of a set of conditional distributions, one for each node in the graph:



$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5).$$

Bayesian Networks

- The joint distribution defined by a graph is given by
  - the product of a conditional distribution for each node,
  - conditioned on the variables corresponding to its parents.
- For a graph with K nodes, the joint distribution is given by

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

where pa_k denotes the set of parents of x_k, and $\mathbf{x} = \{x_1, \ldots, x_K\}$.

Bayesian Networks

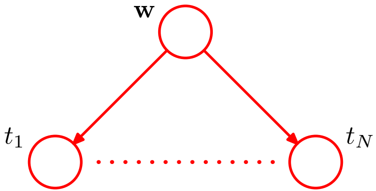
- This key equation expresses the **factorization** properties of the joint distribution for a directed graphical model.
- We can associate sets of variables and vector-valued variables with the nodes of a graph.
- Note that the directed graphs must be with no **directed cycles**. Such graphs are called **directed acyclic graphs**, or **DAGs**.

Bayesian Networks

Example: Polynomial regression

- Directed graphical model representing the following joint distribution corresponding to the Bayesian polynomial regression model

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}).$$



Conditional Independence

Conditional Independence

- An important concept for probability distributions over multiple variables is that of **conditional independence**.
- Suppose that the conditional distribution of a, given b and c, is such that it does not depend on the value of b, so that

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

- We say a is conditional independent of b given c.

Conditional Independence

- An important concept for probability distributions over multiple variables is that of **conditional independence**.
- Suppose that the conditional distribution of a, given b and c, is such that it does not depend on the value of b, so that

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

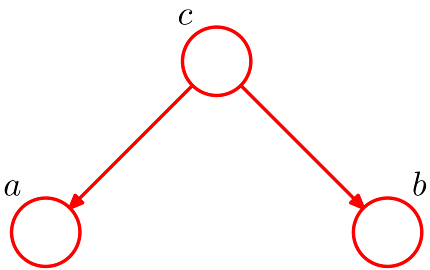
- We say a is conditional independent of b given c.
- This can be expressed in a slightly different way if we consider the joint distribution of a and b conditioned on c:

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}.$$

Conditional Independence

- This says that the variables a and b are statistically independent, given c.
- We use a shorthand notation for conditional independence

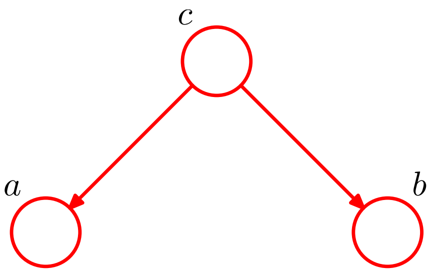
$$a \perp b | c$$



Conditional Independence

- This says that the variables a and b are statistically independent, given c.
- We use a shorthand notation for conditional independence

$$a \perp b | c$$



- Conditional independence properties play an important role by
  - simplifying the structure of a model
  - simplifying the computations needed to perform inference and learning

Conditional Independence

Example graphs

- The joint distribution corresponding to this graph is

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

- If none of the variables are observed, then we can investigate whether a and b are independent by marginalizing both sides

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c).$$

- In general, this does not facorize into the product $p(a)p(b)$, and so

$$a \not\perp\!\!\!\perp b|\emptyset$$

Conditional Independence

Example graphs

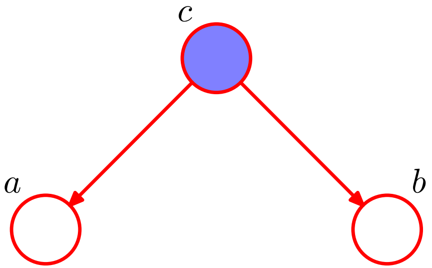
- Now suppose we condition on the variable c, the conditional distribution of a and b, given c, in the form

$$
\begin{aligned}
(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

- so we obtain the conditional independent property

$$a \perp b|c.$$

- The node c is said to be **tail-to-tail**.



