10-708: Probabilistic Graphical Models 10-708, Spring 2012

3 : Representation of Undirected GMs

Lecturer: Eric P. Xing *Scribes: Nicole Rafidi, Kirstin Early*

Last Time

In the last lecture, we discussed directed graphical models (aka Bayesian Networks) and their semantics. We showed how a Directed Acyclic Graph (DAG) can represent the independence relations of a joint probability distribution. We can find these relationships by exploring moralized ancestral graphs, and we learned algorithms that let us do that quickly. We also introduced the concept of I-maps: a DAG is an I-map for a distribution if it encodes all the independence relations of that distribution. Note that intuitively, a complete graph can be an I-map of any distribution. A more useful notion is the minimal I-map, which is the graph that minimally encodes the independence relationships of the distribution. A minimal I-map of a distribution P is defined as a DAG, where removing one edge would render it no longer an I-map of P. A distribution may have several minimal I-maps, which comes from the notion of I-equivalence. A good example is $A \rightarrow B \rightarrow C$ and $A \leftarrow B \leftarrow C$, which both encode the independence relationship $A \perp C | B$. A key issue is that this can make deriving scientific insight from graphical models difficult, because different graphs (which have different domain-specific implications) are indistinguishable via I-map.

We also touched on the issue of perfect maps: a DAG G is a perfect map for a distribution P if I(P) = I(G), where I(A) is the I-map of A. Note that not every distribution has a perfect map that is also a DAG. Here is a simple counter example:

$$A \perp C | \{B, D\}, \quad B \perp D | \{A, C\}$$

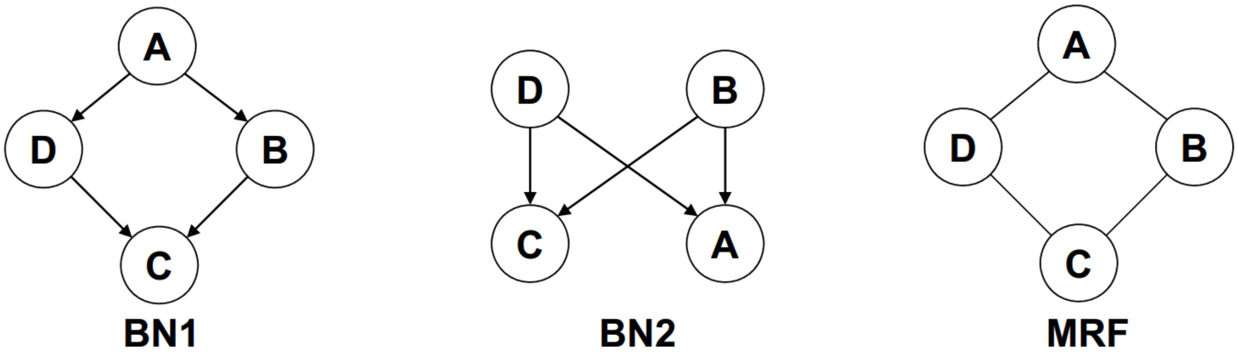

Figure 1: Neither BN1 nor BN2 can represent both of these conditional independencies, but the MRF can.

It is impossible for a DAG to capture both of these independence relations at the same time. Thus, there is a portion of the space of distributions that we cannot express with directed graphical models. This motivates an alternative, complementary approach: undirected graphical models, aka Markov Random Fields.

1 Undirected Graphical Models (UGM) Representation

UGMs are very similar to directed graphical models; however, connections in UGMs capture pairwise relationships that are a symmetric kind of correlation/association. Causation cannot be inferred from the edges of a UGM, but such a graph can be used to capture distributions in an analogous way to Bayesian Networks.

