Chris Bishop's PRML
Ch. 8: Graphical Models

Ramya Narasimha & Radu Horaud

January 24, 2008

Introduction

- Visualize the structure of a probabilistic model
- Design and motivate new models
- Insights into the model's properties, in particular *conditional independence* obtained by inspection
- Complex computations = graphical manipulations

A few definitions

- Nodes (vertices) + links (arcs, edges)
- Node: a random variable
- Link: a probabilistic relationship
- Directed graphical models or Bayesian networks useful to express *causal* relationships between variables.
- Undirected graphical models or Markov random fields useful to express soft constraints between variables.
- Factor graphs convenient for solving inference problems

Chapter organization

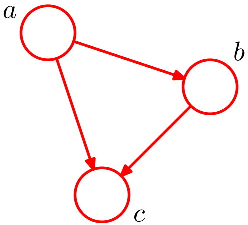
8.1 **Bayesian Networks**: Representation, polynomial regression, generative models, discrete variables, linear-Gaussian models.

8.2 **Conditional independence**: Generalities, D-separation

8.3 **Markov random fields**: conditional independence, factorization, image processing example, relation to directed graphs

8.4 **Inference in graphical models**: next reading group.

Bayesian networks (1)



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

Notice that the left-hand side is symmetrical w/r to the variables whereas the right-hand side is not.

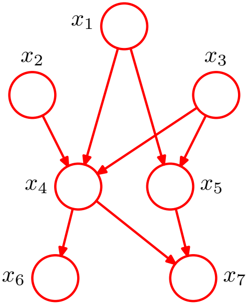
Bayesian networks (2)

Generalization to K variables:

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

- The associated graph is *fully connected*.
- The absence of links conveys important information.

Bayesian networks (3)



It is obvious to obtain the associated joint probability $p(x_1, \ldots, x_7)$.

Bayesian networks (4)

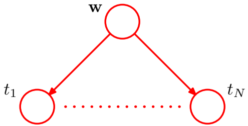
More generally, for a graph with K nodes the joint distribution is:

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

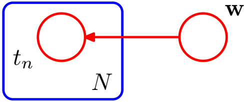
- this key equation expresses the factorization properties of the joint distribution.
- there must be no directed cycles
- these graphs are also called DAGs or *directed acyclic graphs*.
- equivalent definition: there exists an ordering on the nodes such that there are no links going from any node to any lowered numbered node (see example of Figure 8.2).

Polynomial regression (1)

- random variables: polynomial coefficients **w** and the observed data **t**.
- $p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$

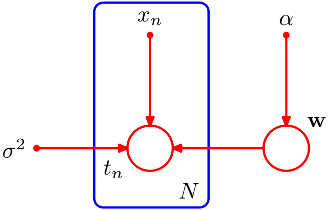


OR

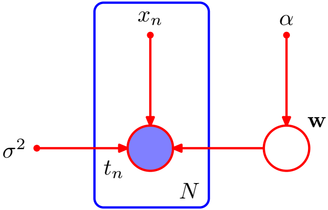


The box is called a plate

Polynomial regression (2)



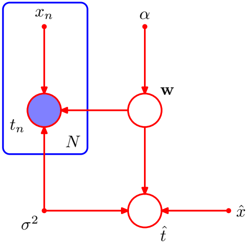
Deterministic parameters shown by small nodes



shaded nodes are set to observed values

Polynomial regression (3)

- the observed variables, $\{t_n\}$, are shown by shaded nodes
- the values of the variables $\mathbf{w}$ are not observed – latent or hidden variables.
- but these variables are not of direct interest
- the goal is to make predictions for new input values, ie the graphical model below:



Generative models

- Back to:

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

- each node has a higher number than any of its parents
- the factorization above corresponds to a DAG.
- goal: draw a sample $\hat{x}_1, \ldots, \hat{x}_K$ from the joint distribution.
- apply ancestral sampling start from lower-numbered nodes, downwards trhough the graph's nodes.
- generative graphical model captures the *causal* process that generated the observed data (object recognition example)

Discrete variables (1)

- The case of a single discrete variable $\mathbf{x}$ with K possible states (**look at section 2.2 on multinomial variables**):

$$p(\mathbf{x}|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

with $\mu = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ and $\sum_k \mu_k = 1$ hence $K - 1$ variables need be specified.

