Graphical Models

Sung-Yub Kim

Dept of IE, Seoul National University

April 2, 2017

Bishop, C. M. Pattern Recognition and Machine Learning *Information Science and Statistics*, Springer, 2006.

- **Properties of Graphical Models**
  1. Provide a simple way to visualize the structure of a probabilistic model
  2. Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graphs.
  3. Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations.
- **Types of Graphical Models**
  1. Bayesian Networks(Directed Graphical Models): useful for expressing causal relationships between random variables.
  2. Markov random field(Undirected Graphical Models): better suited to expressing soft constraints between random variables.

- **Motivation**
  Generally, we can represent arbitrary joint distribution using conditional distribution.

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

  We can represent this as a directed graph having K nodes, one for each conditional distribution on the right side of above equation. We say that this graph is *fully connected*. Of course, it is the *absence* of the links in the graph that conveys interesting information about the properties of the class of distributions that the graph represents.

- **Joint Distribution of Directed Graph models**
  For a graph with K nodes, the joint distribution is given by

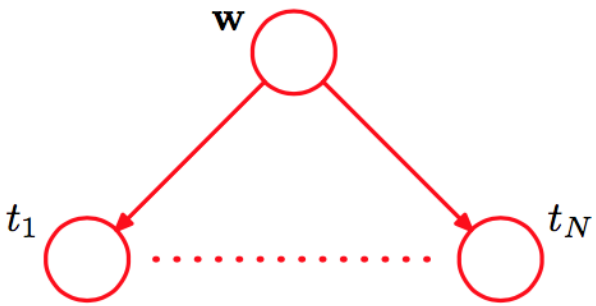
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|pa_k) \tag{2}$$

  where pa_k denotes the set of parents of x_k. Also our interests are focused in Directed Acyclic Graphical models, which have no cycle in graph.

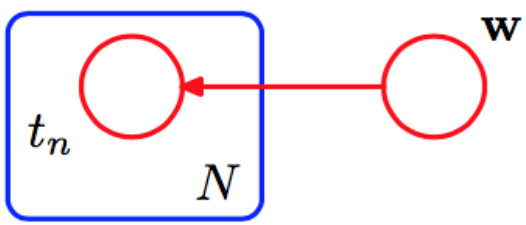
- **Basic Causal relationship**
  We can easily represent this joint distribution

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}) \tag{3}$$

by



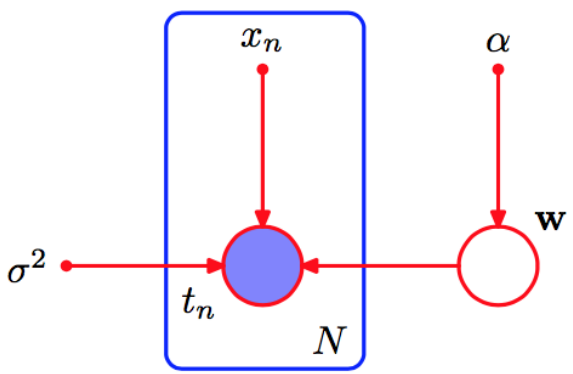
or using plate



- **Deterministic variables and observations**
  We use smaller solid circles to represent deterministic variables. Also we denote observed variables by shading the corresponding nodes. For example,

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2) \tag{4}$$

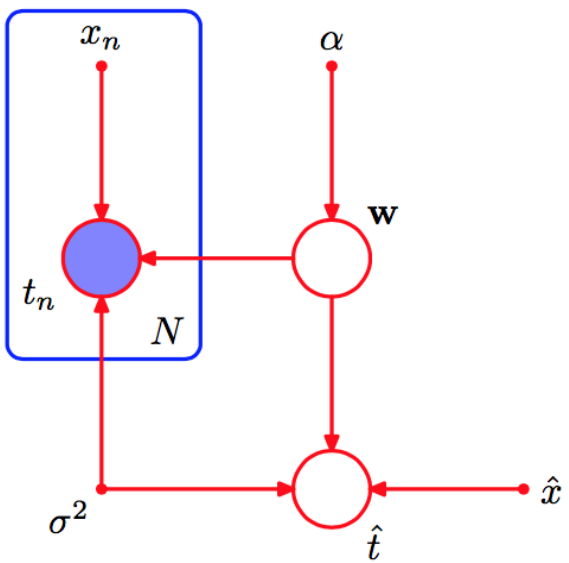
  can be shown as



- **Predictive distribution**
  Our final interest is predictive distribution

$$p(\hat{t}, \mathbf{t}, \mathbf{w}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2) p(\hat{t}|\hat{x}, \mathbf{w}, \sigma^2) \qquad (5)$$

  can be shown as



- **Ancestral sampling**
  First, start with the lowest-numbered node and draw a sample from the distribution $p(x_1)$. We then work through each of nodes in oreder, so that for node n we draw a sample from the conditional distribution $p(x_n|pa_n)$.
- **Marginal Sampling**
  To obtain a sample from some marginal distribution corresponding to a subset of the variables, we simply take the sampled values for the required nodes and **ignore** the sampled values for the remaining nodes.

