10-708: Probabilistic Graphical Models, Spring 2015

3: Representation of Undirected GM

Lecturer: Eric P. Xing *Scribes: Karima Ma, Manu Reddy*

1 Graphical Model Review

In the first couple of lectures, we talked about Bayesian Networks (also called Directed Graphical Models). In this class, we learnt about an alternate specification of GMs called Markov Random Fiels (also called Unidirected Graphical Models).

Directed GMs are structurally represented through a directed acyclic graph. Directed edges in a Directed GM model causal relationships. However, in an undirected GM, edges in the graph are undirected. The undirected edges generally model correlations between variables.

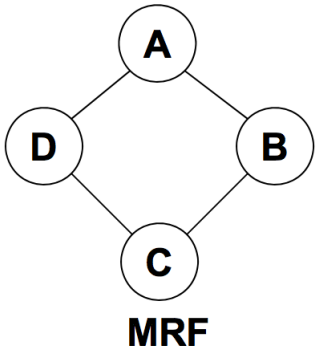
In a DAG we can represent the full joint distribution using the chain rule: $P(X_1, X_2, ...X_n) = \prod_{i=1}^{n} P(X_i|X_{\pi_i})$ where X_{π_i} denotes the parents of X_i.

In an undirected GM, we instead have various mappings (called potential functions) ψ for each clique in the graph. The joint is the normalized product of the potential functions $P(X_1, X_2, ...X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(X_c)$ where $\mathcal{C}$ specifies any set of cliques in the graph.

2 I maps

$I(G)$ is defined as the set of local indecencies encoded by a DAG G. A DAG G is an Independence-map of the distribution P, if $I(G) \subseteq I(P)$. Therefore, a fully connected graph is an I-map for any distribution because there are no conditional independencies encoded in the graph and the null set is a subset of all possible $I(P)$.

A DAG G is a perfect-map (P-map) for a distribution P if $I(G) = I(P)$. A DAG G is a minimal I-map for P if it is an I-map for P and the removal of any single edge in G would make it no longer an I-map of P. A distribution can have several minimal I-maps. However, not every distribution has a perfect map as a DAG. We can use proof by contradiction to show this. Say we have a distribution over four random variables A, B, C, and D, where $A \perp\!\!\!\perp C|\{B, D\}$ and $B \perp\!\!\!\perp D|\{A, C\}$. By drawing out all possible DAGs using these four random variables, we can see that none of them can represent both of these conditional independencies. However, the undirected graph below can:

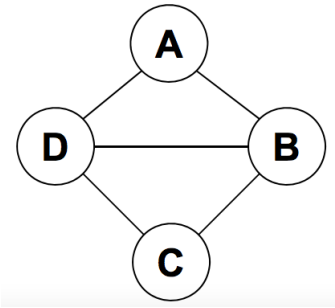


A graph may have multiple P-maps, but they are all I-equivalent, meaning they all specify the same exact set of independencies.

3 Undirected graphs

Undirected graphs are also known as Markov Random Fields or Markov networks. In an undirected graph, we map potential functions ψ to cliques of the graph. The joint is the normalized product of the potential functions. A clique of a graph G is defined as a complete subgraph of G i.e. a subgraph where all pairs of nodes have an edge between them. A clique C of a graph G is a complete subgraph of G. If we choose the cliques to be maximal, the representation becomes unique. A maximal clique of a graph is a clique C such that any superset of the nodes in C is not complete.

In the graph below, nodes A,B, and D belong to one max-clique and nodes B,C, and D belong to another max-clique.



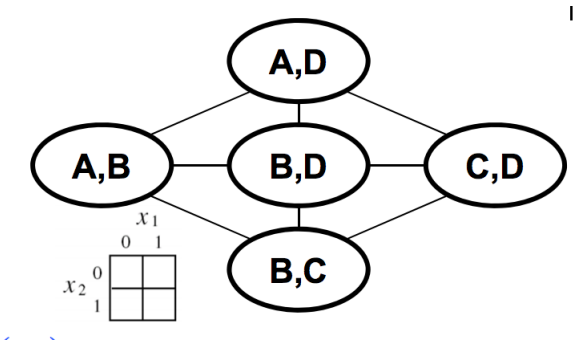
The joint distribution for this graph can be expressed as the normalized produce to potential functions:

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_c(x_{124}) * \psi_c(x_{234})$$

Note that this is only one of multiple possible representations. There are other possible representations as well. If each node in this graph represents a discrete random variable that can take on one of two values, then we can represent the full joint with two 3D tables, one for each potential function instead of one four dimensional table. Here it does not make much of a difference whether or not we choose to factor the joint as the product of two potential functions because the number of nodes is small. However, when the number

of nodes gets large, if we were to model the full joint with only one table, the table size would increase exponentially in the number of nodes. In these cases, it is extremely useful to be able to factor the joint into several potential functions, each with exponentially smaller tables.