- The case of two variables, with similar notations and definitions:

$$p(\mathbf{x}_1, \mathbf{x}_2|\mu) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}$$

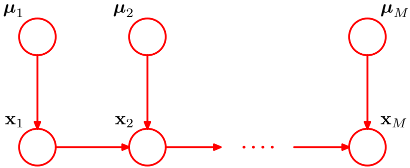
with the constraint $\sum_k \sum_l \mu_{kl} = 1$ there are $K^2 - 1$ parameters.

Discrete variables (2)

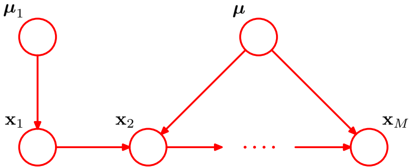
- If the two variables are independent, the number of parameters drops to $2(K-1)$.
- The general case of M discrete variables generalizes to $K^M - 1$ parameters, which reduces to $M(K-1)$ parameters for M independent variables.
- In this example there are $K - 1 + (M-1)K(K-1)$ parameters: 

- the sharing or tying of parameters is another way to reduce their number.

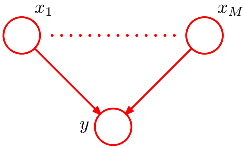
Discrete variables with Dirichlet priors (3)



The same with tied parameters:



Discrete variables (4)

- Introduce parameterizations of the conditional distributions to control the exponential growth: an example with binary variables.
- This graphical model:  requires 2^M parameters representing the probability $p(y = 1)$.

- Alternatively, use a logistic sigmoid function over a linear combination of the parents:

$$p(y = 1|x_1, \ldots, x_M) = \sigma\left(w_0 + \sum_i w_i x_i\right)$$

Linear-Gaussian models (1)

- Extensive use of this section in later chapters...
- Back to DAG: $p(\mathbf{x}) = \prod_{k=1}^{D} p(x_k|pa_k)$
- The distribution of node i:

$$p(x_i|pa_i) = \mathcal{N}\left(x_i|\sum_{j\in pa_i} w_{ij}x_j + b_i, v_i\right)$$

- the logarithm of the joint distribution is a quadratic function in $x_1, \ldots, x_D$ (see equations (8.12) and (8.13)).
- The joint distribution $p(\mathbf{x})$ is a multivariate function.
- The the mean and variance of this joint distribution can be determined recursively, given the parent-child relationships in the graph (see details in the book).

Linear-Gaussian models (2)

- The case of independent variables (no links in the graph): the covariance matrix is diagonal.
- A fully connected graph: the covariance matrix is a general one with $D(D-1)/2$ entries.
- Intermediate level of complexity correspond to partially constrained covariance matrices.
- It is possible to extend the model to the case in which the nodes represent multivariate Gaussian variables.
- Later chapters will treat the case of hierarchical Bayesian models

Conditional Independence

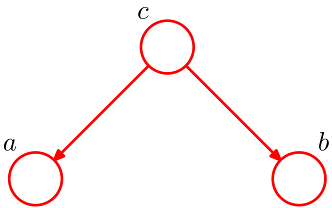
Consider three variable a, b and c

$$p(a|\,b,c) = p(a|\,c) \tag{1}$$

Then a is conditionally independent of b given c

$$p(a,b|\,c) = p(a|\,c)p(b|\,c) \tag{2}$$

a and b are **Statistically independent** given c
Shorthand notation : $a \perp b|c$



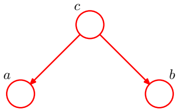
Conditional Independence

- Simplifies the structure of a probabilistic model
- Simplifies the computations needed for inference and learning
- This property can be tested by repeated application of sum and product rules of probability: Time consuming!!

