Markov Random Field

HY Eric Kim

University of Seoul, Department of Statistics

2019-02-14

Outline

- Markov Random Field
  - Conditional Independence Properties
  - Factorization
  - Example: Image De-noising
  - Relation to Directed Graph

Markov Random Field

Markov Random Field

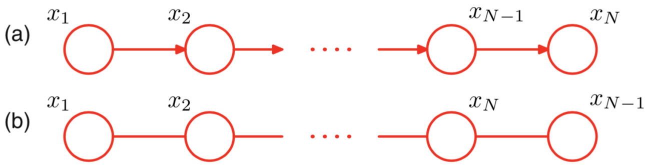

Figure: (a): an example of a directed graph. (b): The equivalent undirected graph

- Directed Graph specify a factorizing of the joint distribution over a set of variables into a product of local conditional distributions.
  - For example, a joint distribution for fig.(a) will be

$$p(X) = p(x_1)p(x_2 \mid 1) \cdots p(x_N \mid X_{N-1})$$

Markov Random Field

- Markov Random Fields a.k.a. "Marcov network" or "undirected graph" has a set of nodes each corresponding to a variable or group of variables as well as links between nodes.
- The links do not carry arrows. "No direction"

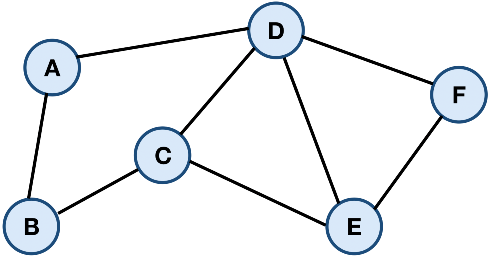

Figure: An undirected graph

Conditional Independence Properties

Conditional Independence Properties

- In directed graph, it is possible to test the independency.
  - By testing the path connecting two nodes are blocked or not.
- Sometimes it is subtle, since there exists "head-to-head" nodes.
  - Thats where the alternative arises.
  - Conditional independence in undirected graph is determined by a single graph seperation.
- By removing directionalities in links,
  THERE ARE NO PARENT, CHILD hence NO HEAD-TO-HEAD nodes.

Conditional Independence Properties

- Suppose in undirected graph, We can identify three sets of nodes A, B, C.
- Consider $A \perp\!\!\!\perp B \mid C$
  - If all possible paths from A to B pass through one or more nodes in C, it is **blocked**.
  - Simply, remove all nodes in C and related links. If two sets of nodes are disconnected, it is conditional independence.

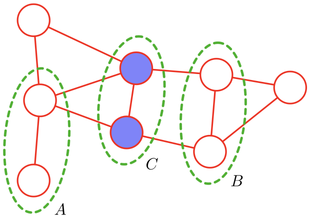

Figure: $A \perp\!\!\!\perp B \mid C$ is described by this graph.

Factorization Properties

Factorization Properties

- Now, let's express the joint distribution $p(X)$ as a product of functions defined over variables that are local to the graph.
- Before we do that, let's define **locality**.

Factorization Properties: Locality

- **Locality**: suppose there are x_i, x_j which are not directly connected. There must be conditional Independence given all other nodes in graph.
  1. This implies there is no direct link between the two nodes.
  2. All other paths goes through nodes that are observed.

  $\rightarrow$ That is to say: Those paths are blocked.

$$p(x_i, x_j \mid X_{\setminus\{i,j\}}) = p(x_i \mid X_{\setminus\{i,j\}})p(x_j \mid X_{\setminus\{i,j\}})$$

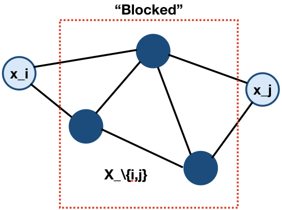

Figure: The concept of locality

Factorization Properties: Clique

Definition: Clique

A subset of nodes in a graph such that there exists a link between all pairs of nodes in the subset. Thus, all nodes in a clique is fully connected

Definition: Maximal Clique

A clique which is not possible to include any other nodes from the graph in the set.

- Therefore we can define the factors in the decomposition of the joint distribution to be **functions of the variables in a clique.**

Factorization Properties: Clique

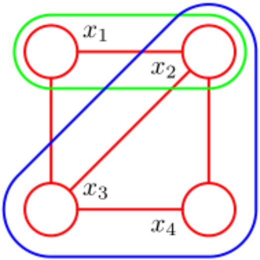

Figure: A four-node undirected graph shoing a clique

- In fact, we can cosider the functions of maximal cliques since all cliques are subset of maximal clique.
  - We can consider an arbitrary function over $\{x_1, x_2, x_3\}$.
  - Including another factor defined over a a subset d these variables would be redundant.

