10-708: Probabilistic Graphical Models, Spring 2020

2: Representation of undirected graphical models

Lecturer: Eric P. Xing *Scribe: T. Wörtwein, H. Lee, V. Rawal, M. Ferdosi*

Building a conditional independence graph (CIG) based on the dependencies of every possible pair of random variables quickly becomes infeasible. Therefore, today we will try something (potentially) easier than building a graph from the bottom up based on observations. Given some graph, how can we understand the model that is represented by it? How can we read the conditional independencies from the graph? What is the class of distributions represented by CIG?

1 Undirected graphical models (UGM)

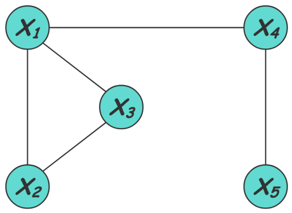

Figure 1: Example of UGM

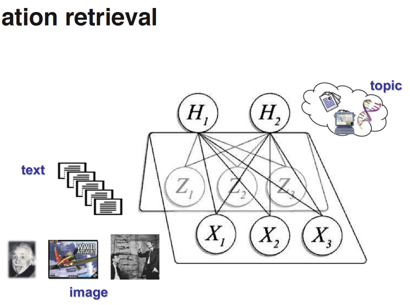

Figure 2: Example Use of UGM in Information Retrieval

Nodes correspond to random variables, while edges correspond to pairwise (non-causal) relationships. Undirected graphical models are $P(\boldsymbol{X}, \theta_G)$, i.e. probability distributions over random variables $\boldsymbol{X}$, whose parameters are determined by the graph G.

Computer vision example (is this blue patch air or water?): Build a grid model of patches in an image, and put a probability distribution on top, based on the assumption that connected nodes are likely to be of the same value.

More examples of UGM: physics model, social networks, protein interaction networks, modeling Go, ...

In domains such as information retrieval, such models can be used to describe relationships between concepts and also entities. For example, in the above figure, this graph models the correlations or alignments between images and texts and how they're aligned towards the hidden topics which give rise to certain words or certain image configurations in a dataset. This leads us to another type of graphical model.

2 Representation of undirected graphical model

An **undirected graphical model** represents a distribution $P(X_1, X_2, ..., X_n)$ defined by an undirected graph H, and a set of positive **potential functions** Ψ_c associated with the cliques of H, such that :

$$P(x_1, x_1, x_2, ..., x_n) = \frac{1}{Z}\prod_{c\in C}\Psi_c(X_c)$$

where Z is known as the partition function:

$$Z = \sum_{x_1,x_2,...,x_n}\prod_{c\in C}\Psi_c(X_c)$$

Given a graph, identify all the ”Cliques” present within the graphical model. This is also known as **Markov Random Fields, Markov Networks, ...** The **potential function** can be understood as a contingency function of its arguments assigning ”pre-probabilistic” score of their joint configurations. They need to be positive to avoid producing a negative probability upon multiplication. Normalizing constant Z is used to convert it into a probability distribution. It is the sum of potential values over all possible Random Variable configurations.

2.1 Quantitative Specification : Cliques

For G = {V,E}, a complete subgraph (clique) is a subgraph $G' = \{V' \subseteq V, E' \subseteq E\}$ such that nodes in V' are fully connected. A maximal Clique is a complete subgraph such that any superset $V'' \supset V'$ is not a clique. A sub-clique is a not-necessarily-maximal clique.

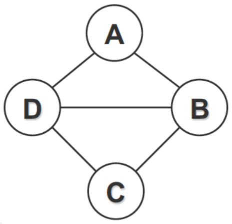

Figure 3: Example Graph

In this example, max-cliques are {A,B,C,D}, {B,C,D}, and sub-cliques are {A,B}, {C,D}, ... all edges and singletons.
Reason for using cliques : Cliques are basic units that capture all possible dependencies that are possible and not to be missed. If we start building from the sub-graphs within the cliques, and begin inter-connecting them, we may risk losing modeling some inter-dependencies.