We can also choose to factor this joint distribution even further, for example, with pairwise potential functions, one for every possible pair of nodes. The graph that represents this factorization is shown below:



This graph is I-equivalent to the graph above, i.e. the set of independencies encoded in these two graphs are the same. However, the graph above represents more dependencies than this graph because we can always factor the potential function over a set of variables such as $\psi(A, B, C)$ as $\psi(A) * \psi(B) * \psi(C)$ because the dependencies between A,B, and C can still be represented by making the separate potential functions correlated with each other.

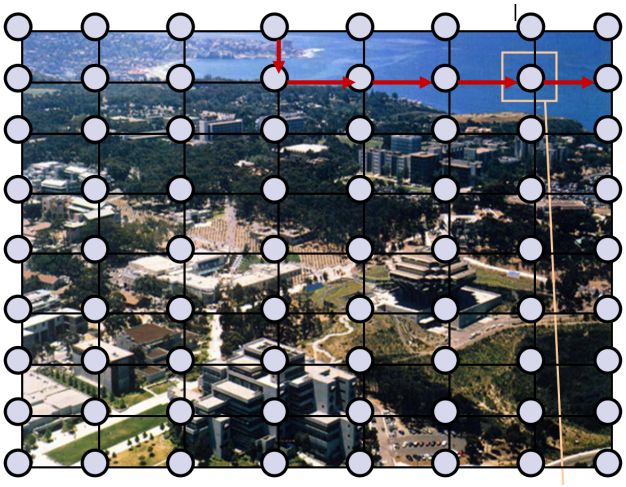
In undirected graphs, we cannot interpret potential functions as conditional or marginal distributions like we can in a DAG. Instead, we can interpret potential functions as scoring mechanisms that give certain configurations of nodes in the graph a higher score if they are more probable. Potential functions assign "pre-probabilistic" scores to the joint configuration of their arguments. Potential functions are only restricted by the fact that they must be strictly positive functions. Since, it is not guaranteed that the product of all the potential function for an undirected graph will equal to one. Thus we use the partition function to compute the normalizing factor with which we divide the product of the potential functions. The equation for the joint distribution of an undirected graph is:

$$P(x_1, ..., x_2) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

The distribution $P(x_1, ..., x_2)$ which factorizes as above is known as the Gibbs distribution. Here Z is the partition function:

$$Z = \sum_{x_1, ..., x_n} \prod_{c \in C} \psi_c(x_c)$$

In real world, undirected graphs are used in various data-mining problems in various fields such as images or social networks. For example, one could use an undirected graph structured as a grid to represent an image, where each node is a pixel. There are edges between neighbouring nodes because adjacent pixels tend to be correlated in terms of color, what object in the image they are a part of, etc. The undirected grid naturally arises in tasks such as image processing and lattice physics.



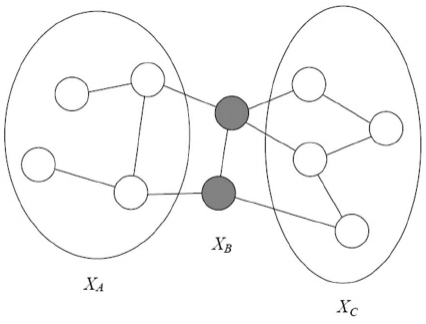
3.1 Independence in an Undirected Graph

In an undirected graph G, two sets of nodes X and Z are independent given another set of nodes Y, if removing all nodes in Y eliminates all paths from any node in X to any node in Y. This is known as the global Markov property.

The Markov blanket for any node X in G is the set of neighbors of X in the graph. Any node X is independent of all other nodes in the graph given its Markov blanket. This is known as local Markov property.

