10-708: Probabilistic Graphical Models, Spring 2017

Lecture 3 : Representation of Undirected Graphical Model

Lecturer: Eric P. Xing *Scribes: Yifeng Tao, Devendra Singh Sachan, Jilong Yu*

1 Directed vs. Undirected Graphical Models

1.1 Two types of GMs

There are two types of graphical models: Directed Graphical Model (or Directed Acyclic Graphs- DAG) and Undirected Graphical Model (UGM). The directed edges in a DAG give **causality** relationships, DAGs are also called **Bayesian Network**. The undirected edges in UGM give **correlations** between variables, UGMs are also called Markov Random Fields (MRF). There are two types of ways to organize and represent relationships between variables. Here is an example of representing the regulatory relationships of proteins and genes in both ways.

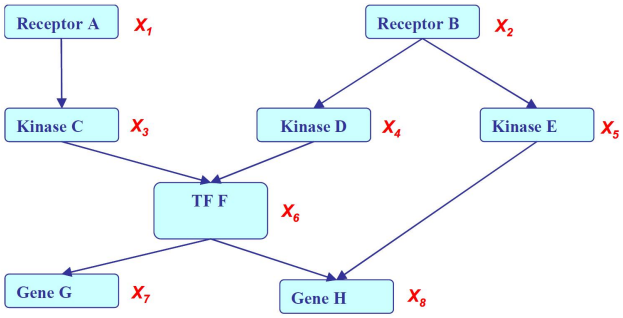

Figure 1: Example of DAG.

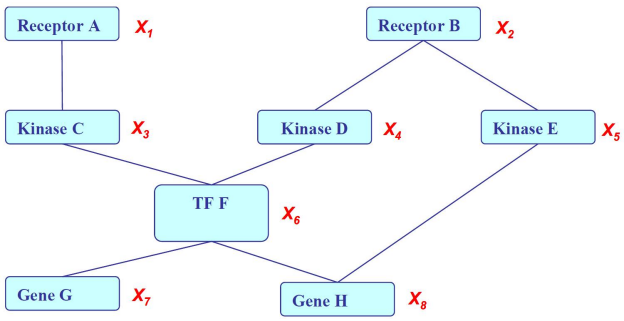

Figure 2: Example of UGM.

The joint probability of the DAG example can be represented as:

$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ =&P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2)P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6) \end{aligned} \tag{1}$$

The joint probability of the UGM example can be represented as:

$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ =&\mathbf{1}/\mathbf{Z}\exp\{E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2) \\ &+ E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\} \end{aligned} \tag{2}$$

Note that the normalization term $1/Z$ is required in the factorization in UGM.

1.2 Review: independence properties of DAGs

In previous lectures, we have seen the following definitions for $I_l(G)$ and I-map:

Definition 1 *let $I_l(G)$ be the set of local independence properties encoded by DAG G, namely,*

$$I(G) = \{X \perp Z|Y : dsep_G(X; Z|Y)\}. \tag{3}$$

Definition 2 *A DAG G is an I-map (independence-map) of P if $I_l(G) \subseteq I(P)$.*

According to the definition, a fully connected DAG G is an I-map for any distribution, since $I_l(G) = \emptyset \subseteq I(P)$ for any P.

We also have the following definition of minimal I-map:

Definition 3 *A DAG G is a minimal I-map for P if it is an I-map for P, and if the removal of even a single edge from G renders it not an I-map.*

A distribution map have several minimal I-maps, with each corresponding to a specific node-ordering.

1.3 P-maps

Based on the definition above, we give the definition of P-maps and its property:

Definition 4 *A DAG G is a **perfect map** (P-map) for a distribution P if $I(P) = I(G)$.*

Theorem 5 *Not every distribution has a perfect map as DAG.*

This can be easily proved by counterexample:

Proof: Suppose we have a model where

$$A \perp C|\{B, D\}, B \perp D|\{A, C\}. \tag{4}$$

This cannot be represented by any Bayes network. ■

Figure 3 list the two possibilities of DAG that satisfy $A \perp C|\{B, D\}$, however, they do not satisfy $B \perp D|\{A, C\}$: BN1 wrongly says $B \perp D|A$ and BN2 wrongly says that $B \perp D$.