2.2 Gibbs Distributions

The following expression in terms of potential functions is also called Gibbs Distribution.

$$P(x_1, x_1, x_2, ..., x_n) = \frac{1}{Z}\prod_{c\in C}\Psi_c(X_c)$$

2.3 Interpretation of Clique Potentials

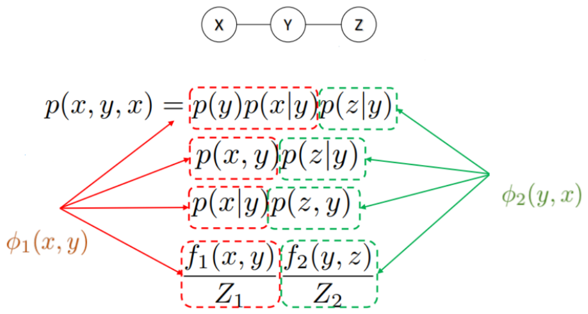
X
Y
Z
$p(x,y,x) = p(y)p(x|y)p(z|y)$
$p(x,y)p(z|y)$
$p(x|y)p(z,y)$
$\frac{f_1(x,y)}{Z_1}\frac{f_2(y,z)}{Z_2}$
$\phi_1(x,y)$
$\phi_2(y,x)$

Figure 4: Interpreting Clique Potentials

Clique potentials are pre-probabilistic, contingency functions offering ways to recover or specify biases on configurations over Random Variables.

In directed graphical models, the joint distribution over the vertices could be factorized as a product of marginal and conditional distributions. In undirected graphical models however, the joint distribution can be factorized into a product of clique potentials. However, these clique potentials are not necessarily marginals or conditionals. They only represent a notion of "goodness" or "compatibility" of the variables. To illustrate this, consider the graph shown in the Figure. While the graph implies $X \perp\!\!\!\perp Z | Y$ and hence, the joint distribution can be represented as $p(x,y,z) = p(y)p(x|y)p(z|y)$, it can also be written in other forms as shown in the figure.

2.4 Examples of Undirected Graphs and Gibbs Representations

The potential functions shown here are symbolic. In order for the clique function to be implemented in a programming language, one would have to use a 3D matrix representation given that we have 3 binary variables at a time. (when using max-cliques as below). So, for discrete nodes, one can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table.

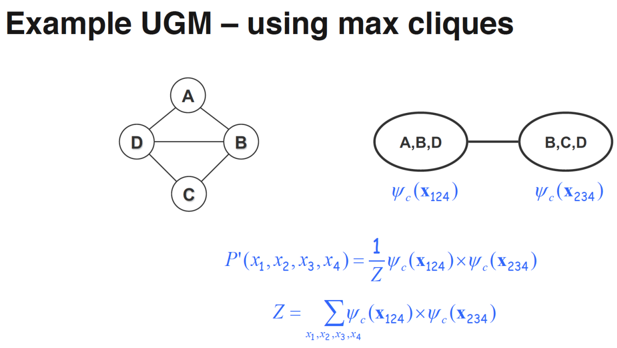
Example UGM – using max cliques
A
B
C
D
A,B,D
B,C,D
$\psi_c(\mathbf{x}_{124})$
$\psi_c(\mathbf{x}_{234})$
$P'(x_1,x_2,x_3,x_4) = \frac{1}{Z}\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$
$Z = \sum_{x_1,x_2,x_3,x_4} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$

