Undirected Graphical Models

Chris Williams

School of Informatics, University of Edinburgh

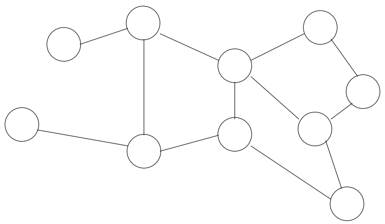
November 2009

Overview

- Undirected graphs
- Potential functions, energy functions
- Conditional independence
- Examples: multivariate Gaussian, MRF
- Boltzmann machines, learning rule
- Reading: Bishop §8.3, Jordan section 2.2.

Undirected Graphs

- graph $G = (X, E)$
- X is a set of nodes, in one-to-one correspondence with a set of random variables
- E is a set of undirected edges between the nodes



Graphs and Cliques

- For directed graphs use $P(\mathbf{X}) = \prod_i P(X_i|Pa_i)$, gives notion of locality
- For undirected graphs, locality depends on the notion of *cliques*
- A clique of a graph is a fully-connected set of nodes
- A maximal clique is a clique which cannot be extended to include additional nodes without losing the property of being fully connected

Parameterization

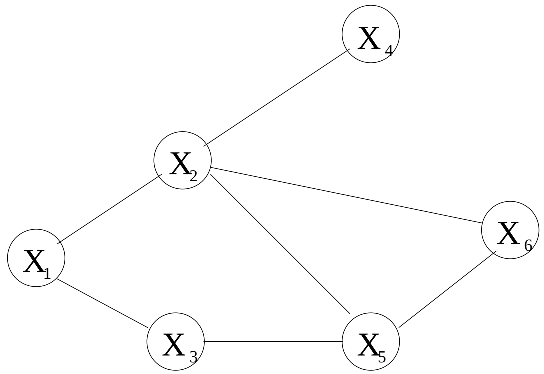
- Joint probability distribution is given as a product of local functions defined on the maximal cliques of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(\mathbf{x}_C)$$

with

$$Z = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_{X_C}(\mathbf{x}_C)$$

- Each $\psi_{X_C}(\mathbf{x}_C)$ is a strictly positive, real-valued function, otherwise arbitrary
- Z is called the partition function



P(x) = Ψ (x1,x2) ψ (x1,x3) ψ (x3,x5) ψ (x2,x5,x6) ψ (x2,x4) /Z

- Potential functions are in general neither conditional or marginal probabilities
- Natural interpretation as agreement, constraint, energy
- Potential function favours certain local configurations by assigning them larger values
- Global configurations that have high probability are, roughly speaking, those that satisfy as many of the favoured local configurations as possible

Energy functions

- Enforce positivity by defining

$$\psi_{X_C}(\mathbf{x}_C) = \exp\{-E_{X_C}(\mathbf{x}_C)\}$$

- Negative sign is conventional (high probability, low energy)

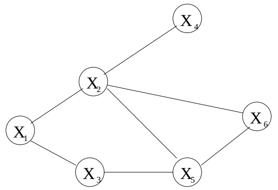
$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(\mathbf{x}_C) = \frac{1}{Z} \exp\{-\sum_{C \in \mathcal{C}} E_{X_C}(\mathbf{x}_C)\}$$

- Energy $E(\mathbf{x}) = \sum_{C \in \mathcal{C}} E_{X_C}(\mathbf{x}_C)$
- Boltzmann distribution

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

Local Markov Property

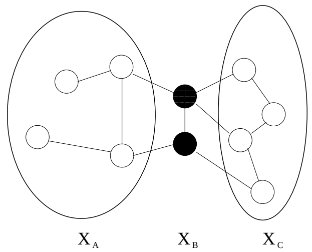
- Denote all nodes by V
- For a vertex a, let ∂a denote the boundary of a, i.e. the set of vertices in $V \backslash a$ that are neighbours of a
- **Local Markov property**: For any vertex a, the conditional distribution of X_a given $X_{V \backslash a}$ depends only on $X_{\partial a}$



P(x) = Ψ (x1,x2) ψ (x1,x3) ψ (x3,x5) ψ (x2,x5,x6) ψ (x2,x4) /Z

Global conditional independence

- Consider arbitrary disjoint index subsets A, B and C
- If every path from a node in X_A to a node in X_C includes at least one node in B then $I(X_A, X_C | X_B)$
- This is a naïve graph-theoretic separation condition (c.f. d-separation)
- Equivalence of conditional independence and clique factorization form is the Hammersley-Clifford theorem



Exact Inference in Undirected Graphical Models

- Triangulate the graph if necessary
- Use the junction tree algorithm discussed earlier

Approximate Inference: Gibbs sampler

Loop T times
    for each unit i to be sampled from
        sample $P(X_i | rest)$
    end for
end loop

- This is a Markov Chain Monte Carlo (MCMC) method. Under general conditions this will converge to the correct distribution as $T \to \infty$
- More general MCMC schemes are possible (e.g. Metropolis-Hastings)