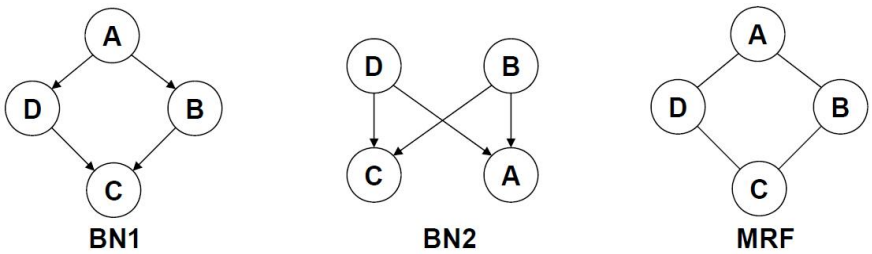

Figure 3: Counterexamples of DAGs.

In addition, the fact that G is a minimal I-map for P is far from a guarante that G captures the independence structure in P.

We also have the property of P-maps: The P-map of a distribution is unique up to I-equivalence between networks. That is, a distribution P can have many P-maps, but all of them I-equivalent.

2 Undirected graphical models (UGM)

As we mentioned in Section 1.1, the UGM reflects the pairwise (non-causal) relationships of variables. We can easily write down the model and score specific configurations of the graph, but there is no explicit way to generate samples from the represented distribution. The contingency is constrained on node configurations.

2.1 A canonical example

Figure 4 shows a canonical example that comes from understanding complex scene: we want to judge whether the cropped piece is air or water.

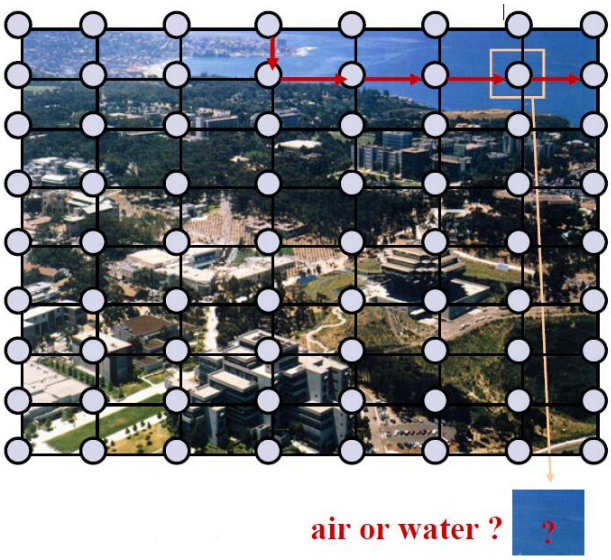

Figure 4: Example of grid model.

In order to solve the problem, we use a grid model here. The grid model has the history from atomic physics. This model naturally arises in image processing, lattice physics, etc. In the model, each node may represent a single "pixel", or an atom. In the above example, each node represents a small square of the image. The states of adjacent or nearby nodes are "coupled" due to pattern continuity or electro-magnetic forces, etc. The most likely joint-configurations usually correspond to a "low-energy" state.

2.2 Other examples of UGM

Apart from the application of understanding complex scene, UGM is applied is a large varity of cases: social networks, protein interaction networks, modeling Go and information retrieval.

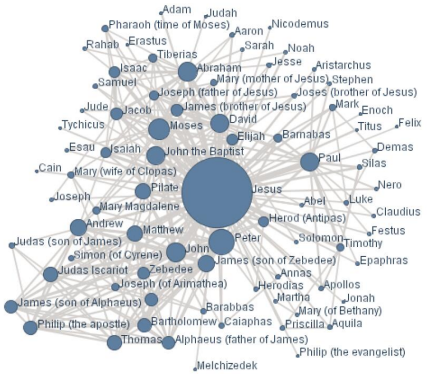

Figure 5: New Testament social networks.

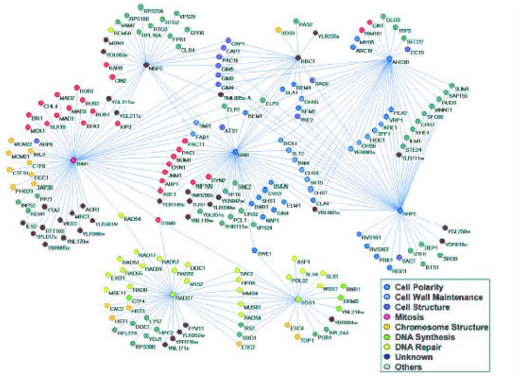

Figure 6: Protein interaction networks.

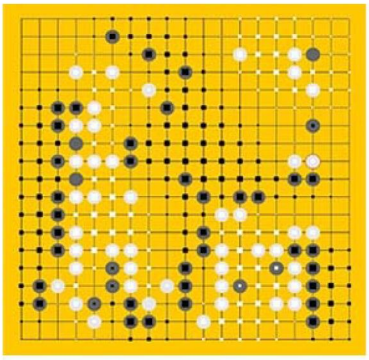

Figure 7: Modeling probability of becoming white or black in Go game.