Figure 5: Example UGM : Using max-cliques

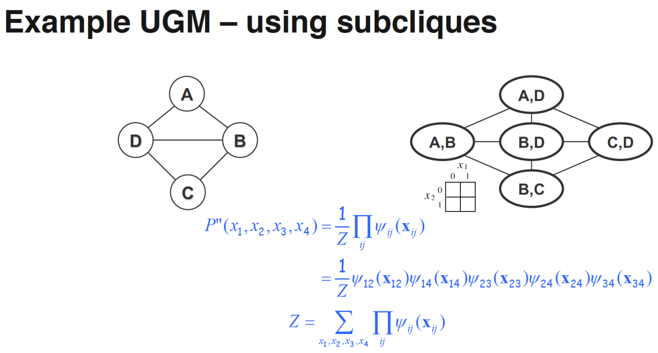
Example UGM – using subcliques
A
B
C
D
A,D
A,B
B,D
C,D
B,C
x_1
x_2
0
1
$P''(x_1,x_2,x_3,x_4) = \frac{1}{Z}\prod_{ij}\psi_{ij}(\mathbf{x}_{ij})$
$= \frac{1}{Z}\psi_{12}(\mathbf{x}_{12})\psi_{14}(\mathbf{x}_{14})\psi_{23}(\mathbf{x}_{23})\psi_{24}(\mathbf{x}_{24})\psi_{34}(\mathbf{x}_{34})$
$Z = \sum_{x_1,x_2,x_3,x_4}\prod_{ij}\psi_{ij}(\mathbf{x}_{ij})$

Figure 6: Example UGM : Using sub-cliques

Instead of using the max-cliques, one could also alternatively use the sub-cliques which gives us a different way of specification. However, one may lose something on doing so : it may certainly be possible for one to have some tricky configurations in the triplets representation which are not necessarily recoverable (or atleast easily recoverable) using the pairwise potential representation. So, in this way, one can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table such as using pair-MRFs, a popular and simple special case.

There is another form of canonical representation which allows one to be over-complete because it uses the potential functions from all possible cliques existent in the graph : triplet, pairwise, and also the singleton cliques. This offers us the ultimate flexibility but the model becomes a little more complicated. The canonical form is thus the most expensive and the richest form and it subsumes the previous two cases.

3 Independence Properties

Now we ask what kinds of distributions, in terms of the set of independence relationships between variables, can be represented by undirected graphs (ignoring the details of the particular parameterization). How do we define placeholders or family of graphs to represent a given set of independence assumptions?

3.1 Definition : Global Markov properties

Global Markov properties of an Undirected Graph H are

$I(H) = \{(X \perp\!\!\!\perp Z|Y) : sep(X; Z|Y)\}$

i.e. The sets X and Y of random variables are independent given the set Z if we have that the two sets are separated from each other by the set Z.

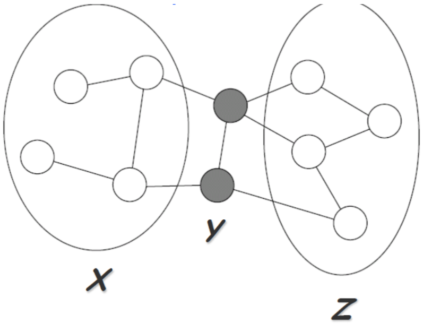

Figure 7: Global Markov Properties

4 I-maps

We define I-maps to establish a formal relationship between a graph H and a distribution P using the conditional independence definition. This will make it easier to describe all conditional independencies of P by representing them as a graph.

Definition: Let P be a distribution over X. We define I(P) as the set of all independence assertion of the form $X \perp\!\!\!\perp Z|Y$ that hold in P (independent of the parameter-values).

X	Y	P(X, Y)
1	1	0.64
1	0	0.16
0	1	0.04
0	0	0.16

Table 1: Are X and Y independent?

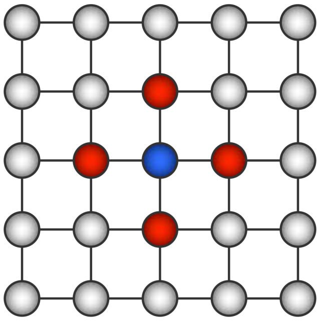

Figure 8: Nodes marked in red are the Markov blanket for the blue node.

Definition: Let H be a graph and $I(H)$ its set of independence assertions. $I(H)$ is an I-map for $I(P)$, if $I(H) \subseteq I(P)$.

If $I(H)$ is an I-map for $I(P)$, it cannot contain independence assertions which are not present in $I(P)$. The contingency Table 1 demonstrates why it is harder to determine conditional independence directly from P compared to from a graph.

