Machine Learning (ML, F16)	Lecture#07 (Thursday Nov. 3rd)
Lecturer: Byron Boots	*Undirected Graphical Models*

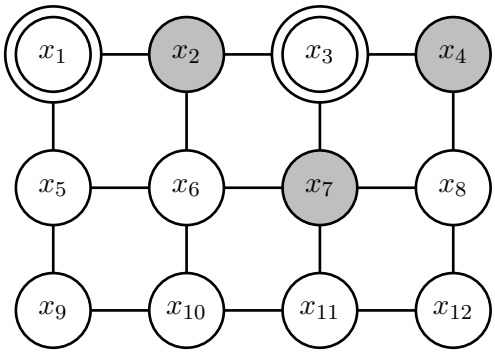
1 Undirected Graphical Models

In the previous lecture, we discussed directed graphical models. In this lecture we will discuss graphical representations of joint probability distributions defined on the basis of undirected graphs. This gives us a different, often simpler perspective on independence within a joint distribution.

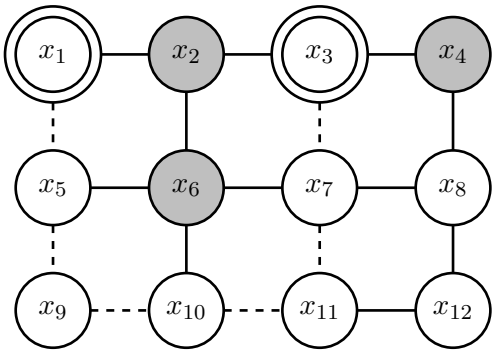
2 Markov Random Fields (MRFs)

A Markov Random Field (MRF) is an **undirected** graphical model that explicitly expresses the **conditional independence relationships between nodes**. As with a Bayes' Net, *fewer* connections means *more* structure.

Two nodes are conditionally independent if all paths between them are blocked by given nodes. See Figures 1(a) and 1(b) for examples. Note that this rule is much simpler than for Bayes' Nets. Due to the way that Markov Random Fields express the relationship between nodes, they make a lot of sense as a representation in problems where there is little structure to guide the construction of a directed graph (e.g. a representation of physical space).



(a) Node x_1 is conditionally independent to node x_3 given x_2, x_4 and x_7. There is no unblocked path between x_1 and x_3.



(b) There is at least one path from x_1 to x_3 given x_2, x_4 and x_6; x_1 and x_3 are not (necessarily) conditionally independent.

Figure 1: Markov Random Field example.

2.1 Rules of Conditional Independence

The rules for conditional independence are the following:

- **Local rule:** A variable is conditionally independent of *all* other variables given all its neighbors. Fig. 1(a) shows the *Markov blanket* as all paths into and out of x_3 are "blanketed".
- **Global rule:** If all paths between two sets of variables A and B pass through another set of nodes S. Then any two subsets of A and B are conditionally independent if S is observed.

For the example in Figure 1(a), we can write a series of conditional independence relationships that are asserted from the Field. For example:

$$\begin{gathered} x_1 \perp x_3 \mid x_2, x_5 \\ x_1 \perp x_7 \mid x_3, x_6, x_{10} \\ x_9 \perp x_{12} \mid x_2, x_7, x_{10} \\ \vdots \end{gathered}$$

An MRF itself is a picture explaining these assumptions, which specify information about the underlying probability distribution. Specifically, it represents, for any given set of observed values, which nodes are conditionally independent. This does not, however, mean that these are the *only* conditional independencies. The fact that there is a path does not necessarily mean that there is a dependence. In fact, a distribution where each variable is fully independent can be represented with any set of edges, since any conditional independencies read from the MRF are definitely true. In general, the *fewer* edged an MRF has, the more information it is providing, since a sparser graph will have blocked paths with fewer observed variables. This means that in a sparser graph, there are more conditional independencies specified.

One important thing to understand is the relationship between the set of possible probability distributions and the set of possible MRFs. Since a distribution of all independent variables satisfies any set of conditional independencies, we know that the independent distribution:

$$P(\mathbf{x}) = \prod_i P(x_i).$$

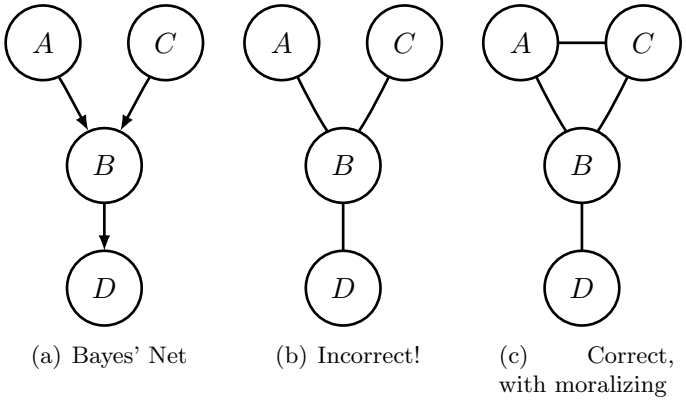
represents every possible MRF. This also means that there can be more than one distribution representing a given MRF.

