Probabilistic Graphical Models

Lecture Notes
Fall 2009

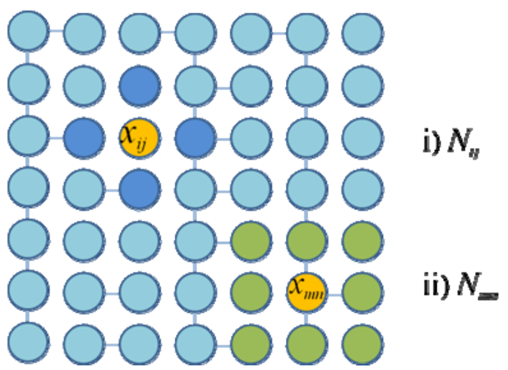
October 28, 2009

Byoung-Tak Zhang
School of Computer Science and Engineering &
Cognitive Science, Brain Science, and Bioinformatics
Seoul National University
http://bi.snu.ac.kr/~btzhang/

Chapter 5. Markov Networks

5.1 Markov Random Fields

- Random field (Ω, Λ, *P*)
  - Ω: a set of all possible configurations of the random variables *X*
  - Λ: a discrete two-dimensional rectangular *lattice*
  - X_{mn}: a random variable defined on Λ that takes on the values x_{mn} at lattice site (m,n).
  - *X*: *configuration* of a lattice system, i.e. the set of values of the random variables
  - *P*: joint probability measure



$$P(X = \mathbf{x}) = \frac{1}{Z} \prod_{s=(i,j)} \psi_{\eta_s}(\mathbf{x}_{\eta_s})$$

i) The order of neighborhood system N_{ij} is 4
ii) The order of neighborhood system N_{mn} is 8

- Neighborhood system
  - No causality
  - Several different *orders* of neighborhood system
  - A neighborhood system N_{ij} associated with a lattice Λ (like undirected graph)

$$N_{ij} = \left\{ \eta_{ij} \subset \Lambda; (i,j) \in \Lambda \right\}$$

$$(i,j) \notin \eta_{ij}$$

$$(k.l) \in \eta_{ij} \Rightarrow (i,j) \in \eta_{kl}$$

η_{ij} is the neighborhood system for lattice site (i, j).

- **Markov random fields**
  - Mathematical generalization of the notion of a one-dimensional **temporal** Markov chain to a two-dimensional (or higher) **spatial** lattice or graph.

  - A generalization of the Ising model.

  - A random field for which the *joint* probability distribution has associated *conditional* probabilities that are local, i.e. having the spatial Markovian relationship like:

    $$P\left(X_{mn}=x_{mn} \mid X_{rs}=x_{rs}, rs \neq mn\right)=P\left(X_{mn}=x_{mn} \mid X_{rs}=x_{rs}, rs \in \eta_{mn}\right)$$

    where η_{mn} is the neighborhood system for lattice site (m,n).

  - The probability distribution is positive definite for all values of the random variable.
  - The conditional probabilities are invariant with respect to neighborhood translations.

5.2 Gibbs Random Fields

- Defining a random field by Gibbs distribution

- Gibbs random field (Ω, Λ, *P*)
  - *Joint* probability distribution is of the form

    $$P\left(X=\mathbf{x}\right)\equiv P\left(\mathbf{x}\right)=\frac{1}{Z}\exp\left(-\frac{1}{T}U\left(\mathbf{x}\right)\right)$$

  - **x**: configuration
  - $U(\mathbf{x})$: potential encapsulating the *global* properties of the system
  - Z: partition function

    $$Z=\sum_{\mathbf{x}\in\Omega}\exp\left(-\frac{1}{T}U\left(\mathbf{x}\right)\right)$$

  - T: temperature (as in simulated annealing)

- Potentials
  - $U(\mathbf{x})$: sum of individual *local* contributions $V_s(x_s)$ from each lattice site

    $$U\left(\mathbf{x}\right)=\sum_{s\in\Lambda}V_s\left(x_s\right)$$

    $$V_s\left(x_s\right)=\sum_{c\in C_s}V_c\left(x_c\right)$$

  - $V_c(x_c)$: clique potential
  - C_s: the set of all cliques associated with the site s, e.g. $s=(i,j)$ in a rectangular lattice

- Gibbs random field is a generalization of the nearest-neighbor interactions of the Ising model
  - In the Ising model
    - Single site denoting interactions of a spin element with an external field
    - Two-body terms denoting adjacent spin elements
  - In the Gibbs random field
    - **Reflecting more elaborate sets of interactions**
    - Single, pair, and three body interactions, …

5.3 Hammersley-Clifford Theorem

- A theorem showing that the *global* character of a Gibbs random field (defined through *local* interactions) is equivalent to the *purely local* character of a Markov random field.

