Probabilistic Graphical Models

Part One: Graphs and Markov Properties

Christopher M. Bishop

Microsoft Research, Cambridge, U.K.

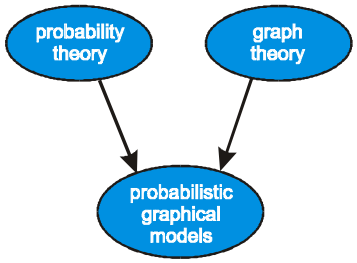
research.microsoft.com/~cmbishop

NATO ASI: Learning Theory and Practice, Leuven, July 2002

Overview of Part One

- Graphs and probabilities
- Directed graphs
- Markov properties
- Undirected graphs
- Examples

Probabilistic Graphical Models



Rules of Probability

- Sum rule

$$P(A) = \sum_B P(A, B)$$

- Product rule

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

Frequentist View of Probabilities

- Limit of infinite number of trials
- Example:

  "*the probability of this coin landing heads is 0.52*"

- Defined as fraction of heads in the limit of an infinite number of trials

Bayesian View of Probabilities

- Quantification of degree of belief, e.g.

  "*the probability that it will rain tomorrow is 0.3*"

- Not possible to repeat "tomorrow"
- *Subjective* and dependent on prior knowledge
- Frequentist probabilities are a special case
- In practice: (i) motivates averaging, (ii) may be computationally intensive (but see later)

Bayes Theorem

- From the product rule

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

- From the sum rule, the denominator can be written

$$P(A) = \sum_B P(A|B)P(B)$$

Role of the Graphs

- New insights into existing models
- Motivation for new models
- Graph based algorithms for calculation and computation (c.f. Feynman diagrams)

Overview of Part One

- Graphs and probabilities
- Directed graphs
- Markov properties
- Undirected graphs
- Examples

Decomposition

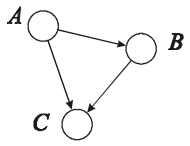
- Consider an arbitrary joint distribution

$$P(A, B, C)$$

- By successive application of the product rule

$$\begin{aligned} P(A,B,C) &= P(A)P(B,C|A) \\ &= P(A)P(B|A)P(C|A,B) \end{aligned}$$

- Diagrammatically:



General Case

- Consider arbitrary joint distribution

$$P(X_1, \ldots, X_d)$$

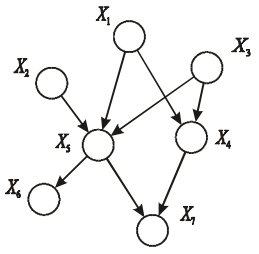
- By successive application of the product rule

$$\begin{aligned} P(X_1, \ldots, X_d) &= P(X_1)P(X_2|X_1) \\ &\quad \ldots P(X_d|X_1, \ldots, X_{d-1}) \end{aligned}$$

- Can be represented by a graph in which each node has links from all lower-numbered nodes (i.e. a fully connected graph)

Directed Acyclic Graphs

- No directed cycles ⇒ can number nodes so that there are no links from higher numbered to lower numbered nodes



Directed Acyclic Graphs (cont'd)

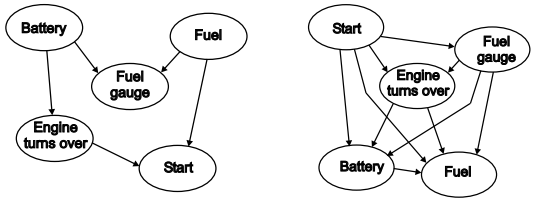
- General factorization

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i|\mathrm{pa}_i)$$

where pa_i denotes the parents of i

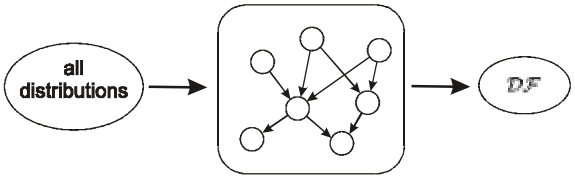
- Missing links imply conditional independencies
- Model specified by graph and by conditional probabilities
- Ancestral simulation can be used to sample from the joint distribution
- If a variable with no children is unobserved it can be removed from the graph to obtain a marginal distribution