- **Exponential family and Directed Graphical Models**
  If we choose the relationship between each parent-child pair in a directed graph to be conjugate, then the models have particularly nice properties. Especially, when they each correspond to discrete variables and Gaussian variables this relationship can be extended hierarchically to arbitrarily complex directed acylcic graphs.
- **Independency**
  If we model full-joint probabily distribution, only two discrete variable has K states needs $K^2 - 1$ parameters

$$p(\mathbf{x}_1, \mathbf{x}_2|\mu) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}} \tag{6}$$

  But if we can add independent assumption between $\mathbf{x}_1$ and $\mathbf{x}_2$, we only need $2(K-1)$ parameters. An alternative way to reduce the number of independent parameters in a model is by *sharing* parameters.

- **Dirichlet prior**
  We can turn a graph over discrete variables into a Bayesian model by introducing Dirichelt priors for the parameters. Thus, each node acquires an additional parent representing the Dirichlet distribution over the parameters associated with the corresponding discerete node.
- **Use parameterized model**
  This strategy can be used to control exponential growth in models of discrete variables. If we consider $p(y|x_1, \ldots, x_M)$, full-discrete model needs 2^M parameters. But if we use logistic sigmoid to model the conditional distribution, then we only need $M + 1$ parameters.

- **Definition**
  a is conditionally independent of b given c means

$$p(a|b,c) = p(a|c) \tag{7}$$

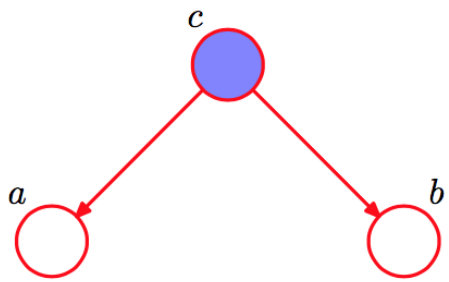
  or equivalently

$$p(a,b|c) = p(a|c)p(b|c) \tag{8}$$

  And we denote this relation by

$$a \perp\!\!\!\perp b|c \tag{9}$$

- **Tail-to-Tail**
  First, consider



By graph, we know

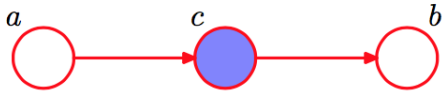
$$p(a, b, c) = p(c)p(a|c)p(b|c) \tag{10}$$

Therefore, we get

$$p(a, b|c) = p(a|c)p(b|c) \tag{11}$$

By the definition of CI, we know that $a \perp\!\!\!\perp b|c$ satisfy.

- **Head-to-Tail**
  Next, consider



By graph, we know

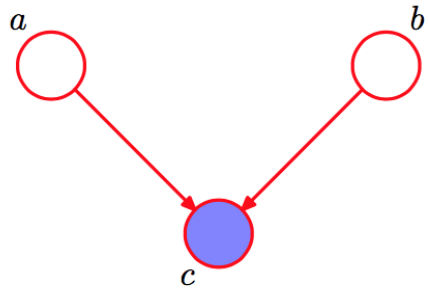
$$p(a, b, c) = p(a)p(c|a)p(b|c) \tag{12}$$

By Bayes Thm, we get

$$p(a, b|c) = p(a|c)p(b|c) \tag{13}$$

By the definition of CI, we know that $a \perp\!\!\!\perp b|c$ satisfy.