- Gibbs random field == Markov random field

- For finite lattices and graphs

- Proofs
  - Polynomial expansion by Besag (Beckerman, Ch. 6.3)
  - Equivalence based on Möbius inversion by Grimmett (Beckerman, Ch. 6.4)
  - Other proofs

- Homework: Prove the HC Theorem

5.4 Markov Networks

- A graphical model in which a set of random variables have Markov property described by an undirected graph.
- Often used interchangeably with the term *Markov random field*

- An undirected graphical model representing a joint probability distribution of random variables by the product of potential functions defined over the maximal cliques of the graph

$$P(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathbf{x}_C)$$ Joint distribution

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$ Potential function

$$Z = \sum_{\mathbf{x}}\prod_{C}\psi_C(\mathbf{x}_C)$$ Partition function (normalization constant)

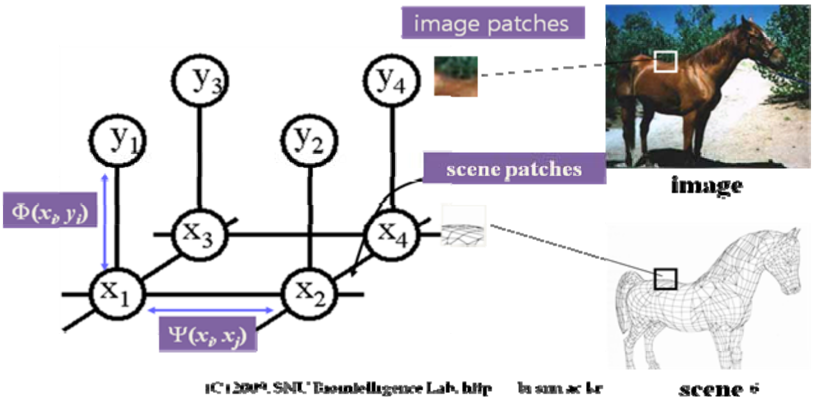
  - Clique C: a subset of the nodes in a graph s.t. there exists a link btw all pairs of nodes in the subset
  - Potential function $\psi_C(\mathbf{x}_C)$: Functions of the maximal cliques that are the factors in the decomposition of the joint distribution

- Potential functions are not restricted to marginal or conditional distribution
- Normalization constant: major limitation of undirected graph. But we can overcome when we focus on *local conditional* distribution

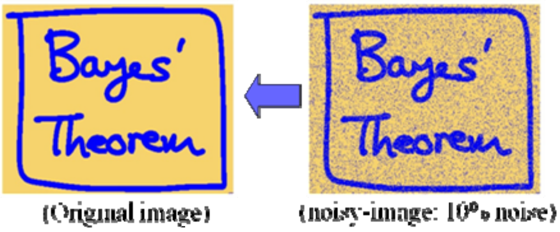
5.5 MRFs for Images

- Two dimensional images highly organized spatial systems. MRFs are used for
  - texture analysis
  - synthesis
  - reconstruction
  - segmentation

MRF nodes as patches



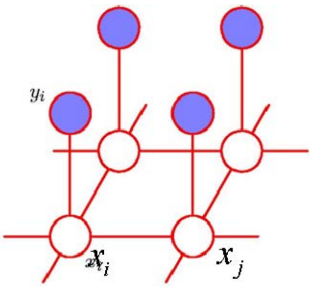
- Image denoising
  - Goal: to recover the original noise-free image



- Setting
  - Image as a set of 'binary pixel values' {-1, +1}
  - In the observed noisy image
  - In the unknown noise-free image
  - Noise: randomly flipping the sign of pixels with some small probability

- Prior knowledge (when the noise level is small)
  - Strong correlation between x_i and y_i
  - Strong correlation between neighboring pixels x_i and x_j

- Corresponding Markov random field
  - A simple energy function for the cliques
    - form $\{x_i, y_i\}$: $-\eta x_i y_i$
    - form $\{x_i, x_j\}$: $-\beta x_i x_j$:
  - Bias (preference of one particular sign): hx_i



y: known noisy image
x: unknown noise-free image

- The complete energy function for the model / joint distribution

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

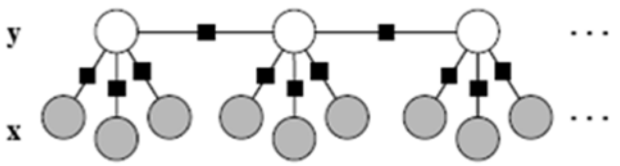
$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}.$$

- Image restoration results
    - Iterated conditional modes (ICM)
        - Coordinate-wise gradient ascent
        - Initialization: $x_i = y_i$ for all i
        - Take one node, evaluate the total energy, change the state of the node if it results in lower energy
        - Repeat till some stopping criterion is satisfied
    - Graph-cut algorithm
        - Guaranteed to find the global maximum in Ising model