Importance of Ordering



Directed Factorization Property

- A distribution which can be factored according to a particular directed graph is said to respect the directed factorization property DF
- View the graph as a filter

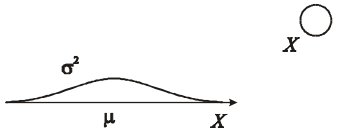


Examples of Directed Graphs

- Hidden Markov models
- Kalman filters
- Factor analysis
- Probabilistic principal component analysis
- Independent component analysis
- Mixtures of Gaussians
- Probabilistic expert systems
- Sigmoid belief networks
- Hierarchical mixtures of experts
- Etc…

Example: Univariate Gaussian

$$P(X) = \mathcal{N}(X|\mu, \sigma^2)$$



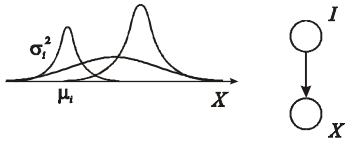
Example: Mixture of Gaussians

- Conditional distributions

$$P(I = i) = \pi_i$$
$$P(X|I = i) = \mathcal{N}(X|\mu_i, \sigma_i^2)$$

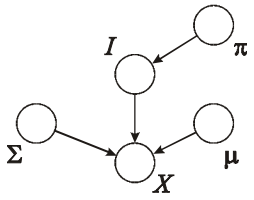
- Marginal distribution

$$P(X) = \sum_I P(X|I)P(I) = \sum_i \pi_i \mathcal{N}(X|\mu_i, \sigma_i^2)$$



Example: Bayesian Mixture of Gaussians

- Priors over parameters



Exponential Family

- How should we choose the conditional distributions?
- Considerable simplifications arise if we choose distributions from the exponential family

$$\ln P(X|\theta) = \phi_X(\theta)^{\mathrm{T}} \mathbf{u}_X(X) + f_X(X) + g_X(\theta)$$

- Includes many well-known distributions (Gaussian, Dirichlet, Gamma, Multi-nomial, Wishart, Bernoulli, …)
- Likelihood function (a function of θ) depends on data set only through *sufficient statistics* of fixed dimension

$$\sum_{n=1}^{N} \mathbf{u}_X(X_n)$$

Illustration: Uni-variate Gaussian

- Precision (inverse variance) $\tau = 1/\sigma^2$

$$P(X|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(X-\mu)^2\right\}$$

- In standard form

$$\begin{aligned} \phi_X &= \begin{bmatrix} \mu\tau \\ -\tau/2 \end{bmatrix} \\ \mathbf{u}_X &= \begin{bmatrix} X \\ X^2 \end{bmatrix} \\ g_X &= \frac{1}{2}\ln\left(\frac{\tau}{2\pi}\right) - \frac{1}{2}\tau\mu^2 \\ f_X &= 0 \end{aligned}$$

Conjugate Priors

- Choose prior distribution so that posterior distribution has same functional form as the prior

$$\ln P(\theta) = \phi_X(\theta)^{\mathrm{T}} \boldsymbol{\nu} + \eta g_X(\theta) + \text{const.}$$

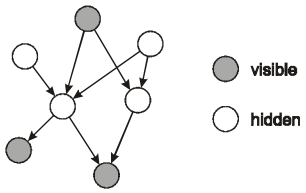
- Hence posterior is of the form

$$\ln P(\theta|X) = \phi_X(\theta)^{\mathrm{T}} [\mathbf{u}_X(X) + \boldsymbol{\nu}] + [1+\eta] g_X(\theta) + \text{const.}$$