Advantage of Graphical models

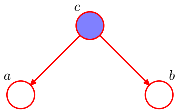
- Conditional independence can be read directly from the graph without having to perform any analytical manipulations
- The framework for achieving this : **D-separation**

Example-I



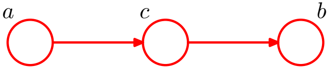
$$
\begin{aligned}
p(a,b,c) &= p(a|\,c)p(b|\,c)p(c) \qquad (3) \\
p(a,b) &= \sum_c p(a|\,c)p(b|\,c)p(c) \neq p(a)p(b) \longrightarrow a \not\perp b|\emptyset
\end{aligned}
$$

Using Bayes' Theorem



$$
\begin{aligned}
p(a,b|\,c) &= \frac{p(a,b,c)}{p(c)} \qquad (4) \\
&= p(a|\,c)p(b|\,c) \longrightarrow a \perp b|c
\end{aligned}
$$

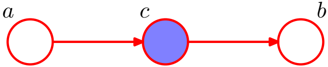
Example-II



$$p(a,b,c) = p(a)p(c|a)p(b|c) \quad (5)$$

$$p(a,b) = p(a)\sum_c p(c|a)p(b|c) = p(a)p(b|a) \longrightarrow a \not\perp b|\emptyset$$

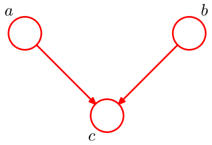
Using Bayes' Theorem



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} \quad (6)$$

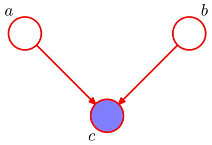
$$= p(a|c)p(b|c) \longrightarrow a \perp b|c$$

Example-III



$$p(a,b,c) = p(a)p(b)p(c|\,a,b) \quad (7)$$
$$p(a,b) = p(a)p(b) \longrightarrow a \perp b|\emptyset$$

Using Bayes' Theorem

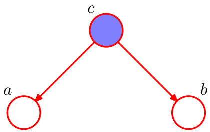


$$p(a,b|\,c) = \frac{p(a)p(b)p(c|\,a,b)}{p(c)} \longrightarrow a \not\perp b|c$$

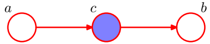
Terminology: x is the *Descendant* of y if there is path from x to y in which each step of the path follows directions of arrows

observed c blocks path a — b

- Tail to Tail nodes

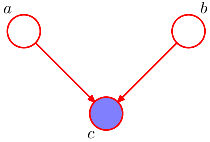


- Head to Tail nodes

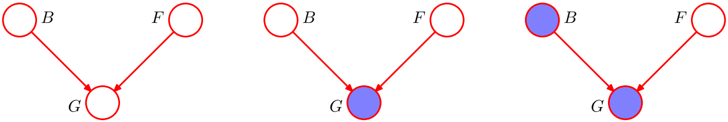


observed c *unblocks* path a — b

- Head to Head nodes



Fuel gauge Example



B : Battery state either 0 or 1

F : Fuel state either 0 or 1

G : Gauge reading either 0 or 1

Observing the reading of the gauge G makes the fuel state F and battery state B dependent

D-separation

D stands for Directed

A, B and C: non-intersecting sets of nodes

To ascertain $A \perp B|C$:

- Consider all paths that are *Blocked* from any node A to any node B
- Path is said to be Blocked path if it includes a node such that
  - 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 d-separated from B by C

Example-I

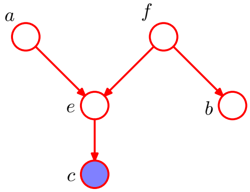


Figure: $a \not\perp b | c$

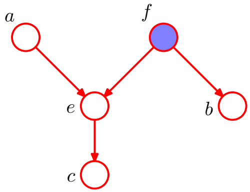
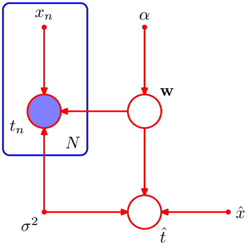