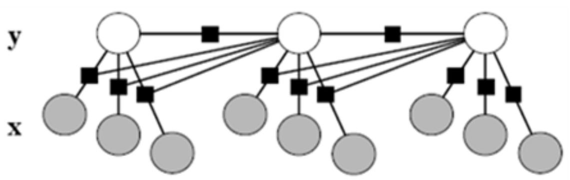
- Inference algorithms for MRFs
    - Iterated conditional modes (ICM)
    - Gibbs sampling
    - Simulated annealing
    - Variational methods
    - Belief propagation
    - Graph cuts

5.6 Conditional Random Fields

- Conditional random fields (CRF) are a probabilistic model that directly models the conditional distribution $p(\mathbf{y}|\mathbf{x})$



Graphical models of an HMM-like linear-chain CRF



Graphical models of a linear-chain CRF in which the transition score depends on the current observation

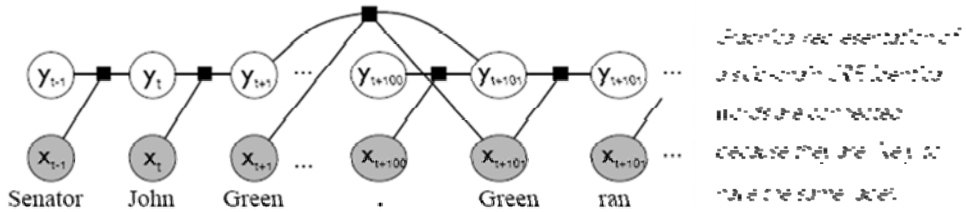
- **A linear-chain conditional random field**

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

- **y**, **x**: random vectors
- $\Lambda = \{\lambda_k\} \in \Re^K$: a parameter vector
- $\{f_k(y, y', \mathbf{x}_t)\}_{k=1}^{K}$: a set of real-valued feature functions
- The main difference from the generative model: a conditional distribution *p*(**y**|**x**) does not include a model of *p*(**x**).
- The difficulty in modeling *p*(**x**) is that it often contains many highly dependent features which are difficult to model.
-

- **Skip-chain CRF**
  - a linear-chain CRF with *additional long-distance edges* between similar words (*skip edges*) to model certain kinds of long-range dependencies between entities.



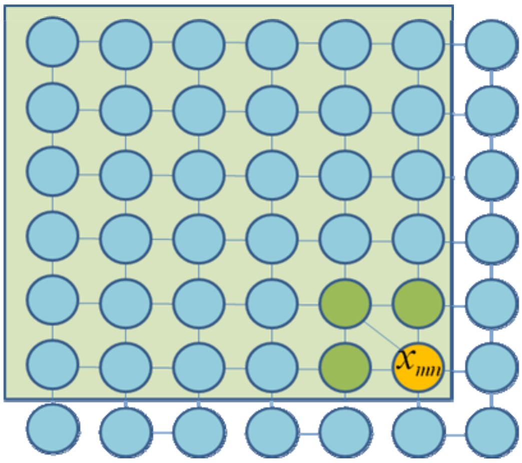
- defined as a general CRF with two clique templates: one for the linear-chain portion, and one for skip edges.

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Psi_t(y_t, y_{t-1}, \mathbf{x}) \prod_{(u,v)\in\mathcal{I}} \Psi_{uv}(y_u, y_v, \mathbf{x})$$

$$\Psi_t(y_t, y_{t-1}, \mathbf{x}) = \exp\left\{\sum_k \lambda_{1k} f_{1k}(y_t, y_{t-1}, \mathbf{x}, t)\right\}$$

$$\Psi_{uv}(y_u, y_v, \mathbf{x}) = \exp\left\{\sum_k \lambda_{2k} f_{2k}(y_u, y_v, \mathbf{x}, u, v)\right\}$$

5.7 Other Random Field Models



- Causal random field models
  - Markov mesh random field (Abend, Harley, and Kanal)
  - Pickard random field
  - Causal random field
    - A random variable X_{ij} in an image conditioned on random variables in upper and left region will depend only on random variables at sites immediately above and to the left.

      $$p\left(x_{ij} \mid x_{kl};(k,l)\in\Phi_{ij}\right) = p\left(x_{ij} \mid x_{kl};(k,l)\in S_{ij}\right)$$

    - Φ_{ij}: the larger segment of the lattice above and to the left of a site (i,j)
    - S_{ij}: the small set of the segment called the support set

- Noncausal models
  - Gauss-Markov random fields
  - Simultaneous autoregressive (SAR) model (can be either causal or noncausal)