- Can interpret prior as η effective observations of value $\boldsymbol{\nu}/\eta$
- Examples:
  - Gaussian for the mean of a Gaussian
  - Gamma for the precision of a Gaussian
  - Dirichlet for the parameters of a discrete distribution

Conditioning on Evidence

- Group variables into hidden (or latent) H and visible (or observed) V



- Hidden variables may have a specific interpretation, or may be introduced to permit a richer class of distribution

Overview of Part One

- Graphs and probabilities
- Directed graphs
- Markov properties
- Undirected graphs
- Examples

Conditional Independence

- Suppose A is independent of B given C

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

- Phil Dawid's notation

$$A \perp\!\!\!\perp B|C$$

- Equivalently

$$\begin{aligned} P(A,B|C) &= P(A|B,C)P(B|C) \\ &= P(A|C)P(B|C) \end{aligned}$$

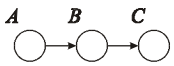
- Conditional independence crucial in practical applications since we can rarely work with a general joint distribution

Markov Properties

- Can we determine the conditional independence properties of a distribution directly from its graph?
- Yes: "d-separation"
- Start by considering three simple examples

Markov Properties: Example 1

- Joint distribution over 3 variables specified by the graph

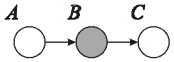


- Note the missing edge from A to C
- Node B is "head-to-tail" with respect to the path A–B–C
- Joint distribution

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

Markov Properties: Example 1 (cont'd)

- Suppose we condition on node B



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

- Hence

$$A \perp\!\!\!\perp C|B$$

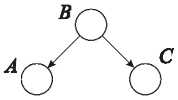
- Note that if B is not observed we have

$$A \not\perp\!\!\!\perp C \mid \emptyset$$

- Informally: observation of B "blocks the path" from A to C

Markov Properties: Example 2

- 3-node graph



- Joint distribution

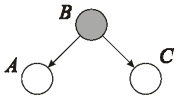
$$P(A,B,C) = P(B)P(A|B)P(C|B)$$

- Node B is "tail-to-tail" with respect to the path A–B–C
- Again, note missing edge from A to C

Markov Properties: Example 2 (cont'd)

- Now condition on node B
- We have

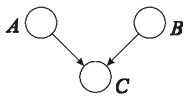
$$P(A,C|B) = P(A|B)P(C|B)$$



- Hence $A \perp\!\!\!\perp C|B$
- Again, if B is not observed $A \not\perp\!\!\!\perp C \mid \emptyset$
- Informally: observation of B "blocks the path" from A to C

Markov Properties: Example 3

- Node C is "head-to-head" with respect to the path A–C–B



- Joint distribution

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

- Note missing edge from A to B
- If C is *not* observed we have

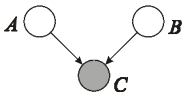
$$P(A, B) = P(A)P(B)$$

and hence $A \perp\!\!\!\perp B | \emptyset$

Markov Properties: Example 3 (cont'd)

- Suppose we condition on node C

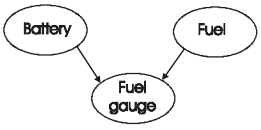
$$P(A, B|C) \neq P(A|C)P(B|C)$$



- Hence $A \not\perp\!\!\!\perp B \mid C$
- Informally: an *unobserved* head-to-head node C "blocks the path" from A to B, but once C is observed the path is unblocked
- Note: observation of any *descendent* of C also unblocks the path

Explaining Away

- Illustration



Observation of "Fuel Gauge" renders "Battery" and "Fuel" conditionally dependent

d-separation

- Stated here without proof
- Consider three *groups* of nodes A, B, C
- To determine whether the conditional independence statement $A \perp\!\!\!\perp B | C$ is true, consider all possible paths from any node in A to any node in B
- Any such path is blocked if there is a node Ω which is head-to-tail or tail-to-tail with respect to the path and $\Omega \in C$ or if the node is head-to-head and neither the node, *nor any of its descendents*, is in C