4.1 Global Markov Independencies

A probability distribution P satisfies the global Markov property of the undirected graph H, if for any disjoint X, Y, and Z such that Y separates X and Z, X is independent of Z given Y:

$$I(H) = \{X \perp\!\!\!\perp Z | Y : \text{sep}_H(X; Z|Y)\}$$

$I(H)$ is a I-map of P.

4.2 Local Markov Independencies

Definition: The set of neighboring nodes of a node X_i in an undirected graph is called Markov blanked (denoted as MB_{X_i}).

Definition: The local Markov independencies associated with H are:

$$I_l(H) = \{X_i \perp\!\!\!\perp V - \{X_i\} - MB_{X_i} | MB_{X_i} : \forall i\}$$

Where B is the set of all nodes in H.

This means that X_i (the blue node in Figure 8) is independent of all other nodes in H (the white nodes in Figure 8) given its Markov blanket MB_{X_i} (the red nodes in Figure 8), i.e., all neighbors of X_i.

4.3 Soundness and Completeness of global Markov Independencies

P is a Gibbs distribution over H, if it can be represented as a normalized product over all potential functions defined on the cliques of H:

$$P(X) = \frac{1}{Z} \prod_{c \in C} \Psi_c(X_c)$$

Soundness (from graph to distribution): If P is a Gibbs distribution over H, then H is guaranteed to be an I-map of P. Given a graph H, we can find a distribution P such that H is an I-map of P, by letting H be a Gibbs distribution over H.

Completeness: If $\mathrm{sep}_H(X; Z|Y)$ not in $I(H)$ (implies that X and Y are dependent given Z), there are still some P that factorize over H in which X is independent of Z given Y ($X \perp_P Z|Y$). Intuitively, this can happen when the concrete probability numbers are numerically independent. Despite that, the reverse (from a distribution to a graph) is 'almost always' possible.

There is no strict equivalence between graphs and distributions!

Definition: The pairwise Markov independencies associated with UG $H = (V, E)$ are

$$I_p(H) = \{X \perp Y | V \setminus \{X, Y\} : \{X, Y\} \notin E\}$$

4.4 Hammersley-Clifford Theorem

Hammersley-Clifford Theorem: Let P be a positive distribution over V, and H a Markov network graph over V. If H is an I-map for P, then P is a Gibbs distribution over H.

Sometimes H and P are perfectly equivalent.

Definition: A Markov network H is a perfect map for P if for any X, Y, Z, we have that

$$\mathrm{sep}_H(X; Z|Y) \iff P(X \perp Z|Y)$$

Theorem: Not every distribution has a perfect map as UGM.

4.5 Exponential Form

Constraining clique potentials to be positive can be inconvenient (e.g. sometimes you want to model physical phenomenons with +1, -1). We represent a clique potential $\psi_c(X_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(X_c)$: $\psi_c(x_c) = \exp(-\phi_c(x_c))$. The joint probability has a nice additive structure

$$p(x) = \frac{1}{Z} \exp\left\{ -\sum_{c \in C} \phi_c(x_c) \right\} := \frac{1}{Z} \exp\{-H(x)\}$$

$H(x)$ is free energy. This is called the Bolzmann distribution in physics, and a log-linear model in statistics.

5 Boltzmann machines

Definition: A fully connected graph with parwise(edge) potentials on binary-valued nodes (-1,1).

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp\left\{\sum_{ij} \phi_{ij}(x_i, x_j)\right\} = \frac{1}{Z} \exp\left\{\sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C\right\}$$

The overall energy function has the form of $H(x) = \sum_{ij}(x_i - \mu)\theta_{ij}(x_j - \mu) = (x - \mu)^\top \Theta (x - \mu)$. Why is this useful : we will cover a technique to learn this model to recover the structure of the graph from data. Graph becomes interesting especially if Θ is sparse. If $\Theta_{ij} = 0$, then there is no edge between x_i and x_j.