Conditional Independence

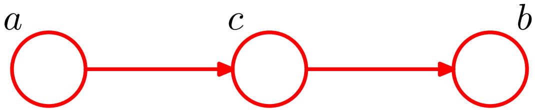
Example graphs

- The joint distribution corresponding to this graph is obtained as:

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

- First of all, suppose that none of the variables are observed.
- Again, we can test if a and b are independent by marginalizing over c:

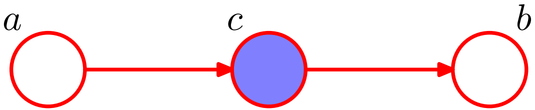
$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$



Conditional Independence

Example graphs

- Now suppose we condition on node c
- Using Bayes' theorem and marginalizing over c:



$$
\begin{aligned}
p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\
&= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

- and so again we obtain the conditional independence property

$$a \perp b|c.$$

Conditional Independence

Example graphs

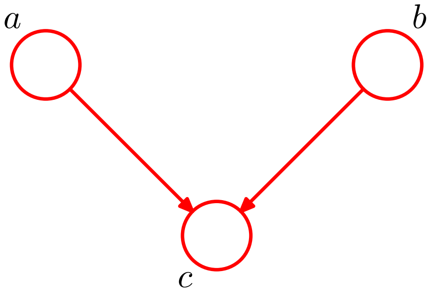
- We consider a 3-node example with has a more subtle behaviour than the two previous graphs.
- The joint distribution can be written down as:

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

- Consider the case where none of the variables are observed.
  - Marginalizing both sides over c we obtain

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

  - so a and b are independent with no variables observed.



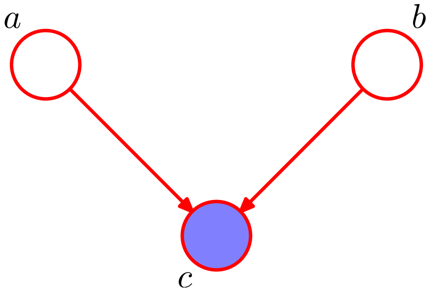
Conditional Independence
Example graphs

- Now suppose we condition on c, as indicated in the right figure.
- The conditional distribution of a and b is then given by

$$
\begin{aligned}
(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\
&= \frac{p(a)p(b)p(c|a, b)}{p(c)}
\end{aligned}
$$

- which in general does not factorize into the product $p(a)p(b)$, and so

$$a \not\perp\!\!\!\perp b | c.$$



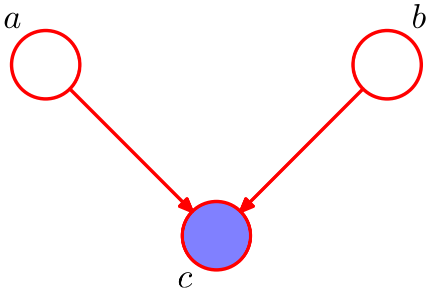
Conditional Independence
Example graphs

- Now suppose we condition on c, as indicated in the right figure.
- The conditional distribution of a and b is then given by

$$(a, b|c) = \frac{p(a,b,c)}{p(c)}$$
$$= \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

- which in general does not factorize into the product $p(a)p(b)$, and so

$$a \not\perp\!\!\!\perp b|c.$$



- Node c is *head-to-head* w.r.t. the paths from a to b.
- When node c is unobserved, it “blocks” the path.

Conditional Independence

- Node y is a *descendant* of node x if there is a path from x to y,
- in which each step of the path follows the directions of the arrows.
- A head-to-head path will become **unblocked** if either the node, or *any of its descendants*, is observed.

Conditional Independence

- Node y is a *descendant* of node x if there is a path from x to y,
- in which each step of the path follows the directions of the arrows.
- A head-to-head path will become **unblocked** if either the node, or *any of its descendants*, is observed.

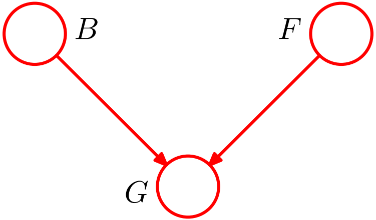
Summary

- A tail-to-tail node or a head-to-tail node leaves a path unblocked unless it is observed in which case it blocks the path.
- A head-to-head node blocks a path if it is unobserved
  - once the node, and/or at least one of its descendants, is observed the path becomes unblocked.