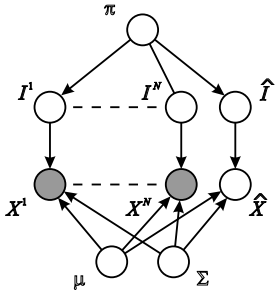
d-separation (cont'd)

- Note that a particular node may, for example, be head-to-head with respect to one particular path and head-to-tail with respect to a different graph
- If all possible paths are blocked then $A \perp\!\!\!\perp C | B$
- Consider this procedure as a filter applied to a probability distribution
- A distribution which satisfies all of the conditional independence properties implied by d-separation on the graph is said to respect the directed global Markov property $\mathcal{DG}$
- Theorem:

$$\mathcal{DF} \Leftrightarrow \mathcal{DG}$$

d-separation (cont'd)

- Illustration: Bayesian mixture of Gaussians



$$\mu \perp\!\!\!\perp \Sigma \mid \emptyset$$
$$\mu \not\perp\!\!\!\perp \Sigma \mid \{X^n\}$$

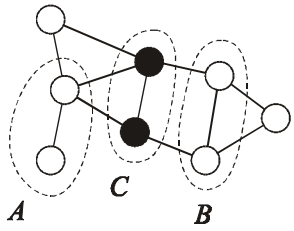
$$\hat{X} \perp\!\!\!\perp \{X^n\} \mid \mu, \Sigma, \pi$$
$$\hat{X} \not\perp\!\!\!\perp \{X^n\} \mid \emptyset$$

Overview of Part One

- Graphs and probabilities
- Directed graphs
- Markov properties
- Undirected graphs
- Examples

Undirected Graphs

- Simpler definition of separation for undirected graphs
- Sets A and B of nodes are separated by a third set C if every path from any node in A to any node in B passes through a node in C



Undirected Graphs (cont'd)

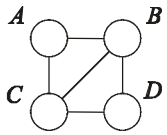
- The property of undirected graph separation says that if A and B are separated by C in the graph then

$$A \perp\!\!\!\perp B | C$$

- We can consider a graph as a filter: any distribution which satisfies all of the conditional independence statements implied by the graph is said to satisfy the undirected global Markov property $\mathcal{G}$ with respect to that graph

Undirected Factorization

- Definitions: (i) a set of nodes is *complete* if there is a link from each node to every other node in the set; (ii) a *clique* is a maximal complete set of nodes
- Example: the following graph has cliques $\{A,B,C\}$ and $\{B,C,D\}$



Undirected Factorization (cont'd)

- A probability distribution is said to factorize with respect to a given undirected graph if it can be expressed as the product of positive functions over the cliques of the graph

$$P(X) = \frac{1}{Z} \prod_C \psi_C(X_C)$$

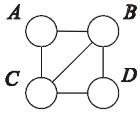
where $\psi_C(X_C)$ are the *clique potentials,* and Z is a normalization constant

Undirected Factorization (cont'd)

- A slightly more general representation (which will come in useful later) is the product of clique potentials divided by the *separator* potentials (a separator between two cliques is the set of nodes they have in common)

$$P(X) = \frac{1}{Z} \frac{\prod_C \psi_C(X_C)}{\prod_S \phi_S(X_S)}$$

- For the previous example the cliques are $\{A,B,C\}$ and $\{B,C,D\}$, and the separator set is $\{B,C\}$

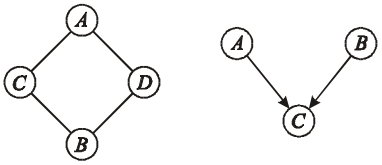


Undirected Factorization (cont'd)