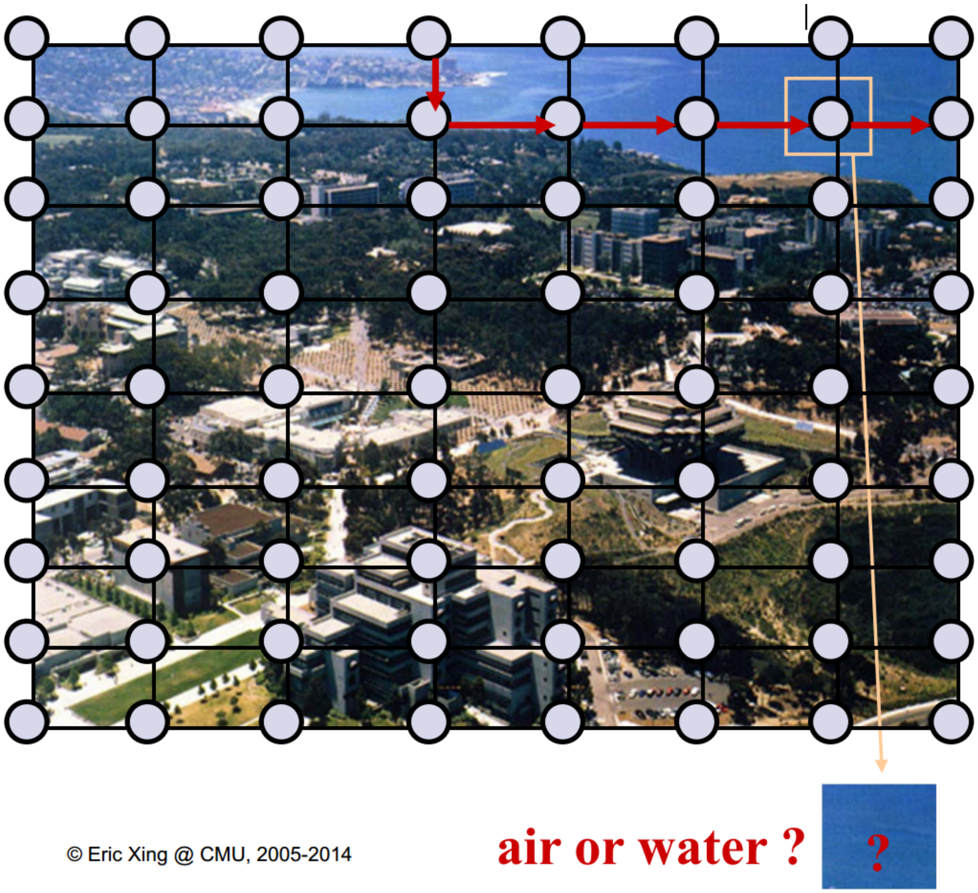

Figure 2: Using an UGM to predict whether a section of a scene is air or water.

Definition An undirected graphical model represents a distribution $P(X_1, ...X_n)$ defined by an undirected graph $\mathcal{H}$, and a set of *positive-valued* potential functions ψ_c corresponding to each clique $c \in C$ of $\mathcal{H}$ such that:

$$P(X_1, ...X_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c)$$

where $Z = \sum_{X_1,...,X_n} \prod_{c \in C} \psi_c(X_c)$ is known as the partition function and acts as a marginalization constant. Without it, the product of potentials will *not* represent a probability distribution. It is important to remember that the potential functions are not conditional probabilities in their own right.

UGMs are generally known as Markov Networks or Markov Random Fields. The potential function represents the coupling strength of the clique, which indicates how much the nodes within that clique covary. Thus we can see that each model has two components: the graph structure itself (which yields the independence properties), and the potential function assignments, which can be used to derive the distribution represented by the graph.

1.1 Global Markov Independencies

The definition of separation for UGMs is simply that B separates A and C if every path from a node in the set A to a node in C passes through a node in set B.

A probability distribution satisfies the *global Markov property* if for any disjoint A, B, C where B separates A and C, A is independent of C given B. This definition is sound and complete:

Completeness: If P is a Gibbs distribution over $\mathcal{H}$, then $\mathcal{H}$ is an I-map of P.

Soundndess: If $\neg\text{sep}_{\mathcal{H}}(X, Z|Y)$, then $\exists P$ that factorizes over $\mathcal{H}$ such that $X \not\perp_P Z|Y$.

Note that we can now express the counterexample from the first section.