4 Independence properties

4.1 Global Markov Independencies



In an undirected Graph H, a set of nodes B separates sets of nodes A and C if there is no active path between A and C when B is observed. This is denoted as $sep_H(A; C|B)$

A probability distribution satisfies the global markov property if for any disjoint sets of nodes A,B,C such that B seperates A and C, A is independant of C given B. The global Markov independencies of graph H are written as

$$I(H) = A \perp C|B : sep_H(A; C|B)$$

4.1.1 Soundness and Completeness of Global Markov property

The above definition is sound and complete.
Soundness: If P is a Gibbs distribution over H then H is an I-map of P.

Completeness: If $\neg sep_H(X; Z|Y)$ then $\exists$ P that factorizes over H such that $X \not\perp_P Z|Y$.

4.2 Local Markov Independencies

The unique Markov blanket of a node X_i, denoted MB_{X_i}, is the set of its neighbours which share an edge with it. The local Markov independency is that a node is independent of every other node in the graph given its Markov blanket. The local Markov independencies associated with a graph H are

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

4.3 Pairwise Markov Independencies

The pairwise Markov independencies associated with an undirected graph $H = (V, E)$ are

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

4.3.1 Relationship between local and global Markov properties

Theorem: If $P \models I_l(H)$ then $P \models I_P(H)$
Theorem: If $P = I(H)$ then $P \models I_l(H)$
Theorem: If $P > 0$ *and* $P \models I_p(H)$ then $I(H)$
Corollary: The following three statements are equivalent for a *positive* distribution P:

$$P \models I_l(H), \text{ local Markov independencies}$$
$$P \models I_p(H), \text{ pairwise Markov independencies}$$
$$P \models I(H) \text{ , global Markov independencies}$$

The above equivalence relies on the positivity assumption and might not hold for non-positive distributions.

4.4 Hammersley-Clifford Theorem

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.

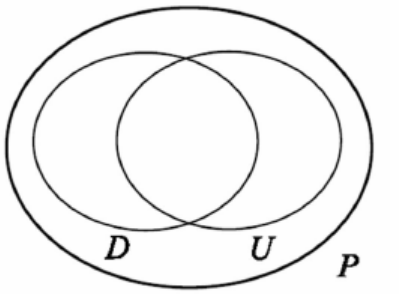
4.5 Perfect Maps

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

$$sep_H(X; Z|Y)P \models (X \perp Z|Y)$$

Theorem: Not every distribution has a perfect map as UGM. The proof is by counter example. The independencies in a v-structure $X \rightarrow Z \leftarrow Y$ cannot be captured by any MRF with the same number of nodes.

There are some distributions which could be captured with Bayesian networks alone and some which could be captured with MRF's alone. There are also some distributions which cannot be captured with either MRF's or BN's. The figure below summarizes the statement.



4.6 Exponential form

Constraining the clique potentials to be positive could be inconvenient as these could be used to model the interactions between a pair of atoms which could be repulsive or attractive. In order, to get around this problem, we represent a clique potential $\Psi_c(x_c)$ in an unconstrained form using real-value energy function $\phi_c(x_c)$.

$$\Psi_c(x_c) = exp\left\{\phi_c(x_c)\right\}$$

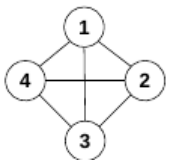
. This gives the joint probability distribution a nice additive structure.

$$p(x) = \frac{1}{Z} * exp\{-\sum \phi_c(x_c)\} = 1/Z * exp\{-H(x)\}$$

where H is the free energy. This model is called the Boltzmann distribution in physics. In statistics this is the log linear model. The reason is that if we take a logarithm we end up with a linear function.

5 Examples

5.1 Boltzmann Machines



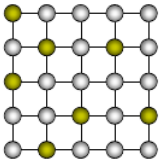
The boltzmann machine is an example of the undirected graphical model.This was originally used to capture dependencies/interactions between atoms. It is a fully connected graph with pairwise(edge) potentials on binary-valued nodes. Its joint probability distribution is given as.