Factorization Properties: The Potential & Partition Function

- Let's denote a clique by C, a set of variables in C by X_C.
- The joint distribution is written as a product of *potential functions* $\psi_C(X_C)$ over the maximal cliques of the graph

$$p(X) = \frac{1}{Z} \prod_C \psi_C(X_C). \tag{1}$$

- The quantity Z is called *partition function* which is a normalization constant and is given by

$$Z = \sum_X \prod_C \psi_C(X_C). \tag{2}$$

Factorization Properties: The Partition Function

- The *partition function* is the limit of undirected graph.
  - For example, M discrete nodes each having K states, then the evaluation of the normalization term involves summing over K^M states.
- The partition function is needed for parameter learning.
  - Since it will be a function of any parameters that govern the potential functions.

Factorization Properties: The Partition Function

- For evaluation of local conditional distributions, the partition function is not needed.
  - Since a conditional is the ratio of two marginals.
  - $\rightarrow$ The partition function cancels between numerator and denominator when evaluating this ratio.
- Similarly, For evaluating local marginal probabilities, we can work with the unnormalized joint distribution and then normalize the marginals explicitly at the end.
- Provided the marginals only involves a small number of variables, the evaluation of their normalization coefficient will be feasible.

Factorization Properties

- Let's consider a graphical model as a filter.

  Consider the set of all possible distributions defined over a fixed set of variables corresponding to the nodes of a particular undirected graph.

- *UI*: the set of such distributions that are consistent with the set of conditional independence statements that can be read from the graph using graph separation.
- *UF*: the set of such distributions that can be expressed as a factorization of the form (1) with respect to the maximal cliques of the graph.
- The Hammersley-Clifford theorem (Clifford, 1990) states that **the sets UI and UF are identical**.

Factorization Properties: The Potential Function

- Let's discuss connection between conditional independence and factorization for undirected graph.

RESTRICT the potential functions $\psi_C(X_C)$ to be strictly positive.

- it is convenient to express $\psi_C(X_C)$ as exponentials, so that

$$\psi_C(X_C) = exp[-E(X_C)] \tag{3}$$

- $E(X_C)$ is called an *energy function*.
- The joint distribution is defined as the product of potentials.
  - The total energy is obtained by adding the energies of each of the maximal cliques.

Factorization Properties: The Potential Function

- The potentials in an undirected graph do not have a specific probabilistic interpretation.
  - Which is contrast to directed graph.
- It gives **greater flexibility** in choosing the potential function.
- Then how do we choose potential function for a particular application..?
  - See potential function as expressing which configurations of the local variables are preferred to others.
  - Global configurations that have a relatively high probability are those that find a good balance in satisfying the influences of the clique potentials.

Illustration: Image De-noising

- Let the observed noisy image be an array of binary pixel values $y_i \in \{-1, +1\}$, for $i = 1, ..., D$.
- Suppose the image is obtained by taking an unknown noise-free image $x_i \in \{-1, +1\}$.
- When the image is observed, it flips x_i's sign with probability 10%.

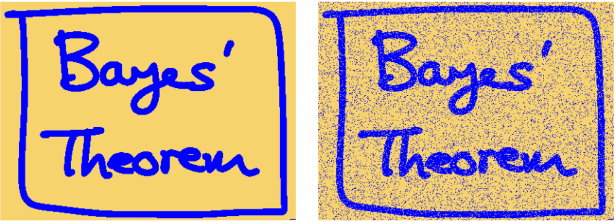

Figure: The original image on the left, the corrupted image on the right

Illustration: Image De-noising

- The noise level is small, so there will be a strong correlation between x_i and y_i.
- Neighbouring pixels x_i and x_j are strongly correlated.
- The graph expressing these relations has **two types of cliques**.
  - Each of which contains two variables.

Illustration: Image De-noising

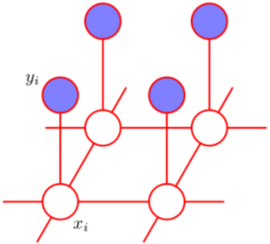

Figure: An undirected graph for image de-noising

- The cliques of the form $\{x_i, y_i\}$ have an associated energy function that expresses the correlation between these variables.
  - Simple energy function for these cliques: $-\eta x_i y_i$, $\eta > 0$
- The other cliques comprise pairs of variables $\{x_i, x_j\}$ consisting of neighbouring pixels.
  - An energy function: $-\beta x_i x_j$, $\beta > 0$.

