

STA414

Week 3 - 1/2: Markov Random Fields

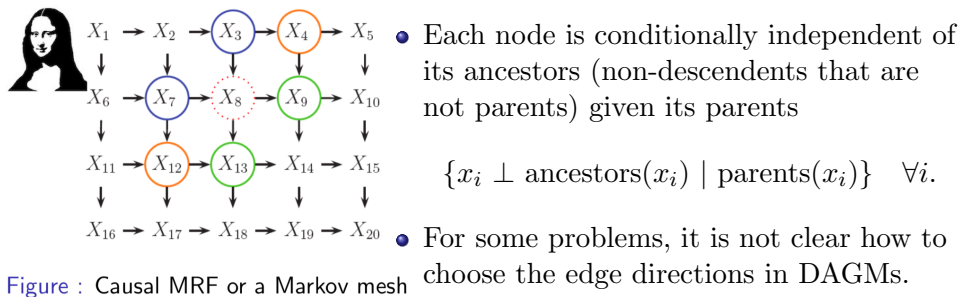
David Duvenaud

University of Toronto. Slide credit to Murat Erdogdu

- Assignment 1 is released later today.
 - ▶ Decision theory
 - ▶ Discrete latent variable model of images
 - ▶ Gradient-based regression
 - ▶ Comprehensive Python starter code in colab.
- In-person lectures pushed back a week by dept. New first in-person lecture: Feb 7th
- Online alternatives may be available
- Preference for midterm?

- Markov Random Fields (MRFs) and Undirected Graphical Models
- Exact Inference
- Tutorial:
 - ▶ Basic MLE derivations (for Q1 of HW1)
 - ▶ Simple neural net demo
 - ▶ Worked examples of UGMs and exact inference

DAGMs are sometimes a bad fit



Markov blanket

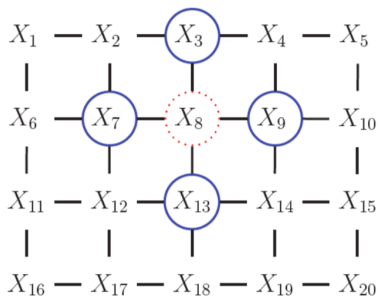
The set of nodes that renders a node conditionally independent of all the other nodes in the graph is called that node's Markov blanket (*mb*).

$$mb(X_8) = \{X_3, X_7\} \cup \{X_9, X_{13}\} \cup \{X_{12}, X_4\}$$

Excercise: What is the Markov blanket for X_8 ? Perhaps counterintuitive.

Markov Random Fields

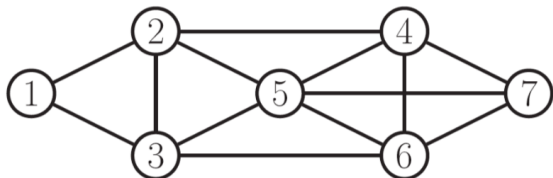
- Undirected graphical models, also called Markov random fields (MRFs), are a set of random variables where the dependencies are described by an undirected graph.
- The nodes in the graph represent random variables. However, in contrast to DAGMs, edges represent unnormalized interactions between neighbors (as opposed to conditional distributions).



Dependencies in MRFs

The following rule expresses which nodes are conditionally independent in MRFs:

- Global Markov Property (G):** $x_A \perp x_B | x_C$ iff x_C separates x_A from x_B . That is, there is no path in the graph between A and B that doesn't go through x_C .



$$\{x_1, x_2\} \perp \{x_6, x_7\} | \{x_3, x_4, x_5\}$$

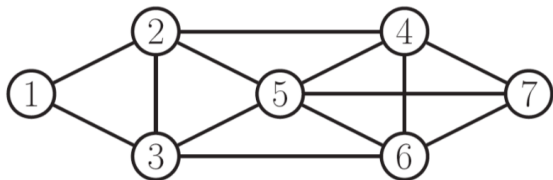
Dependencies in MRFs

A corollary of the Global Markov Property:

- 2 **Markov Blanket** (L): A set of nodes that renders a node t conditionally independent of all the other nodes in the graph

$$t \perp (\text{all} \setminus cl(t)) \mid mb(t)$$

where $cl(t) = mb(t) \cup t$ is the closure of node t .



$$x_1 \perp \text{rest} \mid \{x_2, x_3\}$$

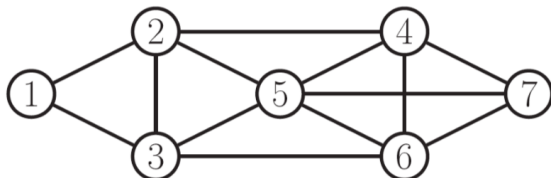
so $mb(x_1) = \{x_2, x_3\}$.