- **Head-to-Head**
  Finally, consider



By graph, we know

$$p(a, b, c) = p(a)p(b)p(c|a, b) \tag{14}$$

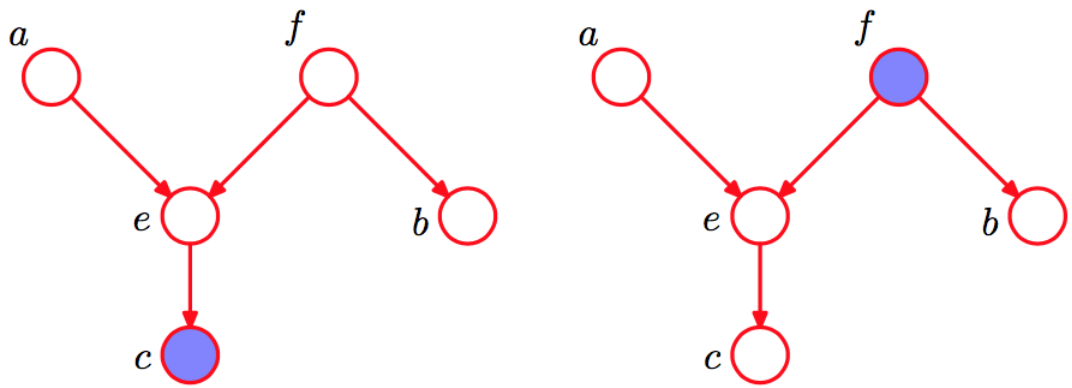
By marginalzing c, we get

$$p(a, b) = p(a)p(b) \tag{15}$$

Therefore, we get $a \perp\!\!\!\perp b|\phi$. However, condition on c we cannot get any independence between a and b. In short, a H2H path will become unblocked if either the node, or *any of its descendants*, is observed.

- **Blocked**
  $A \perp\!\!\!\perp B | C$ satisfy if all possible paths from any node in A to any node in B are blocked. We say that a path is blocked if it includes a node such that either
  1. the arrows on the path meet either H2T or T2T at the node, and the node is in the set C, or
  2. the arrows meet H2H at the node, and neither the node, nor any of its descendants, is in the set C.

  In following diagrams, left case is not blocked and right case is blocked

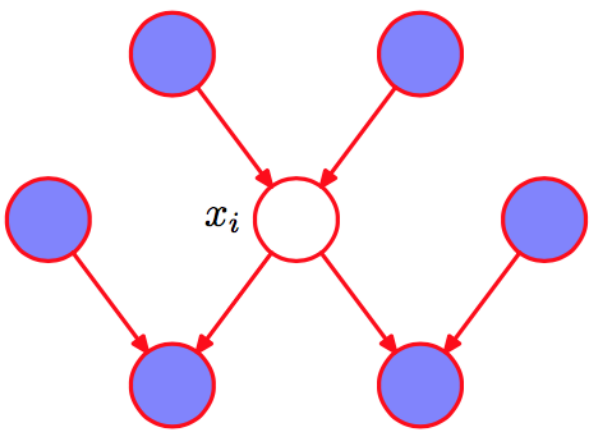


- **Markov Blanket**

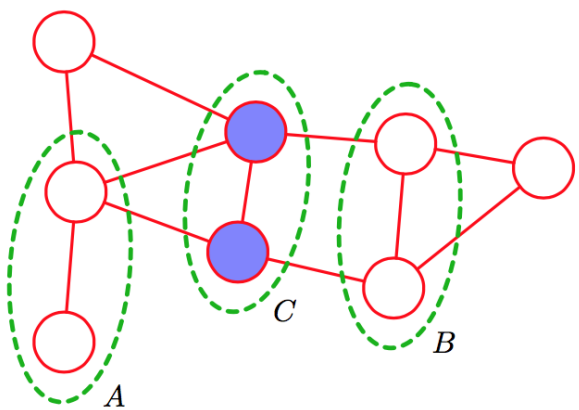
  Using factorization, we can express condistional distribution in the form

$$p(x_i|x_{\{j\neq i\}}) = \frac{p(x_1,\ldots,x_D)}{\int p(x_1,\ldots,x_D)dx_i} = \frac{\prod_k p(x_k|pa_k)}{\int \prod_k p(x_k|pa_k)dx_i} \tag{16}$$

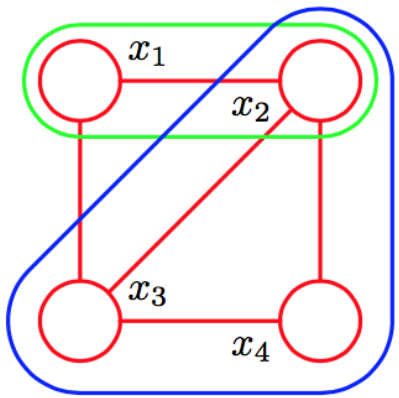
  In this fraction, the only factors that remain will be the conditional distribution $p(x_i|pa_i)$ for node x_i itself, together with the conditional distributions for any nodes x_k such taht node x_i is in the conditioning set of $p(x_k|pa_k)$, in other words for which x_i is a parent of x_k.