6 Ising Models

The concept of the Ising model comes from statistical physics as a model for modeling the energy of a physical system consisting of interactions of atoms which can be represented as nodes in an undirected graph. In the Ising model, nodes are arranged in a regular topology (often as a grid) and connected only to their immediate neighbors. It can be seen as a sparse Boltzmann Machine. We can also define a multi-state Ising model (called Potts model), in which nodes can take multiple values instead of just binary values. One example of Ising model is shown in the following figure:

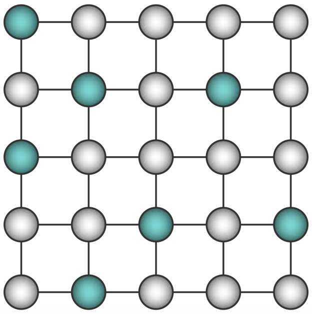

Figure 9: Example of an Ising model in which each node is connected to its neighbouring node

Its probability distribution can be written as:

$$P(X) = \frac{1}{Z} \exp\{\sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i\}$$

7 Restricted Boltzmann Machine (RBM)

The Restricted Boltzmann Machine (RBM) is inspired by the Boltzmann Machine. We can model it as a bipartite graph consisting of visible units and hidden units. One example is shown below:

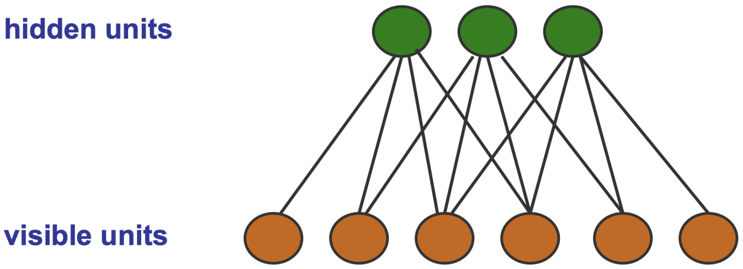

Figure 10: Example of RBM as an undirected graph wit visible and hidden nodes

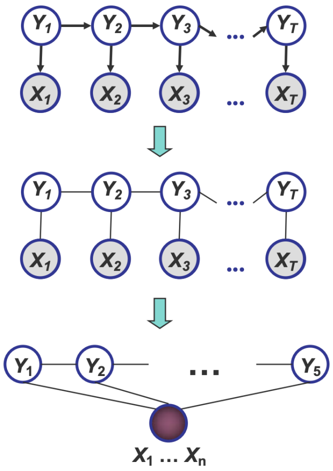

Figure 11: Example of CRF

Its probability distribution can be written as:

$$p(x,h|\theta) = \exp\Big\{\sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) \sum_{i,j} \theta_{i,j}\phi_{i,j}(x_i,h_j) - A(\boldsymbol{\theta})\Big\}$$

Some of the properties of RBM: Factors are marginally dependent. Factors are conditionally independent given observations on the visible nodes. It enables iterative Gibbs sampling. There are some learning algorithms for its estimators.

The conditional probability distribution of one layer given the other layer can be factorized:

$$p(H|X) = \prod_{i=1}^{n} p(h_i|X)\ ,\ p(X|H) = \prod_{i=1}^{m} p(x_i|H)$$

The marginal distribution of hidden and visible layers can be defined as:

$$p_{ind}((H)) \propto \prod_j \exp\{\theta_j g_j(h_j)\}\ ,\ p_{ind}((X)) \propto \prod_i \exp\{\theta_i g_j(x_i)\}$$

Using f and g, we can write the joint distribution in the following form:

$$p(x,h|\theta) = \exp\Big\{\sum_i \theta_i f_i(x_i) + \sum_j \lambda_j g_j(h_j) \sum_{i,j} f_i^t(x_i)\mathbf{W}_{i,j} g_j(h_j)\Big\}$$

8 Conditional Random Field (CRF)

CRFs are undirected graph representation in which we can encode conditional distributions $P(Y|X)$, where Y_i are target variables and X is the set of observed variables.

The probability distribution could be written as:

$$P_\theta(y|x) = \frac{1}{Z(\theta,x)\exp\{\sum_c \theta_c f_c(x,y_c)\}}$$