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

Upon further simplificiation, the expression has the structure of a gaussian distribution.

$$H(X) = \sum_{ij} (x_i - \mu)^T \Theta (x_i - \mu)$$

5.2 Ising Model

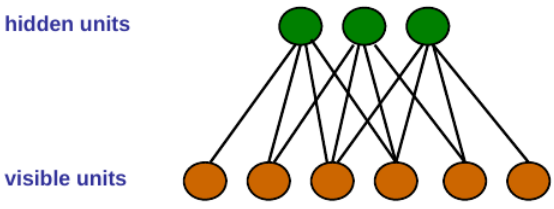


Ising models are another example of undirected graphical models. These are same as sparse Boltzmann machines. The nodes in this model are arranged in a regular topology like a lattice and are only connected to their geometric neighbours. Its potential is given by.

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

A **Potts model** has the same structure as an Ising model but the nodes are allowed to take mulitple values unlike the Ising model.

5.3 Restricted Boltzmann Machines



Another example of Undirected graphical models is Restricted Boltzmann machines(RBM). These are often used as the building blocks for the deep belief networks. An RBM is an assymetric model. The state space for every variable is not the same as in the Potts model and the Boltzmann machines. It consists of a hidden and visible layer. The nodes in the visible layer correspond to observations and the nodes in the hidden layer are features learnt from these observation. These features can in turn be used as input for a supervised learning model. There are edges between the layers but none between the nodes in the same layer. The joint probability distribution for the RBM is

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

where, $A(\theta)$ is the normalization term.

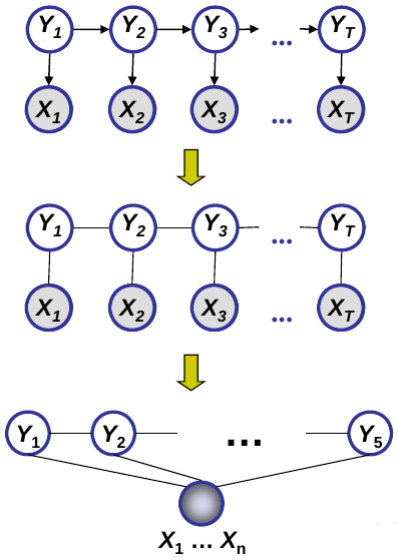
5.3.1 Properties of RBM

The latent variables are marginally dependent but they are conditionally independent given observations on the visible nodes and vice versa. This decoupling allows us to compute the latent variables given the observations in a single step. This also allows us to perform iterative Gibbs Sampling over the model.

$$p(l,w) = \prod_i p(l_i,w)$$

5.4 Conditional Random Fields

Discriminative models are often used for labeling tasks. Unlike generative models like HMMs, CRFs do not assume anything about the way the data or labels were generated. As a result, they use undirected edges. In a CRF, each data point, $\{X_1,...,X_n\}$ has a corresponding label Y_i, but unlike many other labeling models, general CRFs do not have to assume that individual data points are IID, so all of the $X_i s$ can be placed in one super node X. There are edges between a label Y_i and its previous and next label, Y_{i-1} and Y_{i+1} and an edge between each label and the super node X. Due to this structure, when labeling a data point X_i, future and past observations are taken into account. Below is the graphical representation of a CRF.



The goal of a CRF is to model the conditional probability of a label sequence Y given an observation sequence X. We can model this conditional probability, $p(y|x)$, by assigning a feature function f_k to each clique k, i.e. a triplet of two adjacent labels Y_k and Y_{k+1} and X, weighted by a parameter λ_k. We can also assign a function g to each label vector y_k weighted by a parameter μ_k. The conditional probability $p(y|x)$ will be equal to:

$p(y|x) = \frac{1}{Z} exp\{\sum_k \lambda_k f_k(y_k, y_{k+1}, x) + \sum_{v \in Y} \mu_v g_v(y_v)\}$

CRFs are popular for tasks like image analysis where we have less confidence on exactly how each data point affects its neighbors because they allow for arbitrary dependencies between data points. Potential functions provide us with a way to model correlation between variables without any generative story.

5.5 Summary

- An undirected graphical model captures relatedness, coupling and co-occurrence and synergism between data.
- We can identify local and global independence properties from the structure of the graph. In an undirected GM, a node is independent of every other node in the network given its neighbours.
- They could be used to define conditional or joint probability distributions.
- Potential functions and the cliques completely determine the joint probability distribution.
- Example of undirected graphical models.

  Boltzmann Machines

  Ising Model

  Potts Model

  Restricted Boltzmann Machines

  Conditional Random fields.