Figure: $a \perp b | f$

Example-II



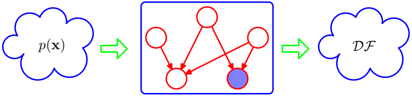
- **w** is a tail-to-tail node with respect to the path from $\hat{t}$ to any one of the nodes $\{t_n\}$
- Hence $\hat{t} \perp t_n | \mathbf{w}$
- Interpretation:
  - First use the training data to determine the posterior distribution over **w**
  - Discard $\{t_n\}$ and use posterior distribution for **w** to make predictions of $\hat{t}$ for new input observations $\hat{x}$

Interpretation as Filter

- Filter-I: allows a distribution to pass through if, and only if, it can be expressed in terms of the factorization implied by the graph

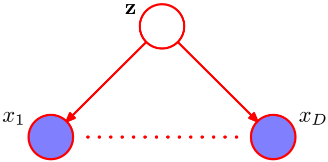
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | pa_k) \quad (8)$$

- Filter-II: allows distributions to pass according to whether they respect all of the conditional independencies implied by the d-separation properties of the graph
- The set of all possible probability distributions $p(\mathbf{x})$ that is passed by *both* the filters is precisely the same
- And are denoted by $\mathcal{DF}$, for *directed factorization*



Naive Bayes Model

- Conditional independence is used to simplify the model structure
- Observed: **x** a D-dimensional vector
- K-Classes: represented as K-dimensional binary vector **z**
- $p(\mathbf{z}|\,\mu)$: Multinomial prior i.e., prior probability of class k
- Graphical representation of naive Bayes model, assumes all components **x** are conditionally independent given **z**
- However this assumption fails when marginalized over **z**



Directed Graphs: Summary

- Represents specific decomposition of a joint probability distribution into a product of conditional probabilities
- Expresses a set of conditional independence statements through d-separation criterion
- Distributions satisfying d-separation criterion are denoted as $\mathcal{DF}$
- Extreme Cases: $\mathcal{DF}$ can contain all possible distributions in case of fully connected graph or product of marginals in case fully disconnected graphs

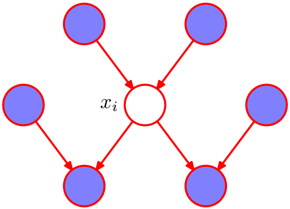
Markov Blanket

Consider a joint distribution $p(\mathbf{x}_1 \ldots \mathbf{x}_D)$

$$p(\mathbf{x}_i | \mathbf{x}_{j \neq i}) = \frac{\prod_k p(\mathbf{x}_k | pa_k)}{\int \prod_k p(\mathbf{x}_k | pa_k) \mathrm{d}\mathbf{x}_i} \quad (9)$$

- Factors not having any functional dependence on $\mathbf{x}_i$ cancel out
- Only factors remaining are
  - Parents and children $\mathbf{x}_i$
  - Also co-parents: corresponding to parents of node $\mathbf{x}_k$ (not $\mathbf{x}_i$)

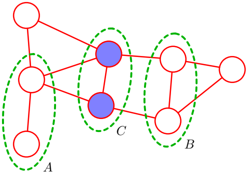
These remaining factors are referred to as The Markov Blanket of node $\mathbf{x}_i$

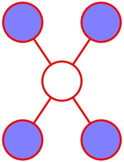


Markov Random Fields

- Also called Undirected Graphical Models
- Consists nodes which correspond to variables or group of variables
- Links within the graph do not carry arrows
- Conditional independence is determined by simple graph separation

Conditional independence properties





Consider three sets of nodes A, B, and C

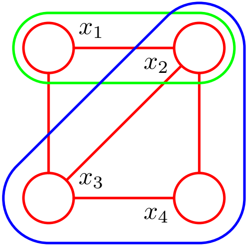
- Consider all possible paths that connect nodes in set A to nodes in set B
- If all such paths pass through one or more nodes in set C, then all such paths are blocked $\rightarrow A \perp B|C$
- Testing for conditional independence in undirected graphs is therefore simpler than in directed graphs
- The Markov blanket: consists of the set of neighboring nodes