Conditional Independence

- Three binary random variables relating to the fuel system on a car.
  - B the state of a battery that is either charged ($B = 1$) or flat ($B = 0$),
  - F the state of the fuel tank that is either full of fuel ($F = 1$) or empty ($F = 0$),
  - G the state of an electric fuel gauge and which indicates either full ($G = 1$) or empty ($G = 0$)

Conditional Independence



- The battery is either charged or flat, and independently the fuel tank is either full or empty, with prior probabilities

$$p(B = 1) = 0.9 \qquad p(F = 1) = 0.9.$$

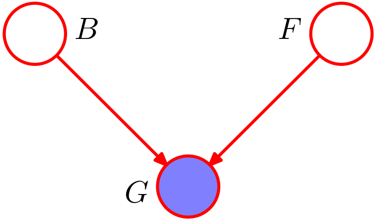
- Given the state of the fuel tank and the battery, the fuel gauge reads full with probabilities given by

$$p(G = 1|B = 1, F = 1) = 0.8 \qquad p(G = 1|B = 1, F = 0) = 0.2$$
$$p(G = 1|B = 0, F = 1) = 0.2 \qquad p(G = 1|B = 0, F = 0) = 0.1$$

- Before we observe any data, the prior probability of the fuel tank being empty is $p(F = 0) = 0.1$.

Conditional Independence

- Suppose that we observe the fuel gauge and discover that it reads empty, i.e., $G = 0$.
- We can use Bayes' theorem to evaluate the posterior probability of the fuel tank being empty.



$$p(F = 0|G = 0) = \frac{p(G = 0|F = 0)p(F = 0)}{p(G = 0)}$$

Conditional Independence

- First we evaluate the denominator for Bayes' theorem given by

$$p(G=0) = \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G=0|B,F)p(B)p(F) = 0.315$$

  and similarly we evaluate

$$p(G=0|F=0) = \sum_{B\in\{0,1\}} p(G=0|B,F=0)p(B) = 0.81$$

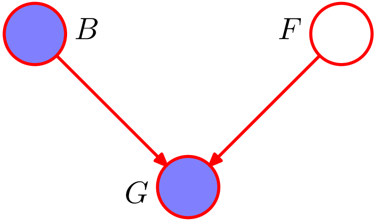
  and using these results we have

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \approx 0.257$$

- so $p(F=0|G=0) > p(F=0)$.
- Thus observing that the gauge reads empty makes it more likely that the tank is indeed empty.

Conditional Independence

- Suppose that we also check the state of the battery and find that it is flat, i.e., $B = 0$.
- We have now observed the states of both the fuel gauge and the battery.



Conditional Independence

- The posterior probability that the fuel tank is empty given the observations of both the fuel gauge and the battery state is then given by

$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \approx 0.111$$

  where the prior probability $p(B=0)$ has cancelled between numerator and denominator.
- The probability that the tank is empty has decreased (from 0.257 to 0.111) as a result of the observation of the state of the battery.
- This accords with our intuition that finding out that the battery is flat **explains away** the observation that the fuel gauge reads empty.