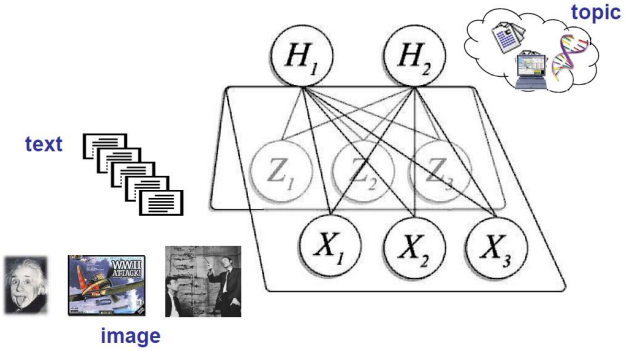

Figure 8: Information retrieval.

The Restricted Boltzmann Machine (RBM) is a simple and important UGM, which is also a type of neural network. Realizing that this is an UGM/MRF is useful and allows us to use the backpropagation algorithm.

3 Representation

We can represent the probability distribution of UGM with **potential functions**:

Definition 6 *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 y_c associated with the cliques of H, s.t.,*

$$P(x_1, x_2, ..., x_n) = \frac{1}{Z} \prod_{c \in C} \Psi_c(x_c), \tag{5}$$

where Z is known as the partition function:

$$Z = \sum_{x_1, x_2, ..., x_n} \prod_{c \in C} \Psi_c(x_c) \tag{6}$$

UGMs are also known as Markov Random Fields, or Markov networks. The potential function can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration. We call this form of distribution in Equation 5 as **Gibbs distribution**.

In the following part, we want to explain the representation from two aspects: clique potentials and independence properties.

3.1 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 interconnected. A (maximal) clique is a complete subgraph such that any superset $V'' \supset V'$ is not complete. A sub-clique is a not-necessarily-maximal clique.

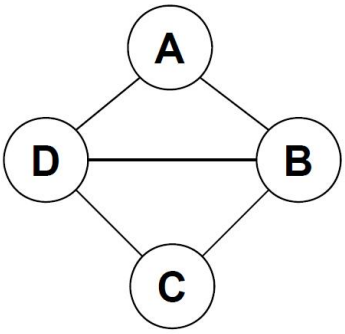

Figure 9: Example of cliques.

Figure 9 shows an example of cliques: $\{A, B, D\}$ and $\{B, C, D\}$ are max-cliques. The sub-cliques contains $\{A, B\}$ $\{C, D\}$ and all edges and singletons, etc.

3.1.1 Interpretation of clique potentials

Does the potential function here have some specific physical meanings? Let's take a concrete example in Figure 10.

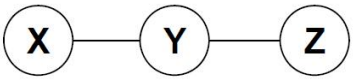

Figure 10: Example for illustrating the meaning of clique potentials.

The model implies $X \perp Z|Y$. This indepence statement implies (by definition) that the joint must factorize as:

$$p(x, y, z) = p(y)p(x|y)p(z|y). \tag{7}$$

We can write this as

$$p(x, y, z) = p(x, y)p(z|y) \tag{8}$$

or

$$p(x, y, z) = p(x|y)p(z, y). \tag{9}$$

However, we cannnot have all potentials be marginals, and cannot have all potential be conditionals. As a matter of fact, the positive clique potentials can only be thought of as the general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

3.1.2 Example UGM-using max cliques

Here we use the example from Figure 9.

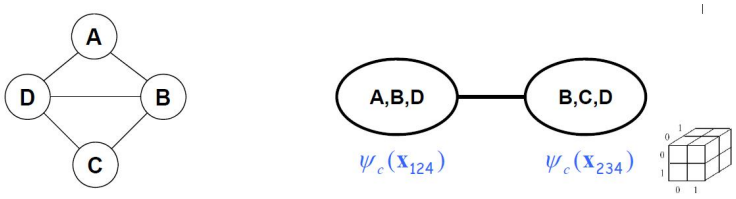

Figure 11: Example of UGM - using max cliques.

We can factorize the graph into two max cliques:

$$P'(x_A, x_B, x_C, x_D) = \frac{1}{Z}\Psi_c(x_{ABD})\Psi_c(x_{BCD}) \tag{10a}$$

$$Z = \sum_{x_A, x_B, x_C, x_D} \Psi_c(x_{ABD})\Psi_c(x_{BCD}). \tag{10b}$$

Both max cliques correspond to 3-D tables. For discrete nodes, in this way, we can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table.