1.2 Local Markov Independencies

The *pairwise Markov independencies* for an undirected graphical model $\mathcal{H} = (V, E)$ are

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

The unique Markov blanket of a node in an undirected graphical model is the set of its neighbors. The *local Markov independency* is that a node is independent of the rest of the graph given its Markov blanket:

$$I_\ell(\mathcal{H}) = \{X \perp V \setminus (X \cup N_{\mathcal{H}}(X)) | N_{\mathcal{H}}(X) : X \in V\}.$$

1.3 Relationship between local and global Markov properties

Theorem: $P \models I_\ell(\mathcal{H}) \implies P \models I_p(\mathcal{H})$

Theorem: $P = I(\mathcal{H}) \implies P \models I_\ell(\mathcal{H})$

Theorem: $P > 0$ and $P \models I_p(\mathcal{H}) \implies P \models I(\mathcal{H})$

Corollary: For a *positive* distribution P, global, local, and pairwise indepedencies are equivalent.

1.4 Cliques

A clique is a complete subgraph. A maximal clique is the largest possible complete subgraph. We call the maximal clique a max-clique. Non-maximal cliques are called sub-cliques.

1.5 Interpretation of Clique Potentials

Clique potentials in their original form cannot be interpreted as probability distributions. This is related to the symmetry of independence relations in the graph (e.g., Figure 3). Since we can't interpret potentials as probabilities, we may as well make them positive numbers and not worry about having things sum to one.

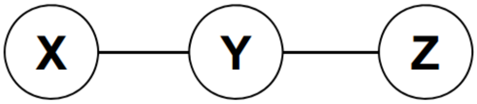

Figure 3: A symmetric independence relation in an undirected graph.

1.6 Calculation of the Distribution

Using max-cliques lets us calculate the Gibbs distribution represented by the graph with very few terms. The expense and difficulty in this calculation lies in the computation of the partition function, which must capture all the possible configurations of the cliques. Alternatively, you could use sub-cliques (e.g. pairwise cliques). There are more terms in the distribution product, but the partition functions are much easier to compute.

Are these two methods equivalent? The I-maps are certainly equivalent, because we have not changed the graph structure, only our clique divisions. However the distributions calculated are not the same. In general, the space of values that can be calculated by the max-clique representation is larger than that calculated by the sub-clique representation. The method you choose depends on your design constraints, and the kinds of relationships. The most general method is canonical representation, which does not use cliques but calculates over individual nodes. The canonical form and the sub-clique form are special cases of the max-clique form.

This is summarized in the *Hammersley-Clifford Theorem*: If a Markov network $\mathcal{H}$ is an I-map for a **positive** distribution P, then P is a Gibbs distribution that factorizes over $\mathcal{H}$.

2 Perfect Maps Revisited

A Markov Network can also be a perfect map for a distribution so long as its separation properties capture the independence relations of the distribution. However, like with Bayesian Networks, not every distribution can be captured in a UGM. In fact, the space of distributions cannot be completely captured by the union of UGMs and Bayesian Networks.

3 Exponential Form

Constraining clique potentials to be positive could be annoying. We can get around this by using a positive, real-valued energy function, which is typically a negative exponential of the original potential. This has nice properties when we calculate products for distributions:

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

where $H(\mathbf{x})$ is the "free energy." This model is called the Boltzmann distribution in physics; log-linear in statistics.

4 Example Models

4.1 Boltzmann Machines

A Boltzmann Machine is a fully connected graph with pairwise potentials on binary-valued nodes. The energy function for this is expressed in sub-clique form, which comes from the physics tradition. Here is the joint distribution for the Boltzmann machine in Figure 4:

$$\begin{aligned}
p(x_1, x_2, x_3, x_4) = & \frac{1}{Z} \exp \left\{ \sum_{i,j} \phi_{ij}(x_i, x_j) \right\} \\
= & \frac{1}{Z} \exp \left\{ \sum_{i,j} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \right\} \\
= & \frac{1}{Z} \exp \left\{ (x - \mu)^T \Theta (x - \mu) \right\}
\end{aligned}$$

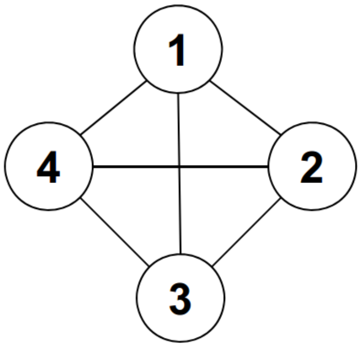

Figure 4: An example Boltzmann machine.