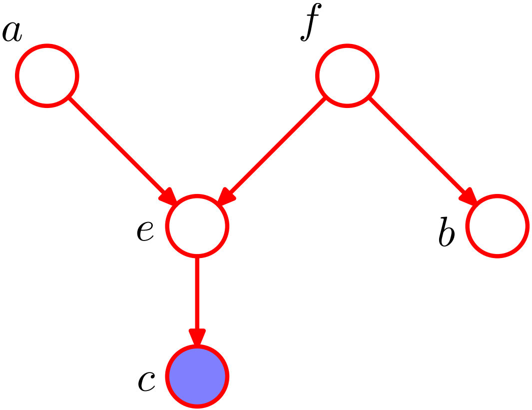
Conditional Independence

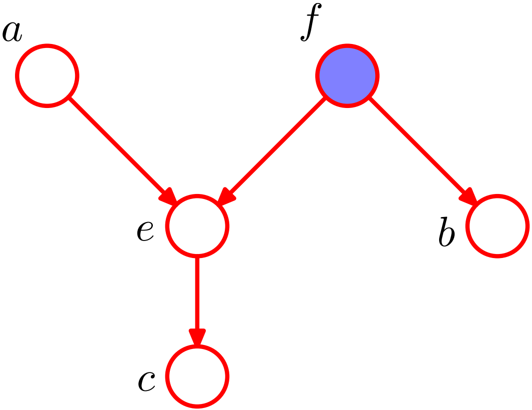
D-separation: $A \perp B | C$

- Consider a general directed graph in which A, B, and C are arbitrary nonintersecting sets of nodes.
- Consider all possible paths from any node in A to any node in B
- Any such path is said to be **blocked** if it includes a node such that either
  - the arrows on the path meet either *head-to-tail* or *tail-to-tail* at the node, and the node is in the set C, or
  - the arrows meet *head-to-head* at the node, and neither the node, nor any of its descendants, is in the set C.
- If all paths are blocked, then A is said to be **d-separated** from B by C.

Conditional Independence

D-separation Example





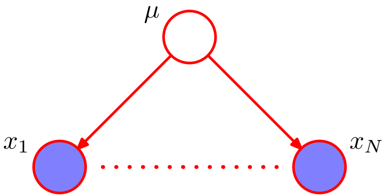
- Path a to b w.r.t. f.
- Path a to b w.r.t. e.
  - Node e has a descendant c.

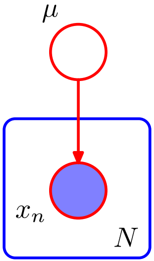
- Path a to b w.r.t. f
  - $a \perp b | f$

Conditional Independence

D-separation: i.i.d.

- Finding the posterior distribution for the mean of a univariate Gaussian distribution.
- The joint distribution is defined by
  - a prior $p(\mu)$ together with
  - a set of conditional distributions $p(x_n|\mu)$

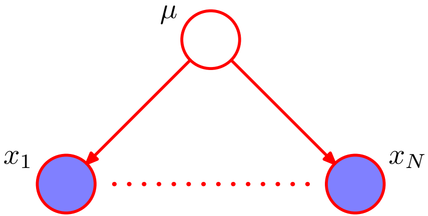




Conditional Independence

D-separation: i.i.d.

- In practice, we observe $\mathcal{D} = \{x_1, \dots, x_N\}$
- Our goal is to estimate μ.
- Path from x_i to $x_{j \neq i}$ is tail-to-tail w.r.t. node μ.



$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu).$$

- However, observations are no longer independent after integration

$$p(\mathcal{D} = \int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n).$$

Markov Random Fields

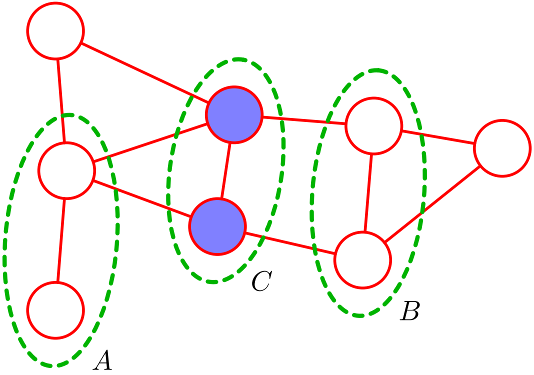
Markov Random Fields

- A **Markov random field**, also known as a **Markov network** or an **undirected graphical model**, has
  - *a set of nodes* each of which corresponds to a variable or group of variables,
  - as well as *a set of links* each of which connects a pair of nodes.

Markov Random Fields

Conditional independence property

- All possible paths that connect nodes in set A to nodes in set B.
- If all such paths through one or more nodes in set C, then all such paths are "blocked", so the conditional independency property holds.
- However, if there is *at least one* such path that is *not* blocked, then the property does not necessarily hold.



Conditional independence property

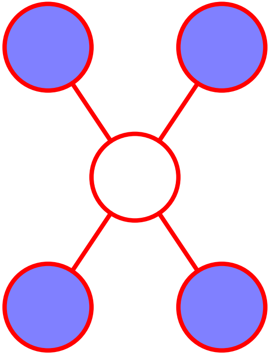
$$A \perp B | C$$

Markov Random Fields

- A **factorization rule** for undirected graphs that will correspond to the above conditional independence test.
- Two nodes x_i and x_j that are not connected by a link, then these variables must be conditionally independent given all other nodes in the graph.

$$p(x_i, x_j | \mathbf{x}_{-\{i,j\}}) = p(x_i | \mathbf{x}_{-\{i,j\}}) p(x_j | \mathbf{x}_{-\{i,j\}})$$

where $\mathbf{x}_{-\{i,j\}}$ denotes the set $\mathbf{x}$ of all variables with x_i and x_j removed.