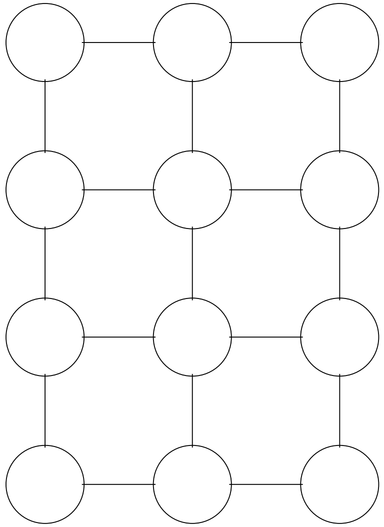
Example I—Multivariate Gaussian

$$p(\mathbf{x}) \propto \exp\{-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}\}$$

- It is the zeros in Σ^{-1} that define the missing edges in the graph and hence the conditional independence structure

Example II—Markov Random Field

- Discrete random variables
- Ising model in statistical physics (spins up/down)
- MRF models used in image analysis, e.g. segmentation of regions. Define energies such that blocks of the same labels are preferred (Geman and Geman, 1984)



Example: GrabCut

- C. Rother, V. Kolmogorov, A. Blake. GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. SIGGRAPH'04, 2004
- Builds Gaussian mixture models of foreground and background pixels, and uses MRF prior on foreground label field

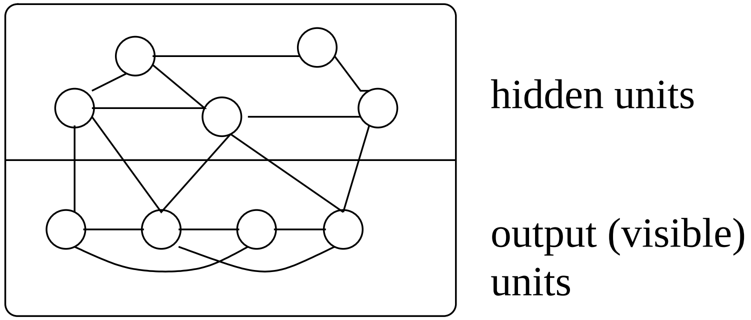
Figure acknowledgement: MSR Cambridge GrabCut page

Boltzmann machines

- Hinton and Sejnowski (1983)
- Binary units ± 1

$$p(\mathbf{x}) = \frac{1}{Z} \exp\{\frac{1}{2} \sum_{ij} w_{ij} x_i x_j\}$$

- $w_{ij} = w_{ji}$ and $w_{ii} = 0$
- set $x_0 = 1$ (bias unit)
- $\frac{1}{2} \sum_{ij} w_{ij} x_i x_j = \sum_{i<j} w_{ij} x_i x_j$
- Can have *hidden* units
- Potential function is not arbitrary function of cliques, but only based on pairwise links (can generalize)
- $P(X_i = 1 | rest) = \sigma(2h_i)$ where $h_i = \sum_j w_{ij} x_j$



Boltzmann machine learning rule

Denote visible units by $\mathbf{x}$, hidden units by $\mathbf{y}$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{\sum_k \theta_k \phi_k(\mathbf{x}, \mathbf{y})\}$$

This is the general form of a *log linear* model.

- Features $\phi_k(\mathbf{x}, \mathbf{y})$ are the pairwise potentials for a Boltzmann machine
- Parameters θ_k correspond to weights in the Boltzmann machine

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{\sum_k \theta_k \phi_k(\mathbf{x}, \mathbf{y})\}$$

$$p(\mathbf{x}) = \frac{1}{Z} \sum_{\mathbf{y}} \exp\{\sum_k \theta_k \phi_k(\mathbf{x}, \mathbf{y})\}$$

$$\log p(\mathbf{x}) = \log \sum_{\mathbf{y}} \exp\{\sum_k \theta_k \phi_k(\mathbf{x}, \mathbf{y})\} - \log Z$$

$$\frac{\partial \log p(\mathbf{x})}{\partial \theta_l} = \sum_{\mathbf{y}} \phi_l(\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}) - \sum_{\mathbf{x},\mathbf{y}} \phi_l(\mathbf{x}, \mathbf{y}) p(\mathbf{x}, \mathbf{y})$$

$$\stackrel{def}{=} \langle \phi_l(\mathbf{x}, \mathbf{y}) \rangle^+ - \langle \phi_l(\mathbf{x}, \mathbf{y}) \rangle^-$$

- $+$ denotes the *clamped* phase (with $\mathbf{x}$ clamped on visible units), $-$ denotes the *free-running* phase (all unclamped)
- Learning stops when statistics match in both phases
- Statistics could be computed exactly (using junction tree algorithm) but often this is intractable—use stochastic sampling
- Boltzmann machine learning can be slow due to the need to use MCMC techniques. Gradient is the *difference* of two noisy estimates
- In Restricted Boltzmann Machines (RBMs), where there is a layer of visible units and a layer of hidden units with bipartite connections, learning can be more efficient (Hinton, 2002)