4.2 Ising Models

This is the grid model that we saw previously, which is equivalent to what is called a sparse Boltzmann machine. Another variant in the Potts model, which is a multi-state Ising model.

4.3 Restricted Boltzmann Machines

This is inspired by the Boltzmann Machine and is responsible for much of the deep learning craze. An RBM consists of many layers. Within each layer, there are two sublayers: one of hidden units (factors, h_j), and one of visible units (x_i). The probability function for an RBM is

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

In the undirected form of the RBM, factors are marginally dependent, but conditionally independent given observations on the visible nodes. This allows us to do iterative Gibbs sampling over the model.

We can define the RBM with “a constructive definition.” Sometimes it becomes too onerous to have to express pairwise joint probabilities. Maybe we want the local conditional distributions we work with to have a nice form, e.g. Gaussian. To accomplish this, we start by writing down the distribution in our desired form. The weights learned in this new setting are slightly different than those in the individual settings, but the relationship can be derived in closed form.

4.4 Conditional Random Fields

A conditional random field is a discriminative UGM that models the conditional probability of a label sequence (hidden) given an observation sequence. This model does not assume independence among the features on the observations. The probability distribution is given by

$$p_\theta(y|x) = \frac{1}{Z(x)} \exp\left\{ \sum_{e\in E,k} \lambda_k f_k(e, y|_e, s) + \sum_{v\in V,k} \mu_k g_k(v, y|_v, x) \right\},$$

where x is a data sequence, y is a label sequence, v is a vertex from the set V of label random variables, e is an edge from the edges E over V, k is the number of features, f_k is a given fixed Boolean edge feature, g_k is a given fixed Boolean vertex feature, $\theta = (\lambda_1, ..., \lambda_n; \mu_1, ..., \mu_n)$ are parameters to be estimated, $y|_e$ is the set of components of y defined by edge e, and $y|_v$ is the set of components of y defined by vertex v.

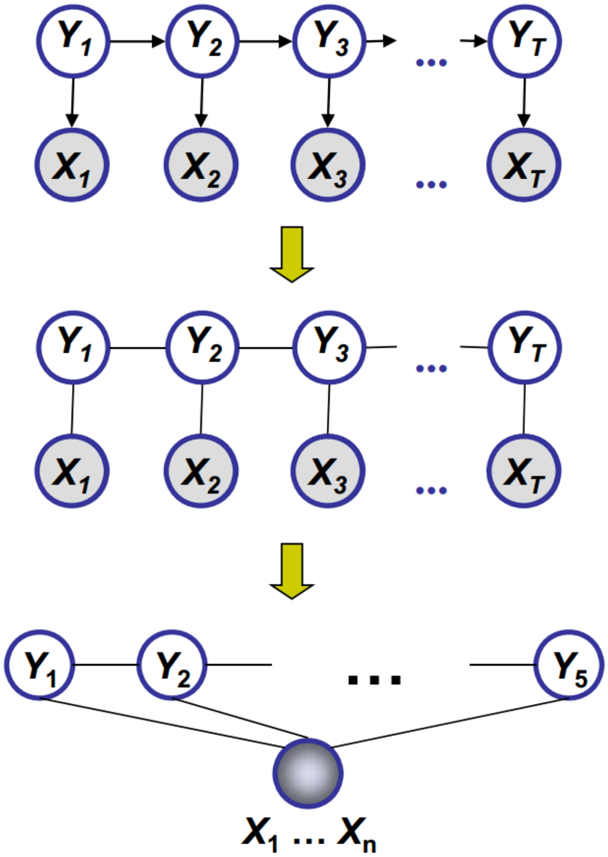

Figure 5: A conditional random field.

5 Summary

- Undirected graphical models indicate relatedness between random variables (rather than causality).
- The graph separation criteria characterize local and global independencies in UGMs.
- Clique potentials quantitatively define UGMs.
- The partition function usually make it intractable to compute the likelihood for UGMs, complicating both inference and likelihood-based learning.
- UGMs can define either joint or conditional distributions.