Dependencies in MRFs

Another corollary of the Global Markov Property:

3 Pairwise (Markov) Property (P):

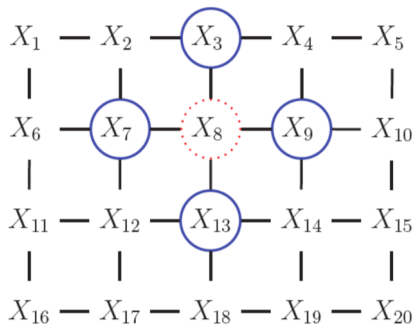
$s \perp t | (\text{all} \setminus \{s, t\}) \Leftrightarrow$ No edge between s & t .



$$x_1 \perp x_7 | \text{rest}$$

iff there are no edges between x_1 and x_7 . E.g. $1 \not\perp 2 | \text{rest}$

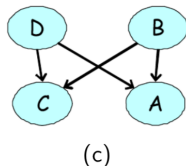
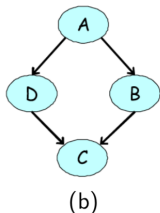
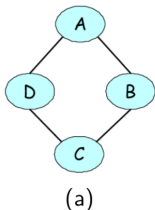
Image MRF



- Global: $\{X_1, X_2\} \perp \{X_{15}, X_{20}\} | \{X_3, X_6, X_7\}$
- Local: $X_1 \perp \text{rest} | \{X_2, X_6\}$
- Pairwise: $X_1 \perp X_{20} | \text{rest}$

Not all MRFs can be represented as DAGMs

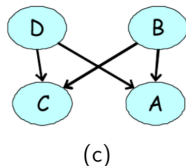
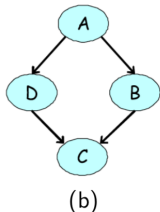
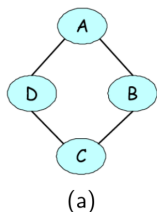
Take the following MRF for example (a) and our attempts at encoding this as a DAGM (b, c).



- Two conditional independencies in (a):
 - ▶ 1. $A \perp C | D, B$
 - ▶ 2. $B \perp D | A, C$

Not all MRFs can be represented as DAGMs

Take the following MRF for example (a) and our attempts at encoding this as a DAGM (b, c).

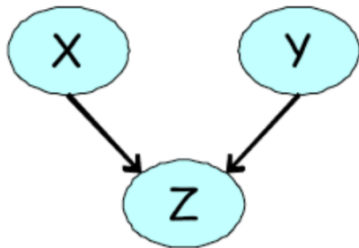


- Two conditional independencies in (a):
 - ▶ 1. $A \perp C | D, B$ 2. $B \perp D | A, C$
- In (b), we have the first independence, but not the second.
- In (c), we have both independencies, but B and D are marginally independent.

Not all DAGMs can be represented as MRFs

Not all DAGMs can be represented as MRFs.

E.g. explaining away:

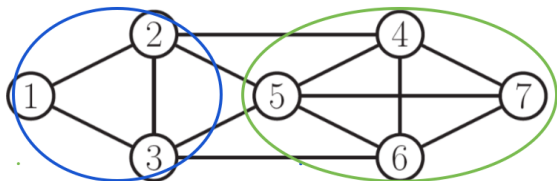


An undirected model is unable to capture the marginal independence, $X \perp Y$ that holds at the same time as $X \not\perp Y | Z$.

Cliques

A **clique** is a subset of nodes such that every two vertices in the subset are connected by an edge.

- **Maximal clique** is a clique that cannot be extended by including one more adjacent vertex.
- **Maximum clique** is a clique of the largest possible size in a given graph.



Above, maximal clique is shown in blue, while a maximum clique is shown in green.

Distributions Induced by MRFs

Let $x = (x_1, \dots, x_m)$ be the set of all random variables in our graph.

- Unlike in DAGMs, there is no topological ordering in MRFs, so the chain rule cannot be used to simplify the joint dist $p(x)$.
- We associate potential functions with each maximal clique: given a maximal clique c , the potential function (or factor)

$$\psi_c(x_c|\theta_c)$$

is a non-negative function, where x_c is the subset of variables in c and θ_c is some parameter.

- Joint distribution is proportional to product of clique potentials

$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$$

where \mathcal{C} is the set of all maximal cliques.

Any distribution whose conditional independencies are represented with an MRF can be represented this way.