Illustration: Image De-noising

- Recall that a potential function is **an arbitrary, nonnegative function** over a maximal clique.
  - We can multiply it by any nonnegative functions of subsets of the clique.
  - Equivalently we can add the corresponding energies.
- In this example, add an extra term hx_i for each pixel i.
- The complete energy function for the model then takes the form:

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

which defines a joint distribution over x and y given by

$$p(X,Y) = \frac{1}{Z} exp[-E(X,Y)]$$

Illustration: Image De-noising

- Fix the elements of y to the observed values.
  - This implicitly defines a conditional distribution $p(x \mid y)$ over noise-free images.
- To find an image x having a high probability, adopt a simple iterative technique called *iterated conditional modes* (ICM).
  1. Initialize the variables $\{x_i\}$, by simply setting $x_i = y_i$ for all i.
  2. Take one node x_j at a time, evaluate the total energy for $x_j = +1$ and $x_j = -1$
  3. Set x_j to whichever state has the lower energy.
  4. Repeat the update for another site, and so on, until some stopping criterion is satisfied.

Illustration: Image De-noising

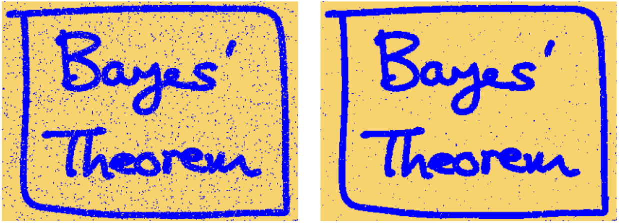

Figure: The restored image using (ICM) on the left and the graph-cut algorithm on the right

- In this example, the author has set $\beta = 1, \eta = 2.1$ and $h = 0$
  - $h = 0$ means that the prior probabilities of the two states of x_i are equal.
  - If we set $\beta = 0$, which removes the links between neighbouring pixels.
- There are many other more effective algorihms.
  - Such as *max-product algorithm*.

Relation to Directed Graph

Relation to Directed Graph: Converting Directed Graph

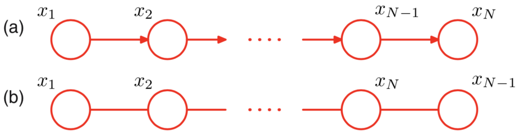

Figure: (a): an example of a directed graph. (b): The equivalent undirected graph

- Consider converting directed graph to undirected graph.
- Let's Convert (a) to (b). (See the above figure)
  - The joint distribution for (a):

$$p(X) = p(x_1)p(x_2 \mid 1) \cdots p(x_N \mid X_{N-1})$$

  - Converting (a) into undirected graph:

$$p(X) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3) \cdots \psi_{N-1,N}(x_{N-1}, x_N)$$

Relation to Directed Graph: Converting Directed Graph

$$p(X) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

- This is easily done by identifying:

$$\begin{aligned}
\psi_{1,2}(x_1, x_2) &= p(x_1)p(x_2 \mid x_1) \\
\psi_{2,3}(x_2, x_3) &= p(x_3 \mid x_2) \\
&\vdots \\
\psi_{N-1,N}(x_{N-1}, x_N) &= p(x_N \mid x_{N-1})
\end{aligned}$$

  - We observed the marginal $p(x_1)$ for the first node into the first potential function.
  - The partition function $Z = 1$

Relation to Directed Graph: Converting Directed Graph

- To generalize this construction, **The clique potentials are given by the conditional distributions** of the directed graph.
- For this to be valid, we must ensure the following:

  the set of variables that appears in each of the conditional distributions is a member of at least one clique of the undirected graph.
- For nodes having just one parent:

  Simply **drop the directionality**.
- For nodes having more than one parent, this is not sufficient:

  These nodes are called "head-to-head" node.

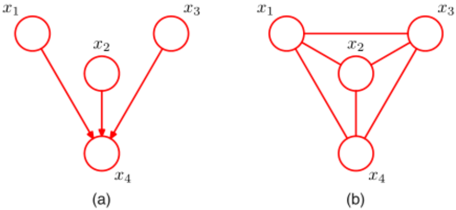

Figure: (a):a simple directed graph, (b):the corresponding moral graph

- The joint distribution for (a) is:

$$p(X) = p(x_1)p(x_2)p(x_3)p(x_4 \mid x_1, x_2, x_3)$$

- The 4th factor involves 4 variables.
  - To absorb this conditional distribution to a clique potential, it must belong to a single clique.