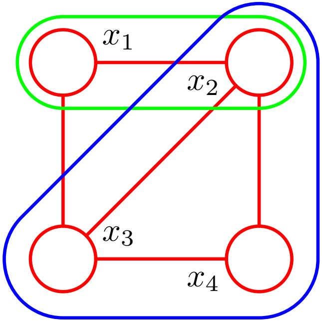


Markov Random Fields

A clique is defined as a subset of the nodes in a graph such that there exists a link between all pairs of nodes in the subset.

A maximal clique is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.

- Five cliques of two nodes
- Two maximal cliques given by $\{x_1, x_2, x_3\}$, and $\{x_2, x_3, x_4\}$.



Markov Random Fields

- Let us denote a clique by C and the set of variables in that clique by $\mathbf{x}_C$.
- The joint distribution is written as a product of **potential functions** $\psi_C(\mathbf{x}_C)$ over the maximal cliques of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C)$$

- The quantity Z, sometimes called the **partition function**, is the normalization constant and is given by

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C)$$

- Note: a model with M discrete nodes each having K states.

Markov Random Fields

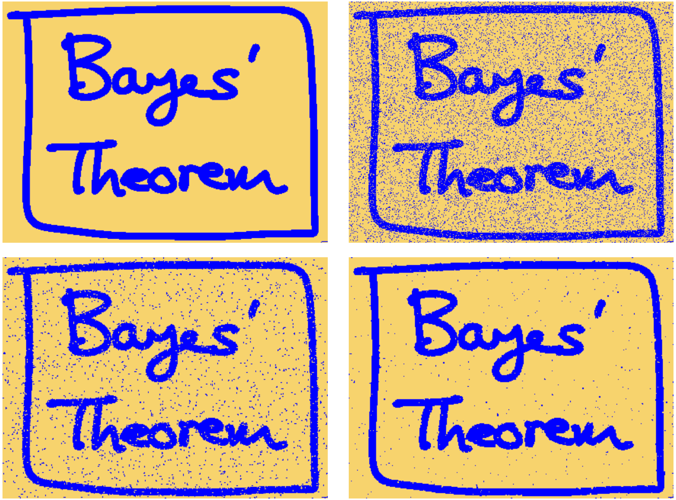
- Because potential functions are strictly positive,
- It is convenient to express them as exponentials,

$$\psi_C(\mathbf{x}_C) = \exp\Big[-E(\mathbf{x}_C)\Big]$$

where $E(\mathbf{x}_C)$ is called an **energy function**, and the exponential representation is called the **Boltzmann distribution**.

Markov Random Fields

Image denoising: iterated conditional models vs graph-cut



Markov Random Fields

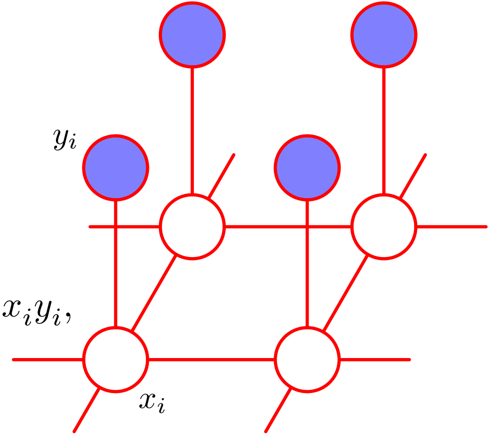
- Two types of cliques
  - $\{x_i, y_i\}$ with energy function $-\eta x_i y_i$ expressing the correlation between these variables.
  - $\{x_i, x_j\}$ of neighboring pixels with an energy $-\beta x_i x_j$.
- The complete energy function for the model then takes the form

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i,$$

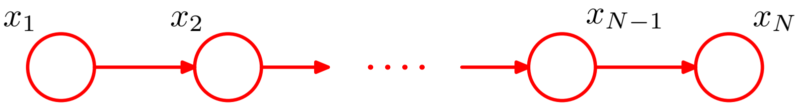
- which defines a joint distribution

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}.$$

Prior knowledge of image



Markov Random Fields

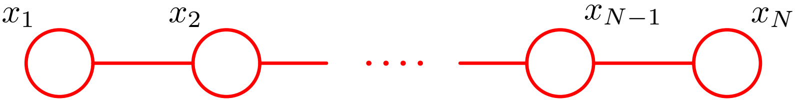


- The joint distribution of the directed graph is

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1}).$$

- Convert it to an undirected graph representation.