In the other direction, we can also see that there are many possible MRFs for a given distribution. For any distribution, we can draw a fully connected MRF, but this is completely uninformative. In general, it may be incredibly difficult to determine the most useful MRF given a distribution, but we can typically draw the MRF using assumptions, for example "each node is affected only by its neighbors"

2.2 Moralizing: Converting a Bayes' Net to a MRF

Consider the Bayes' Net in Figure 2(a). Simply removing the arrows (Figure 2(b)) to create a MRF is not sufficient! In particular, in the resulting MRF, observing node B causes nodes A and C to become independent. This is the opposite of what the original Bayes' Net represented!

Instead, we need to *moralize* the graph. Whenever there are two parents that are not connected (married), we connect them. Thus, Figure 2(c) shows the correct representation of the original Bayes net. Note that during this conversion we actually lose information, namely that A and C are marginally independent.



(a) Bayes' Net (b) Incorrect! (c) Correct, with moralizing

Figure 2: Moralizing process.

3 Gibbs Fields

A Gibbs Field is another undirected graphical representation of a set of random variables and their relationships that was developed separately from MRFs. A simple example of a Gibbs Field is given in Figure 3; edges are undirected, and connote some correlation between the connected nodes. Gibbs Fields are powerful because they imply a way to write the joint probability of the random variables as functions over cliques in the graph.

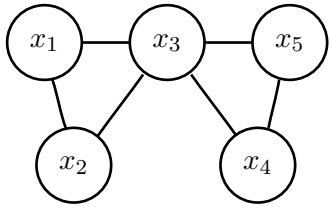


Figure 3: A Gibbs Field with nodes x_1, x_2, x_3, x_4, x_5.

Because the interaction between random variables is not directed, there is no reason to use conditional probability distributions were we represent one node given others. Instead we want to capture some more general notion of affinity between variables.

3.1 Cliques and Joint Probability

A clique is any fully connected subset of the the graph (e.g. $\{x_4\}$, $\{x_1, x_2, x_3\}$, or $\{x_3, x_5\}$). We denote the set of all cliques in a graph as C, with a clique $c_i \in C$ comprising its nodes (e.g. $c_3 = \{x_3, x_5\}$). The joint probability for any set of random variables $\mathbf{x} = \{x_1, \ldots x_n\}$ represented

by a Gibbs Field can be written as the product of clique potentials ϕ_i:

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c_i \in C} \phi_i(c_i), \tag{1}$$

with $\phi_i(c_i)$ the ith clique potential, a function only of the values of the clique members in c_i. Each potential function ϕ_i must be positive, but unlike probability distribution functions, they need not sum to 1. A normalization constant Z is required in to create a valid probability distribution $Z = \sum_x \prod_{c \in C} \phi_i(c_i)$.

For any Gibbs Field, there is a subset $\hat{C}$ of C consisting of only *maximal cliques* which are not proper subsets of any other clique. For example, the Gibbs Field in Figure 3 has two maximal cliques: $\hat{c}_1 = \{x_1, x_2, x_3\}$ and $\hat{c}_2 = \{x_3, x_4, x_5\}$. We can write a clique potential $\hat{\phi}$ for each maximal clique that is the product of all the potentials of its sub-cliques. In this way, we can write the joint probability as only a product over these maximal clique potentials:

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c_i \in \hat{C}} \hat{\phi}_i(c_i). \tag{2}$$

We usually take these potentials to be only functions over the maximal cliques, as in (2).

3.2 Clique Potentials as Energy Functions

Often, clique potentials of the form $\phi_i(c_i) = \exp(-f(c_i))$ are used, with $f_i(c_i)$ an energy function over values of c_i. The energy assigned by the function $f_i(c_i)$ is an indicator of the likelihood of the corresponding relationships within the clique, with a higher energy configuration having lower probability and vice-versa. If this is the case, (1) can be written as

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left[- \sum_{c_i \in C} f_i(c_i) \right]. \tag{3}$$

For example, we can write energy functions over the cliques in the example graph from Figure 3. Let $f_1(\{x_1, x_2, x_3\}) = x_1^2 + (x_2 - 5x_3 - 3)^2$, and $f_2(\{x_3, x_4, x_5\}) = (x_3 - x_4)^2 + (x_3 + x_4 + x_5)^2$. Then the joint probability can be written as

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left[-(x_1^2 + (x_2 - 5x_3 - 3)^2) - ((x_3 - x_4)^2 + (x_3 + x_4 + x_5)^2) \right].$$

In this form (f_i quadratic in the xs), the Gibbs Field is known as a Gaussian Gibbs Field.

3.3 Relationship to MRFs

MRFs and Gibbs Fields both "look" the same in the sense that they are undirected graphs. Gibbs fields have an implicit probability function ϕ for each clique, while MRFs only specify the conditional independence. A natural question to ask is weather we can use the same graph that represents an MRF and come up with a set of ϕ functions which represent the same distributions.