3.1.3 Pairwise MRF

In this example, the probability distribution is proportional to the product of pairwise factor product between connecting nodes. It can be expressed as:

$$P(x_A, x_B, x_C, x_D) = \frac{1}{Z}\prod_{ij \in E}\psi_{ij} \tag{11a}$$

$$= \frac{1}{Z}\psi_{AB}(x_{AB})\psi_{AD}(x_{AD})\psi_{BD}(x_{BD})\psi_{BC}(x_{BC})\psi_{CD}(x_{CD}) \tag{11b}$$

$$Z = \sum_{x_A, x_B, x_C, x_D}\prod_{ij \in E}\psi_{ij} \tag{11c}$$

One can can represent $P(X_{A:D})$ using 5 2D tables instead of one 4D table.

3.1.4 Canonical Representation

A distribution P_Ψ with $\Psi = \{\psi_1(\mathbf{D_1}, .., \mathbf{D_k})\}$ factorizes over a Markov Network H if each $\mathbf{D_k}$ (k = 1,..,K) is a complete sub-graph of H. For example, in the above graph, it can be expressed as:

$$P(x_A, x_B, x_C, x_D) = \frac{1}{Z}\psi_{ABD}\psi_{BCD}\psi_{AB}\psi_{AD}\psi_{BD}\psi_{BC}\psi_{CD}\psi_A\psi_B\psi_C\psi_D \tag{12a}$$

$$Z = \sum_{x_A, x_B, x_C, x_D}\psi_{ABD}\psi_{BCD}\psi_{AB}\psi_{AD}\psi_{BD}\psi_{BC}\psi_{CD}\psi_A\psi_B\psi_C\psi_D \tag{12b}$$

3.2 Independence properties

Global Independencies: A set of nodes Z separates X and Y in H, denoted $sep_H(X : Y|Z)$, if there is no active path between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$. Global independencies associated with H are defined as:

$$I(H) = \{X \perp Y|Z : sep_H(X : Y|Z)\} \tag{13}$$

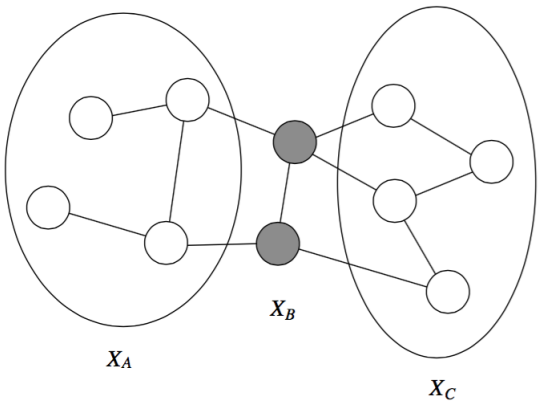

Figure 12: In this, the set X_B separates X_A from X_C. All paths from X_A to X_C pass through X_B

In Figure 12, B separates A and C if every path from a node in A to a node in C passes through a node in B. It is written as $sep_H(A : C|B)$. A probability distribution satisfies the **global Markov property** if for any disjoint A, B, C such that B separates A and C, A is independent of C given B.

$$I(H) = \{A \perp C|B : sep_H(A : C|B)\} \tag{14}$$

3.2.1 Local Markov Independencies

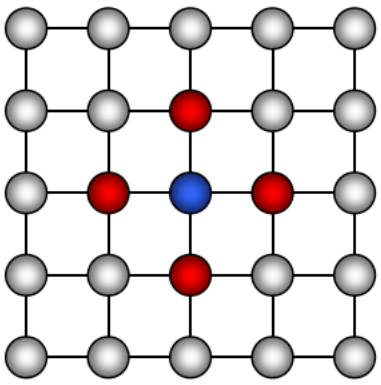

Figure 13: Illustration of Markov Blanket in undirected graph

For a given graph H, for each node $X_i \in V$, there is a unique Markov blanket of X_i, denoted as MB_{X_i}, which is the set of neighbors of X_i in the graph H (those that share an edge with X_i).

$$X_i \perp X - X_i - X_{MB_i} | X_{MB_i}$$
$$P(X_i|X_{-i}) = P(X_i|X_{MB_i})$$

Mathematically, the local Markov independencies associated with H is defined to be:

$$I_l(H) : \{X_i \perp V - \{X_i\} - MB_{X_i} | MB_{X_i} : \forall i\} \tag{15}$$

In other words, local independencies state that X_i is independent of the rest of the nodes in the graph given its immediate neighbors.