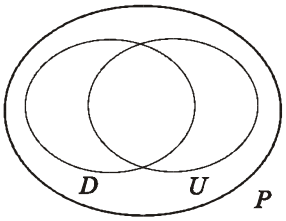
- A distribution which factorizes according to a particular graph is said to respect the undirected factorization property $\mathcal{F}$
- Theorem: for any graph and any distribution $\mathcal{F} \Rightarrow \mathcal{G}$
- Theorem (Hammersley-Clifford): for strictly positive distributions and arbitrary graphs $\mathcal{G} \Leftrightarrow \mathcal{F}$
- Also $\mathcal{G} \Leftrightarrow \mathcal{F}$ for any distribution if, and only if, the graph is *triangulated* (see later)

Directed versus Undirected

- There exist undirected graphs which cannot be re-expressed as directed graphs, and vice versa
- Examples:

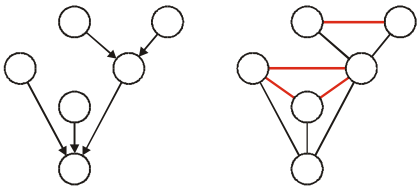


Directed versus Undirected (cont'd)



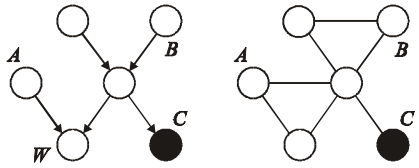
Directed Markov Revisited

- We can now formulate a simpler definition of $\mathcal{DG}$ with the aid of undirected graph separation
- Simply dropping the arrows and applying undirected graph separation clearly fails because of explaining away
- We can resolve this by adding links which connect all of the parents for each node – called *moralization*
- Example:



Directed Markov Revisited (cont'd)

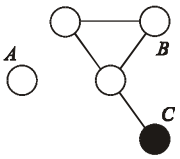
- However, moralization alone would suppress some of the conditional independencies, e.g. $A \perp\!\!\!\perp B | C$ in the graph



- The problem arises from the node W since it is not part of the conditioning set

Directed Markov Revisited (cont'd)

- Definition: a subset of nodes within a DAG is an *ancestral* set if, for every node in the set, all ancestors of that node are also in the set
- Theorem: if a probability distribution factorizes according to a directed acyclic graph, then $A \perp\!\!\!\perp B | C$ whenever A and B are separated by C in the moral graph of the smallest ancestral set containing $A \cup B \cup C$

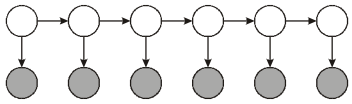


Overview of Part One

- Graphs and probabilities
- Directed graphs
- Markov properties
- Undirected graphs
- Examples

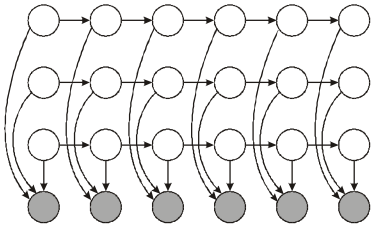
Example: State Space Models

- Hidden Markov model
- Kalman filter



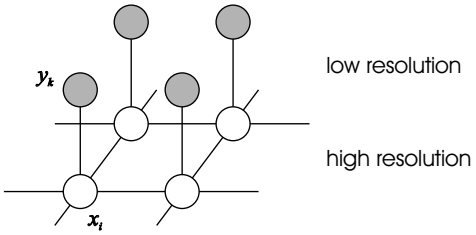
Example: Factorial HMM

- Multiple hidden sequences
- Avoid exponentially large hidden space



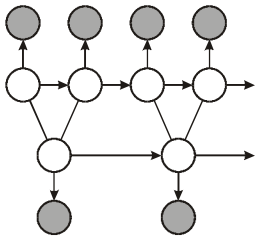
Example: Markov Random Field

- Application to image super-resolution (Freeman *et al.*)



Example: Coupled HMM

- Fusion of audio and video signals



Summary of Part One

- Probabilistic graphs provide insights into existing models and motivate the design of new models
- Directed and undirected graphs together encompass a wide range of models of practical interest
- The conditional independence (Markov) properties of complex distributions can be determined graphically

Viewgraphs and tutorials available from:

research.microsoft.com/~cmbishop