Factorization properties

- Consider two nodes x_i and x_j that are not connected by a link then these are conditionally independent given all other nodes
- As there is no direct path between the nodes
- All other paths are blocked by nodes that are observed

$$p(x_i, x_j | \mathbf{x}_{\setminus\{i,j\}}) = p(x_i | \mathbf{x}_{\setminus\{i,j\}}) p(x_j | \mathbf{x}_{\setminus\{i,j\}}) \tag{10}$$

Maximal cliques



- *Clique*: A set of fully connected nodes
- *Maximal Clique*: clique in which it is not possible to include any other nodes without it ceasing to be a clique
- Joint distribution can thus be factored it terms of maximal cliques
- Functions defined on maximal cliques includes the subsets of maximal cliques

Joint distribution

For clique $\mathcal{C}$ and set of variables in that clique $\mathbf{x}_\mathcal{C}$
The joint distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{C}} \Psi_\mathcal{C}(\mathbf{x}_\mathcal{C}) \tag{11}$$

Where Z is the partition function

$$Z = \sum_{\mathbf{x}} \prod_{\mathcal{C}} \Psi_\mathcal{C}(\mathbf{x}_\mathcal{C}) \tag{12}$$

- With M node and K states, the normalization term involves summing over K^M states
- So (in the worst case) is exponential in the size of the model
- The partition function is needed for parameter learning
- For evaluating local marginal probabilities the unnormalized joint distribution can be used

Hammersley and Clifford Theorem

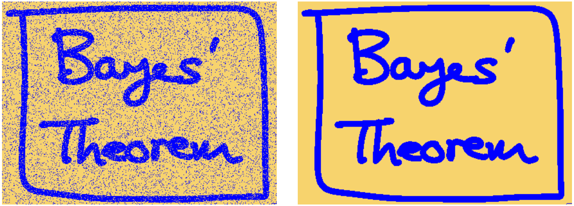
Using filter analogy

- $\mathcal{U}\mathcal{I}$: the set of distributions that are consistent with the set of conditional independence statements read from the graph using graph separation
- $\mathcal{U}\mathcal{F}$: the set of distributions that can be expressed as a factorization described with respect to the maximal cliques
- The Hammersley-Clifford theorem states that the sets $\mathcal{U}\mathcal{I}$ and $\mathcal{U}\mathcal{F}$ are identical if $\Psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$ is strictly positive
- In such case

$$\Psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}) = \exp\{-E(\mathbf{x}_{\mathcal{C}})\} \tag{13}$$

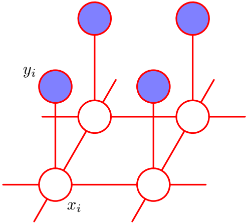
- Where $E(\mathbf{x}_{\mathcal{C}})$ is called an energy function, and the exponential representation is called the Boltzmann distribution

Image Denoising Example



- Noisy Image: $y_i \in \{-1, +1\}$ where i runs over all the pixels
- Unknown Noise Free Image: $x_i \in \{-1, +1\}$
- Goal: Given Noisy image recover Noise Free Image

The Ising Model



Two types of cliques

- $-\eta x_i \, y_i$: giving a lower energy when x_i and y_i have the same sign and a higher energy when they have the opposite sign
- $-\beta x_i \, x_j$: the energy is lower when the neighboring pixels have the same sign than when they have the opposite sign

The Complete energy function and joint distribution

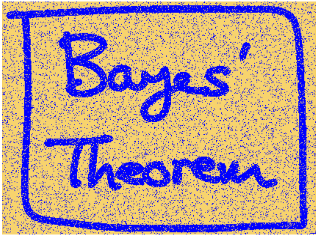
$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i \, x_j - \eta \sum_i x_i \, y_i \tag{14}$$