3.2.2 Soundness and completeness of global Markov property

Definition 7 *An undirected graph H is an I-map for distribution P if $I(H) \subseteq I(P)$, i.e., P entails $I(H)$.*

Definition 8 *P is a **Gibbs distribution** over H if it can be represented as*

$$P(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}) = \frac{1}{Z} \prod_{c \in C} \Psi_c(\mathbf{x_c}) \tag{16}$$

Theorem 9 *(**Soundness**): If P is a Gibbs distribution that factorizes over H, then H is an I-map of P.*

Theorem 10 *(**Completeness**): If X and Y are not separated given Z in H ($\neg sep_H(X; Z|Y)$), then X and Y are dependent given Z, in some distributon P represented as ($X \not\perp_P Z|Y$) that factorizes over H.*

For proof of the above theorems, one can refer to Koller and Friedman [3].

3.2.3 Other Markov properties

For directed graphs, we define I-maps in terms of local Markov properties, and derive global independence while in undirected graphs, we define I-maps in terms of global Markov properties, derive local independence.

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

$$I_p(H) = \{(X \perp Y | V - \{X, Y\}) : \{X, Y\} \notin E\} \tag{17}$$

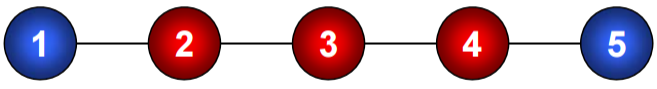

Figure 14: Illustration of pairwise independence in undirected graph. Nodes coloured in red are observed.

Example: In Fig 14, we have the following independence

$$X_1 \perp X_5 | \{X_2, X_3, X_4\} \tag{18}$$

3.2.4 Relationship between local and global Markov properties

- For any Markov Network H, and any distribution P, we have that if $P \models I_l(H)$ then $P \models I_p(H)$
- For any Markov Network H, and any distribution P, we have that if $P \models I(H)$ then $P \models I_l(H)$
- Let P be a positive distribution. If $P \models I_p(H)$, then $P \models I(H)$

Corollary: The following three statements are equivalent for a positive distribution P

- $P \models I_l(H)$
- $P \models I_p(H)$
- $P \models I_(H)$

Above equivalence relies on the positivity assumption of P. For nonpositive distributions, there are examples of distributions P, there are examples which satisfies one of these properties, but not the stronger property.

3.2.5 Hammersley-Clifford Theorem

If arbitrary potentials are utilized in the following product formula for probabilities, then the expression for P is:

$$P(x_1, x_2, ..., x_n) = \frac{1}{Z} \prod_{c \in C} \Psi_c(\mathbf{x_c}) \tag{19a}$$

$$Z = \sum_{x_1,.,x_n} \prod_{c \in C} \Psi_c(\mathbf{x_c}) \tag{19b}$$

Then the family of probability distributions obtained is exactly that set which respectes the qualitative specification (the conditional independecence relations) described earlier.

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.

3.2.6 Perfect maps

A Markov network H is a perfect map for P if for any $X; Y; Z$ we have that

$$sep_H(X; Z|Y) \Leftrightarrow P \models (X \perp Z|Y) \tag{20}$$

Theorem 11 *Not every distribution has a perfect map as undirected graphical model.*

Proof: This is proof by counterexample. There is no undirected graphical model which can encode the independencies in a v-structure $X \rightarrow Y \leftarrow Z$. ■

3.2.7 Exponential Form

Clique potential $\Psi_c(x_c)$ is represented in an unconstrained form using a real-value "energy" function $\phi_c(x_c)$. $\phi_c(x_c)$ a also called potential function.

$$\Psi_c(\mathbf{x_c}) = \exp\{-\phi_c(x_c)\} \tag{21}$$

This gives the joint an additive structure as defined below:

$$P(\mathbf{x}) = \frac{1}{Z}\exp\{-H(x)\} \tag{22}$$

where the $H(x)$ in the exponent is called the "free energy" and is given as:

$$H(x) = \sum_{c \epsilon C} \phi_c x_c \tag{23}$$

The exponential representation ensures that the distribution is positive. In physics, this is called the "Boltzmann distribution" while in statistics, it is known as log-linear models.