Distributions Induced by MRFs

- A distribution $p(x) > 0$ satisfies the conditional independence properties of an undirected graph iff $p(x)$ can be represented as a product of factors, one per maximal clique, i.e.,

$$p(x|\theta) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$$

where \mathcal{C} is the set of all (maximal) cliques of G , and $Z(\theta)$ the **partition function**, defined as

$$Z(\theta) = \sum_x \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c).$$

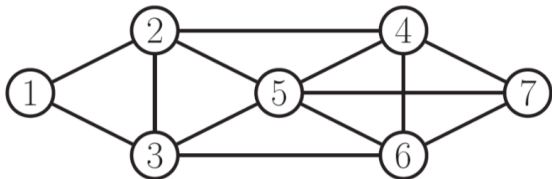
- The factored structure of the distribution makes it possible to more efficiently do the sums/integrals needed to compute it.

MRFs as Exponential Families

- If we use exp to enforce positivity, we can write the distribution as an exponential family:

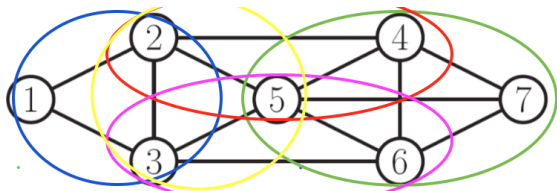
$$p(x|\theta) = \exp \left\{ \sum_{c \in \mathcal{C}} \log \psi_c(x_c|\theta_c) - \underbrace{\log Z(\theta)}_{=A(\theta)} \right\}$$

Example:



- How to factorize the undirected graph of our running example?
- How many maximal cliques are there ?

Example:



Lets see how to factorize the undirected graph of our running example:


$$p(x) \propto \psi_{1,2,3}(x_1, x_2, x_3) \psi_{2,3,5}(x_2, x_3, x_5) \psi_{2,4,5}(x_2, x_4, x_5) \psi_{3,5,6}(x_3, x_5, x_6) \\ \times \psi_{4,5,6,7}(x_4, x_5, x_6, x_7)$$

Representing potentials

If the variables are discrete, we can represent the potential functions as tables of (non-negative) numbers

$$p(A, B, C, D) = \frac{1}{Z} \psi_{A,B}(A, B) \psi_{B,C}(B, C) \psi_{C,D}(C, D) \psi_{A,D}(A, D)$$

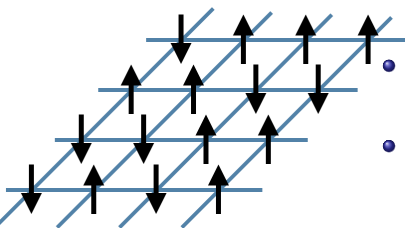
where



	$\phi_1[A, B]$	$\phi_2[B, C]$	$\phi_3[C, D]$	$\phi_4[D, A]$
a^0, b^0	30	100	1	100
a^0, b^1	5	1	100	1
a^1, b^0	1	1	100	1
a^1, b^1	10	100	1	100

Note that these potentials are not probabilities, but instead encode relative affinities between the different assignments. For example, in the above table, a^0, b^0 is taken to be 30X more likely than a^1, b^0 .

Example: Ising model



- The Ising model is an MRF that is used to model magnets.
- The nodes variables are spins, i.e., we use $x_s \in \{-1, +1\}$.

- Define the pairwise clique potentials as

$$\psi_{st}(x_s, x_t) = \begin{pmatrix} e^{W_{st}} & e^{-W_{st}} \\ e^{-W_{st}} & e^{W_{st}} \end{pmatrix}$$

where W_{st} is the coupling strength between nodes s and t .

- For $(x_s, x_t) = (1, -1)$ we have $\psi_{st}(1, -1) = e^{-W_{st}}$ or in compact form $\psi_{st}(x_s, x_t) = e^{x_s x_t W_{st}}$.
- If two nodes are connected we set $W_{st} = J$ and if they are not connected $W_{st} = 0$.
- This captures: it is more likely to have the same neighboring spins.

Ising model

- In compact form, for all pairs (s, t) , we can write

$$\psi_{st}(x_s, x_t) = e^{x_s x_t W_{st}} = \text{pairwise potential}$$

- This only encodes the pairwise behavior. We might want to add unary node potentials as well

$$\psi_s(x_s) = e^{-b_s x_s}$$

- The overall distribution becomes

$$p(x) \propto \prod_{s \sim t} \psi_{st}(x_s, x_t) \prod_s \psi_s(x_s) = \exp \left\{ J \sum_{s \sim t} x_s x_t + \sum_s b_s x_s \right\}.$$

- If $x_s = x_t$, i.e. the neighboring spins are the same, then the likelihood is larger. We will use this Ising model for image denoising in the Assignment 2!

Summary

- In certain cases, DAGMs cannot encode the proper conditional independences.
- MRFs are useful if there is no topological ordering in the graph (image).
- Cliques are key to parametrizing distributions induced by MRFs.
- Ising model is one useful MRF example.
- Next week: how to do exact inference?