Markov Random Fields



- As the maximal cliques are pairs of neighboring nodes, the joint distribution is

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N).$$

- So we have

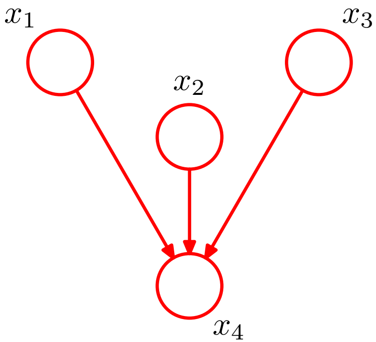
$$\begin{aligned}
\psi_{1,2}(x_1, x_2) &= p(x_1)p(x_2|x_1) \\
\psi_{2,3}(x_2, x_3) &= p(x_3|x_2) \\
&\vdots \\
\psi_{N-1,N}(x_{N-1}, x_N) &= p(x_N|x_{N-1}).
\end{aligned}$$

Markov Random Fields

- The joint distribution

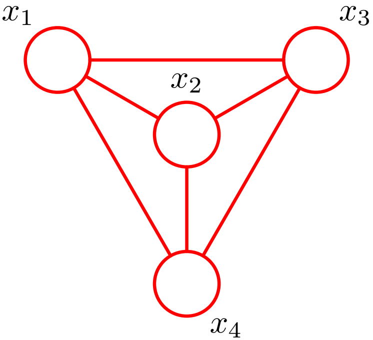
  $$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3).$$

- The factor $p(x_4|x_1, x_2, x_3)$ involves all four variables.
- They must all belong to a single clique.



Markov Random Fields

- We add extra links between all pairs of parents of the node x_4.
- This process of “marrying the parents” is known as **moralization**, the resulting graph is called the **moral graph**.
- Converting from an undirected to a directed representation is much less common.

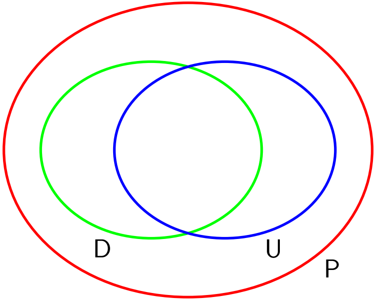


Markov Random Fields

A D map (for “dependency map”) if every conditional independence statement satisfied by the distribution is reflected in the graph.

An I map (for “independence map”) if every conditional independence statement implied by a graph is satisfied by a specific distribution.

A perfect map every conditional independence property reflected, and vice versa.

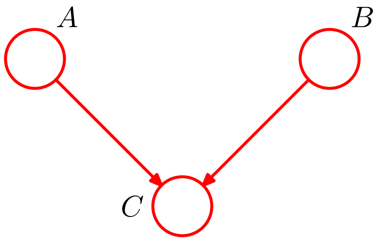


Markov Random Fields

- A directed graph that is a perfect map for a distribution satisfying the conditional independence properties

$$A \perp B | \emptyset$$
$$A \not\perp B | C$$

- There is no corresponding undirected graph over the same three variables that is a perfect map.

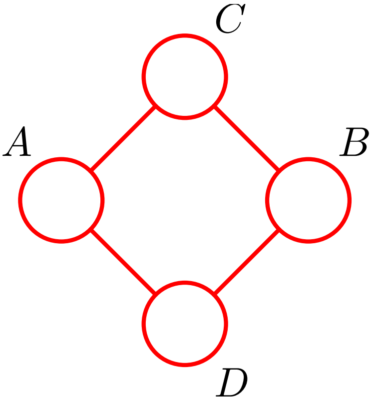


Markov Random Fields

- The undirected graph over four variables has properties

$$A \perp B | \emptyset$$
$$C \perp D | A \cup B$$
$$A \perp B | C \cup D$$

- No directed graph implies the same set of conditional independencies.