3.3 Boltzmann machines

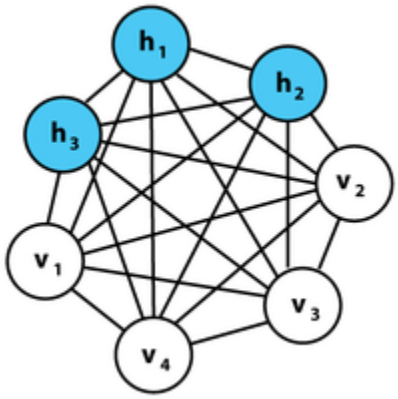

Figure 15: Boltzmann machine in which each node is connected to every other node. Image has been taken from Wikipedia

A Boltzmann machine is represented as a fully connected graph with pairwise (edge) potentials on binary-valued nodes (for $x_i \in \{-1, +1\}$ or $x_i \in \{0, 1\}$). The probability distribution is represented as:

$$P(\mathbf{X}) = \frac{1}{Z}\exp\left(\sum_{i,j} w_{i,j}x_i x_j + \sum_i u_i x_i + C\right) \tag{24}$$

Energy of the above configuration can be represented in vector form using parameters μ and Θ as:

$$H(x) = (x-\mu)^\intercal \Theta (x-\mu) \tag{25}$$

Boltzmann machines have a close relationship with activation function of neurons. Probability distribution of X given the observed connected nodes is given by a sigmoid function similar to the concept of RBM as discussed in the next sections. This model is also a very simple form of a probabilistic recurrent neural network.

3.4 Ising models

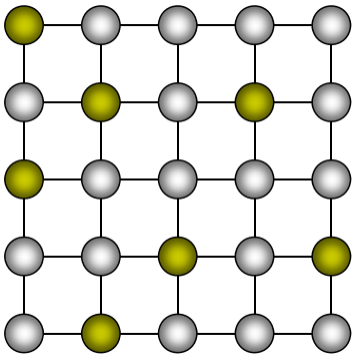

Figure 16: Ising model in which each node is connected to its neighbouring node

The concept of the Ising model came from statistical physics as a model for energy of a physical system which consisted of interactions among atoms which can be represented as nodes in an undirected graph (16). Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors. In this model, each node is represented by a random variable X_i which can take only 2 states $\{-1, +1\}$. Energy function is represented by the following distribution:

$$P(\mathbf{X}) = \frac{1}{Z} \exp\left(\sum_{i<j, j \epsilon N_i} w_{i,j} x_i x_j + \sum_i u_i x_i \right) \tag{26}$$

The first term represents the energy function associated with the edges. When the neighbouring nodes have the same state, the effect is positive correlation and vice versa. Parameters u_i represent bias or node potentials.

Ising models are a special form of Boltzmann machine where $w_{ij} \neq 0$ if i, j are neighbors. These models find applications in image processing such as image denoising where nodes are assumed to be image pixels and edge potential encourages nearby pixels to have similar intensities.

Also, a Potts model is defined as a multi-state Ising model.

3.5 Restricted Boltzmann Machines (RBM) [2]

3.5.1 Introduction

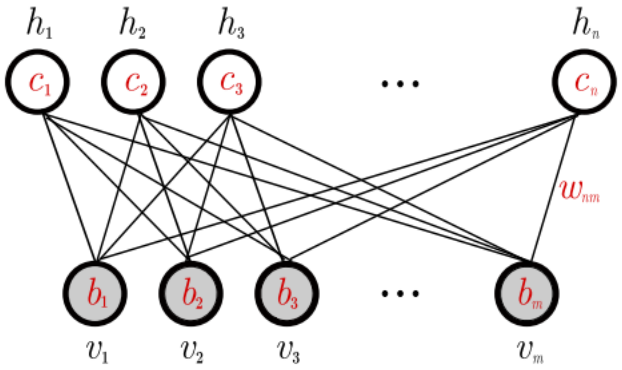

Figure 17: RBM as an undirected graph with m visible nodes and n hidden nodes

The Restricted Boltzmann Machine (RBM) is directly inspired by the Boltzmann Machine. It is a bipartite graph consisting of visible units and hidden units. As can be seen from Figure 17, it has only connections between the hidden and visible variables but not between two variables of the same layer. Hidden units can represent latent features in data. In this model, the joint probability distribution of the model involves pairwise potential between hidden and visible units and their respective biases. It is given as:

$$p(\mathbf{v}, \mathbf{h}|\theta) = \exp\{\sum_i \theta_i \phi_i(v_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(v_i, h_j) - A(\theta)\} \tag{27}$$

The conditional probability distribution of one layer given the other layer can be factorized in terms of the products of CPD of nodes in a layer. Mathematically, it can be stated as:

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{n} p(h_i|\mathbf{v}) \tag{28}$$

$$p(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{m} p(v_i|\mathbf{h}) \tag{29}$$