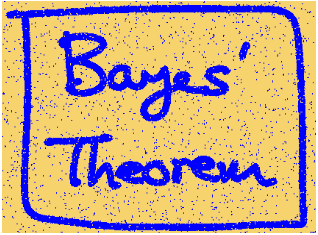
The joint distribution

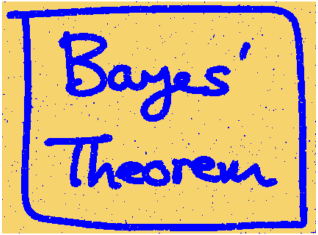
$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\} \quad (15)$$

Fixing **y** as observed values implicitly defines $p(\mathbf{x}|\,\mathbf{y})$
To obtain the image **x** with ICM or any other techniques

- Initialize the variables $x_i = y_i$ for all i
- For x_j evaluate the total energy for the two possible states $x_j = +1$ and $x_j = -1$ with other node variables fixed
- set x_j to whichever state has the lower energy
- Repeat the update for another site, and so on, until some suitable stopping criterion is satisfied


Bayes'
Theorem


Bayes'
Theorem


Bayes'
Theorem

Relation to directed graphs

Distribution for directed graph

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1}) \qquad (16)$$

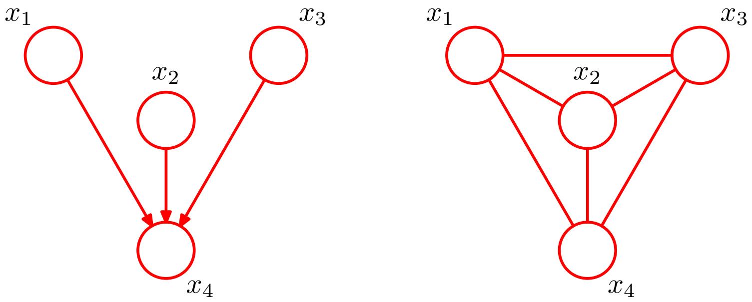
For undirected

$$p(\mathbf{x}) = \frac{1}{Z}\Psi_{1,2}(x_1, x_2)\Psi_{2,3}(x_2, x_2)\cdots\Psi_{N-1,N}(x_{N-1}, x_N) \qquad (17)$$

where

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

Another Example



- In order to convert directed graph into undirected graph add extra links between all pairs of parents
- Anachronistically, this process of 'marrying the parents' has become known as *moralization*
- The resulting undirected graph, after dropping the arrows, is called the *moral graph*

Moralization Procedure

- Add additional undirected links between all pairs of parents for each node in the graph
- Drop the arrows on the original links to give the moral graph
- Initialize all of the clique potentials of the moral graph to 1
- Take each conditional distribution factor in the original directed graph and multiply it into one of the clique potentials
- There will always exist at least one maximal clique that contains all of the variables in the factor as a result of the moralization step
- Going from a directed to an undirected representation discards some conditional independence properties from the graph

D-map and I-maps

Directed and Undirected graphs express different conditional independence properties

- D-map of a distribution: every conditional independence statement satisfied by the distribution is reflected in the graph
- A graph with no links will be trivial D-map
- I-map of a distribution: every conditional independence statement implied by a graph is satisfied by a specific distribution
- Fully connected graph will give I-map for any distribution
- Perfect map: is both D-map and I-map

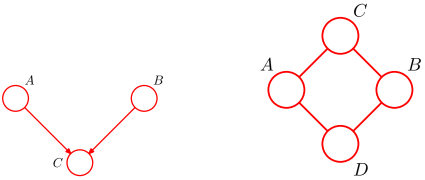


Figure: (a) Directed (b)Undirected

- Case(a)
  - A directed graph that is a perfect map
  - Satisfies the properties $A \perp B|\emptyset$ and $A \not\perp B|C$
  - Has no corresponding undirected graph that is a perfect map
- Case(b)
  - A undirected graph that is a perfect map
  - Satisfies the properties $A \not\perp B|\emptyset$, $C \perp D|A \cup B$ and $A \perp B|C \cup D$
  - Has no corresponding directed graph that is a perfect map