Inference in Graphical Models

Inference in Graphical Models

a) Decompose into a product of factors

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

b) Obsered the value of y, using the sum and product rules

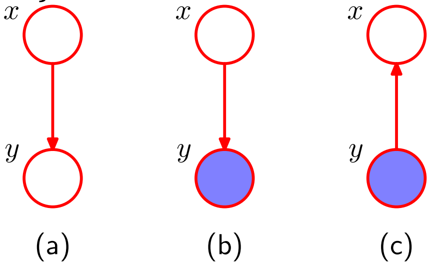
$$p(y) = \sum_{x'} p(y|x')p(x')$$

apply Bayes' theorem

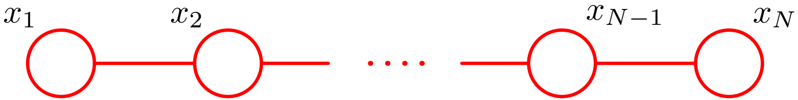
$$p(x|y) = p(y|x)p(x)/p(y).$$

c) The joint distribution $p(x, y) = p(y)p(x|y)$.

A graphical representation of Bayes' theorem



Inference in Graphical Models



- The joint distribution is

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N).$$

- Consider the case of N nodes representing discrete variables each having K states
  - Each potential function $\psi_{n-1,n}(x_{n-1}, x_n)$ comprises an $K \times K$ table, indicating $(N-1)K^2$ parameters.

Inference in Graphical Models

- To find the marginal distribution $p(x_n)$
- By definition, it can be obtained by summing the joint distribution over all variables except x_n,

$$p(x_n) = \sum_{x_1} \cdots \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x}).$$

Inference in Graphical Models

- Rearrange the order of the summations and the multiplications to allow the marinalization be computed more efficiently.
- As potential $\psi_{N-1,N}(x_{N-1}, x_N)$ is the only one depends on x_N, so we can perform the summation

$$\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)$$

  first to give a function of x_{N-1}.
- Then sum over x_{N-1} and similarly we can make summation over x_1 which involves only the potential $\psi_{1,2}(x_1, x_2)$.

Inference in Graphical Models

- If we group the potentials and summations together in this way, we can express the desired marginal in the form

$$p(x_n) = \frac{1}{Z}$$

$$\underbrace{\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3) \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \cdots \right]}_{\mu_\alpha(x_n)}$$

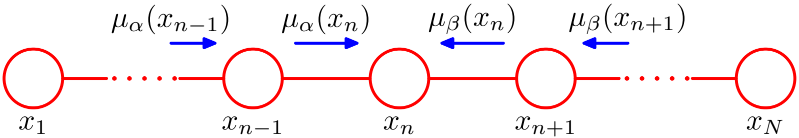
$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots \right]}_{\mu_\beta(x_n)}.$$

Inference in Graphical Models

- Passing of local messages around on the graph

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

  - $\mu_\alpha(x_n)$ a message passed forwards
  - $\mu_\beta(x_n)$ a message passed backwards



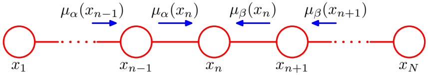
Inference in Graphical Models

- The message $\mu_\alpha(x_n)$ can be evaluated recursively because

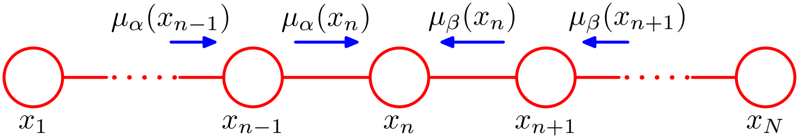
$$\mu_\alpha(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \left[\sum_{x_{n-2}} \cdots \right]$$
$$= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \mu_\alpha(x_{n-1}).$$

- Similarly, the message $\mu_\beta(x_n)$ can be evaluated recursively by starting with node x_N and using

$$\mu_\beta(x_n) = \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n) \mu_\beta(x_{n+1}).$$



Inference in Graphical Models



- Graphs of the form called **Markov chains**
- the corresponding message passing equations represent an example of the **Chapman- Kolmogorov** equations for Markov processes.