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

$$p_{ind}(\mathbf{h}) \propto \prod_j \exp\{\theta_j g_j(h_j)\} \tag{30}$$

$$p_{ind}(\mathbf{v}) \propto \prod_i \exp\{\theta_i f_i(v_i)\} \tag{31}$$

Then the joint can be written in the following form using functions f and g:

$$p(\mathbf{v}, \mathbf{h}|\theta) = \exp\{\sum_i \vec{\theta}_i \vec{f}_i(v_i) + \sum_j \vec{\lambda}_j \vec{g}_j(h_j) + \sum_{i,j} \vec{f}_i^T(v_i) \mathbf{W}_{i,j} \vec{g}_j(h_j)\} \tag{32}$$

The conditional probability density of a visible layer node conditioned upon the hidden layer is given as follows:

$$p(v_i|\mathbf{h}) = \exp\{\sum_a \hat{\theta}_{ia} f_{ia}(v_i) + A_i(\{\hat{\theta}_{ia}\})\} \tag{33a}$$

$$\hat{\theta}_{ia} = \theta_{ia} + \sum_{jb} W_{ia}^{jb} g_{jb}(h_j) = \theta_{ia} + \sum_j \vec{W}_{ia}^{j} \vec{g}_j(h_j) \tag{33b}$$

The conditional probability density of a hidden layer node conditioned upon the visible layer is given as follows:

$$p(h_j|\mathbf{v}) = \exp\{\sum_b \hat{\lambda}_{jb} g_{jb}(h_j) + B_j(\{\hat{\lambda}_{jb}\})\} \tag{34a}$$

$$\hat{\lambda}_{jb} = \lambda_{jb} + \sum_{ia} W_{jb}^{ia} f_{ia}(v_i) = \lambda_{jb} + \sum_i \vec{W}_{jb}^{i} \vec{f}_i(v_i) \tag{34b}$$

We observe that the probability of a hidden unit having state of 1 (turned "*on*") is independent of the states of other hidden units given the visible units. Similarly, visible units are independent of each other given the hidden units. This property of RBM's makes inference using blocked Gibbs Sampling extremely efficient. This allows all hidden units to be sampled independently followed by sampling of visible units.

Contrastive Divergence (CD) algorithm is used to learn the weights of RBM's as:

$$\nabla w_{ij} = \epsilon_w(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model})$$

Above, $\langle\rangle$ is the expectation operator. $\langle v_i h_j \rangle_{model}$ is calculated by alternate sampling of visible layer states given the hidden layer followed by sampling of hidden layer units given the visible layer. CD is an approximation to the true gradient of the likelihood function but works well in practice.

3.6 RBM for text modeling

One application of RBM is in modeling text data where hidden units can represent semantic topics when the visible units are trained on a bag of words representation of documents. The conditional probability distribution of topic given word counts can be represented by a Gaussian distribution and the conditional distribution of word given topics can be represented by binomial distribution.

$$p(\mathbf{h}|\mathbf{v}) = \prod_j Normal_{h_j}[\sum_i \vec{W}_{ij}\vec{x_i}, 1] \tag{35}$$

$$p(\mathbf{v}|\mathbf{h}) = \prod_i Bi_{v_i}[N, \frac{\exp\Big(a_j + \sum_j W_{ij} h_j\Big)}{1 + \exp\Big(a_j + \sum_j W_{ij} h_j\Big)}] \tag{36}$$

which give rise to the following marginal:

$$p(\mathbf{v}) \propto \exp\{(\sum_i a_i x_i - log\Gamma(x_i) - log\Gamma(N - x_i)) + \frac{1}{2}\sum_j (\sum_i W_{i,j} x_i)^2\} \tag{37}$$

3.6.1 Replicated Softmax: An undirected Topic Model [4]

RBMs can be used to extract low dimensional latent semantic representations from documents using their text content similar to topic models. Say the training data consists of bag of words representation for document texts. Let's say that the size of vocabulary is K, and there are N documents. So, we have N x K matrix M, where each entry M_{ij} represents the count of word j in document i.

The hidden units of the RBM represent binary topic features and the visible units represents the probability of count of words in a document (softmax). In this, the bias term of hidden variables is scaled by length of document. The conditional probability of hidden units given the visible units is

$$p(H_i = 1|\mathbf{v}) = \sigma(\sum_{k=1}^{K} W_{ki} v_k + N c_i) \tag{38}$$

The conditional probability of visible units is given by a softmax expression as:

$$p(V_k = 1|\mathbf{h}) = \frac{\exp\left(\sum_{i=1}^{n} W_{ik} h_i + b_k\right)}{\sum_{k=1}^{K} \exp\left(\sum_{i=1}^{n} W_{ik} h_i + b_k\right)} \tag{39}$$

In this model also we apply Contrastive Divergence to learn the weights.