Relation to Directed Graph: Converting Directed Graph

- To do this, we add extra links between all pairs of parents of the node x_4
  - This process of 'marrying the parents' is known as ***moralization***.
  - The resulting graph is called ***moral graph***.
- In summary, converting directed graph to undirected graph is done by following:
  1. Add undirected links between all pairs of parents for each node.
  2. Drop the arrows on the original links.
  3. Take each conditional distribution factor and multiply it into one of the clique potentials.
  4. The partition function is given by $Z = 1$

Relation to Directed Graph: Converting Directed Graph

- Converting from an undirected to a directed representation is much less common.
  - In general, it presents problems due to the normalization constraints.
- In the process of converting, we had to discard some conditional independence properties.
  - Making fully connected undirected graph seems simple, but it discards all conditional independence properties.
- Moralization adds the fewest extra links and retains the maximum number of independence properties.

Relation to Directed Graph: Mapping Theory

- In determining the conditional independence properties, there are two types of graph can express different conditional independence properties.
- The **D-Map** and the **I-Map**

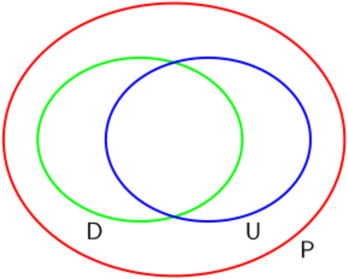

Figure: Venn diagram illustrating the set of all distributions

Relation to Directed Graph: Mapping Theory

- Let V be a set of variables and let $\perp\!\!\!\perp$ be an independence relation defined on V.
- Let $G = (V(G), E(G))$ be an undirected graph, then for each $X, Y, Z \subseteq V$:
  - G is called an undirected **dependence map(D-map)**, if:

$$X \perp\!\!\!\perp Y \mid Z \Rightarrow X \perp\!\!\!\perp_G Y \mid Z$$

  - G is called an undirected **independence map(I-map)**, if:

$$X \perp\!\!\!\perp_G Y \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$$

  - G is an undirected **perfect map(P-map)**, if G is both a D-map and an I-map, or, equivalently:

$$X \perp\!\!\!\perp_G Y \mid Z \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$$

Relation to Directed Graph: Mapping Theory

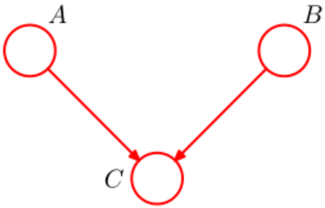

Figure: A directed graph is perfect map for a distribution

- A directed graph whose conditional independence properties cannot be expressed using an undirected graph over the same three variables.
- The graph is a perfect map for a distribution satisfying the conditional independence properties:

  $A \perp\!\!\!\perp B \mid \emptyset$ and $A \not\!\perp\!\!\!\perp B \mid C$.
- There is no corresponding undirected graph that is a perfect map.

Relation to Directed Graph: Mapping Theory

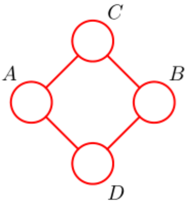

Figure: An undirected graph is perfect map for a distribution

- An undirected graph whose conditional independence properties cannot be expressed in terms of a directed graph over the same variables.
- The graph is a perfect map for a distribution satisfying the conditional independence properties:

  $A \not\perp\!\!\!\perp B \mid \emptyset$, $C \perp\!\!\!\perp D \mid A \cup B$ and $A \perp\!\!\!\perp B \mid C \cup D$.
- There is no corresponding directed graph that is a perfect map.

Summary

- An undirected graph can handle "head-to-head" nodes.
  - By removing directionalities in links from a directed graph.
- The conditional independence is quite different from directed graph
- Undirected graphs utilize the concept of clique.
  - Clique: a subset of nodes in a graph s.t. all nodes are fully connected.
- With clique, we can define potential and the partition function.
  - Just remember $p(x)$ is an arbitrary function that is strictly positve and Z is the normalization constant.
- Using $p(x)$ and Z, the joint distribution is defined as the product of potentials.
- It is possible to convert a directed graph to an undirected graph.
  - This process is called moralization and the resulting graph a moral graph.