- **CI in Undirected Graphical Models**
  To thet whether $A \perp\!\!\!\perp B|C$ or not, we 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.
  An alternative way to view the conditional independence test is to imagine removing all nodes in set C from the graph together with any links that connect to the nodes. If there exists a path that connects any node in A to any node in B.



- **Clique**
  A subset of nodes in a graph such that there exists a link between all pairs of nodes in the subset. In other words, the set of nodes in a clique is fully connected.
- **Maximal Clique**
  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.



- **Potential function and partition function**
  Denote a clique by C and the set of variables in that clique by x_C. Then the joint distribution is written as a product of potential functions $\psi_C(\mathbf{x}_C)$ over the maximal cliques of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C) \tag{17}$$

  Here the quantity Z, sometimes called the partition function, is a normalization constant is given by

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C) \tag{18}$$

  The presence of this normalization constant is one of the major limitations of undirected graphs. However, for evaluation of local conditional distributions, the partition function is not needed because **a conditional is the ratio of two marginals**. Similarly, for evaluating local marginal probabil- ities we can work with the unnormalized joint distribution and then normalize the marginals explicitly at the end.

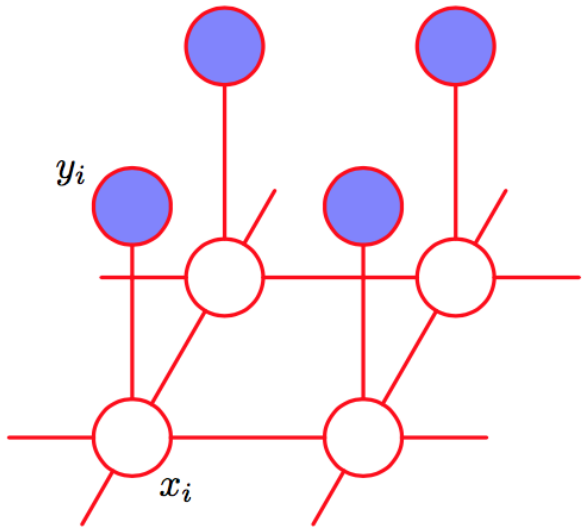
- **Hammersley-Clifford Theorem**
  $\mathcal{UI}$: set of such distributions that are consistent with the set of conditional independence statements that can be read from the graph using graph separation.
  $\mathcal{UF}$: set of such distributions that can be expressed as a factorization with respect to the maximal cliques of the graph.
  H-C Theorem states that the sets $\mathcal{UI}$ and $\mathcal{UF}$ are identical.

- **Energy function and Boltzmann distribution**
  Since potential functions are strictly positive it is convenient to express them as exponentials, so that

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\} \tag{19}$$

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

- **Problem**
  1. $y_i \in \{-1, +1\}$: noisy-image
  2. $x_i \in \{-1, +1\}$: noise-free-image
- **UGM**



Note that this UGM has two types of cliques.

1. $\{x_i, y_i\}$: choose a very simple energy function for these cliques of the form $-\eta x_i y_i$ where η is a positive constant. This has the desired effect of giving a lower energy when xi and yi have the same sign.
2. $\{x_i, x_j\}$: choose an energy given by $-\beta x_i x_j$ where β is a positive constant.

- **Energy function**

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i \tag{20}$$

- **Converting Process**
  Clique potentials of the undirected graph are given by the conditional distributions of the directed graph. For nodes on the directed graph having just one parent, this is achieved simply by replacing the directed link with an undirected link. And for nodes in the directed graph having more than one parent, add extra links between all pairs of parents of the node. This process of **marrying the parents** has become known as **moralization**.
- **Meaning of moralization**
  We saw that in going from a directed to an undirected representation we had to discard some conditional independence properties from the graph. The process of moralization **adds the fewest extra links** and so retains the maximum number of independence properties.
- **Map**
  1. D map : If every conditional independence statement satisfied by the distribution is reflected in the graph.(ex. completely disconnected)
  2. I map: If every conditional indepen- dence statement implied by a graph is satisfied by a specific distribution.(ex. fully connected)
  3. Perfect map: every conditional independence property of the distribution is reflected in the graph, and vice versa.