The *Hammersley-Clifford theorem* (also called the *Fundamental theorem of random fields*) proves that a Markov Random Field and Gibbs Field are equivalent with regard to the same graph.[1] In other words:

- Given any Markov Random Field, all joint probability distributions that satisfy the conditional independencies can be written as clique potentials over the maximal cliques of the corresponding Gibbs Field.
- Given any Gibbs Field, all of its joint probability distributions satisfy the conditional independence relationships specified by the corresponding Markov Random Field.

4 Factor Graphs

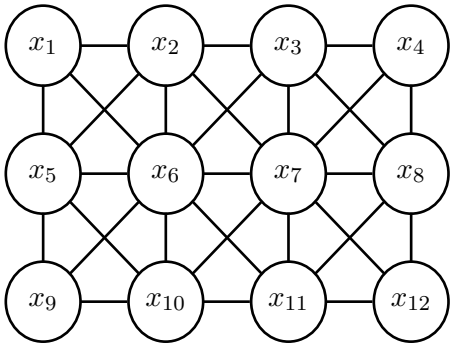


Figure 4: Diagonal connections are ambiguous, does the whole block of 4 nodes need to be considered together, or is this distribution still defined pair-wise between neighbors?

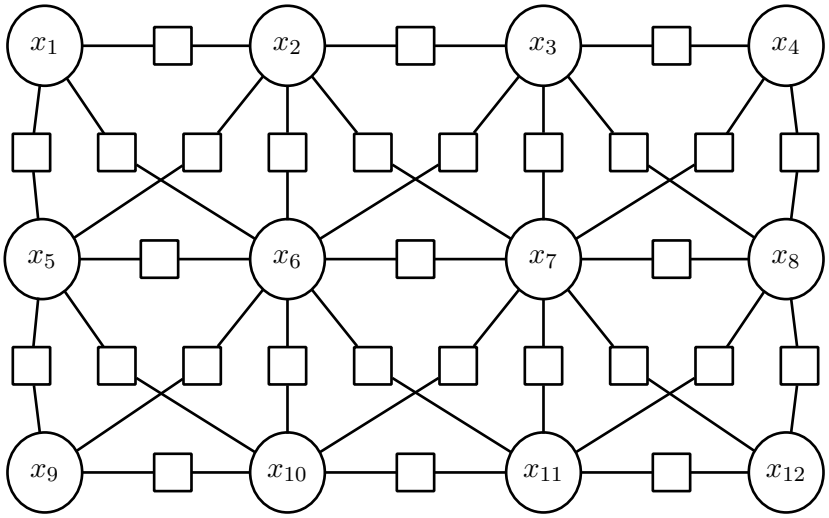


Figure 5: In the factor graph, the square nodes clearly represent that only neighboring nodes are dependent on each other and the squares to not need to all be considered simultaneously

[1] Actually, this is true only as long as $P(\mathbf{x}) \geq 0 \; \forall \mathbf{x}$; that is, as long as all configurations of values are possible.

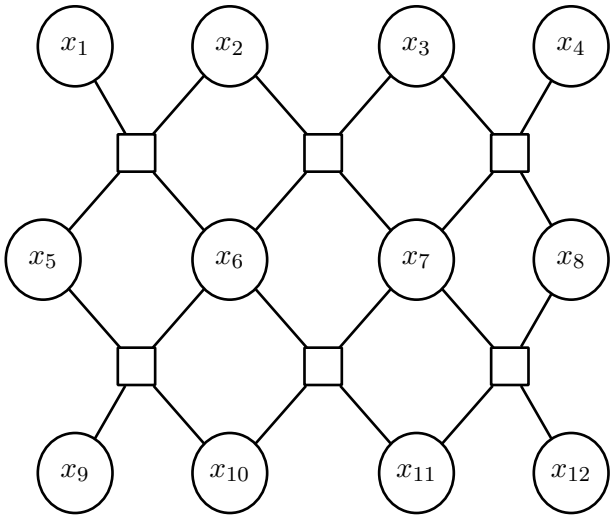


Figure 6: In this graph, each square of four nodes must indeed be considered simultaneously.

One problem with MRFs and Gibbs fields is that they don't have the granularity we might need. Consider, for example, the Gibbs field in Figure 4. From this picture, we would have to define ϕ over maximum cliques, which in this case are the squares. When this was a 4-connected grid, we only had to define functions over pairs. In order to make it clear that functions need only be defined over each pair in Figure 4, we define a factor graph. A *factor graph* has two types of nodes: random variables (same as the nodes in MRFs or Gibbs fields) and factors, which are square nodes and represent that all connected random variable nodes are dependent on each-other (as if they had a line in a traditional Gibbs field). This allows us to specify directly which vertices the ϕ functions need be defined on. In Figure 5 we can see that the factor graph representation of Figure 4 removes the ambiguity and shows us that, indeed, only neighboring pairs should be connected. In Figure 6 we see an alternate factor graph that shows us that the each node in each square is indeed dependent on each other node. The Factor Graphs in both Figures 5 and 6 are both consistent with the Gibbs field in Figure 4.