3.7 Conditional Random Fields (CRF)

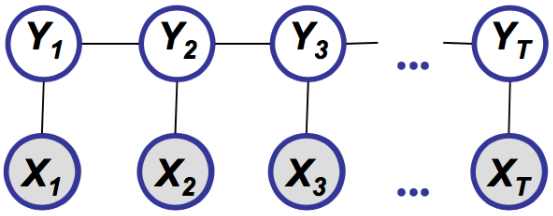

Figure 18: Graphical Model for CRF

CRFs are undirected graph representation in which parameters are used to encode a conditional distribution $P(\mathbf{Y}|\mathbf{X})$, where Y_i are target variables and X is a disjoint set of observed variables X_i as shown in Figure 18. CRFs are pretty flexible models in that they allow the features X to be non-independent, thus it relaxes strong independence assumptions commonly assumed in generative models such as Naive Bayes. The probability of a transition between labels in CRF may depend on both past and future observations.

If the graph $G = (V, E)$ of $\mathbf{Y}$ is a tree, the conditional distribution over the label sequence $Y = y$, given $X = x$, according to Hammersley-Clifford theorem of random fields is given as:

$$P_\theta(\mathbf{Y}|\mathbf{X} = x) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x)\} \tag{40}$$

In the above equation: x is a data sequence, y is a label sequence, v is a vertex from vertex set V, e is an edge from edge set E over V, f_k and g_k are given and fixed, g_k is a Boolean vertex feature, f_k is a Boolean edge feature, k is the number of features, λ_k, μ_k are parameters to be estimated, Z is normalization constant.

In some applications, we allow arbitrary dependencies on input x, then the conditional expression can be represented as:

$$P_\theta(\mathbf{Y}|\mathbf{X}=x) = \frac{1}{Z(\theta,x)} \exp\{\sum_c \theta_c f_c(x, y_c)\} \tag{41}$$

One can use approximate inference techniques for querying the graph.

3.8 Summary: Conditional independence semantics in MRF

- In Markov Random Fields, the structure of the probability distribution is represented by an undirected graph.
- A node is conditionally independent of every other nodes in the network given its connected neighbors.
- Local contingency functions (potentials) and the cliques in the graph completely determine the joint distribution.
- Graph structure can give correlation between variables but does not have an explicit way to generate samples.

4 Structure Learning

The objective in structure learning is to find the most probable graph representation of the probability distribution given a set of observed samples. The number of possible graph structures over n nodes is of order $O(2^{n^2})$. Also, the number of trees over n nodes is of complexity $O(n!)$. So we can't use brute force search for structure learning as the number of possible search cases is exponential.

Using the property that in a tree, each node has only one parent, the Chow-Liu algorithm [1], as explained below, can be used to find the exact solution for the optimal tree.

4.1 Chow-Liu Tree Learning Algorithm

This is an algorithm for finding the best tree-structured network. Let $P(X)$ be true distribution and $T(X)$ be a tree structured network. The Chou-Liu algorithm minimizes KL Divergence between $P(X)$ and $T(X)$.

$$KL(P(X)||T(X)) = \sum_k P(\mathbf{X}=\mathbf{k}) log \frac{P(\mathbf{X}=k)}{T(\mathbf{X}=\mathbf{k})} \tag{42}$$

In order to minimize KL divergence, it suffices to find T, that maximizes the sum of mutual information over edges. Empirical distribution can be computed as:

$$P(X_i, X_j) = \frac{count(x_i, x_j)}{M} \tag{43}$$

Mutual Information is given as:

$$I(X_i, X_j) = \sum_{x_i, x_j} p(x_i, x_j) log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{44}$$

General steps in computing optimal tree BN are:

- Compute maximum weight spanning tree.
- For finding direction in Bayesian Network, pick any node as root, do breadth-first-search to define directions.

4.2 Structure Learning for General Graphs

Theorem 12 *The problem of learning a Bayesian Network structure with at most d parents is an NP-hard problem for any fixed d >= 2.*

Most structure learning algorithms use heuristics which exploit score decomposition. Some examples of this method are:

- Greedy search through space of node-orders
- Local search of graph structures

References

[1] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968.

[2] Asja Fischer and Christian Igel. *An Introduction to Restricted Boltzmann Machines*, pages 14–36. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[3] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[4] Ruslan Salakhutdinov and Geoffrey Hinton. Replicated softmax: An undirected topic model. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 1607–1614, USA, 2009. Curran Associates Inc.