Markov Random Fields

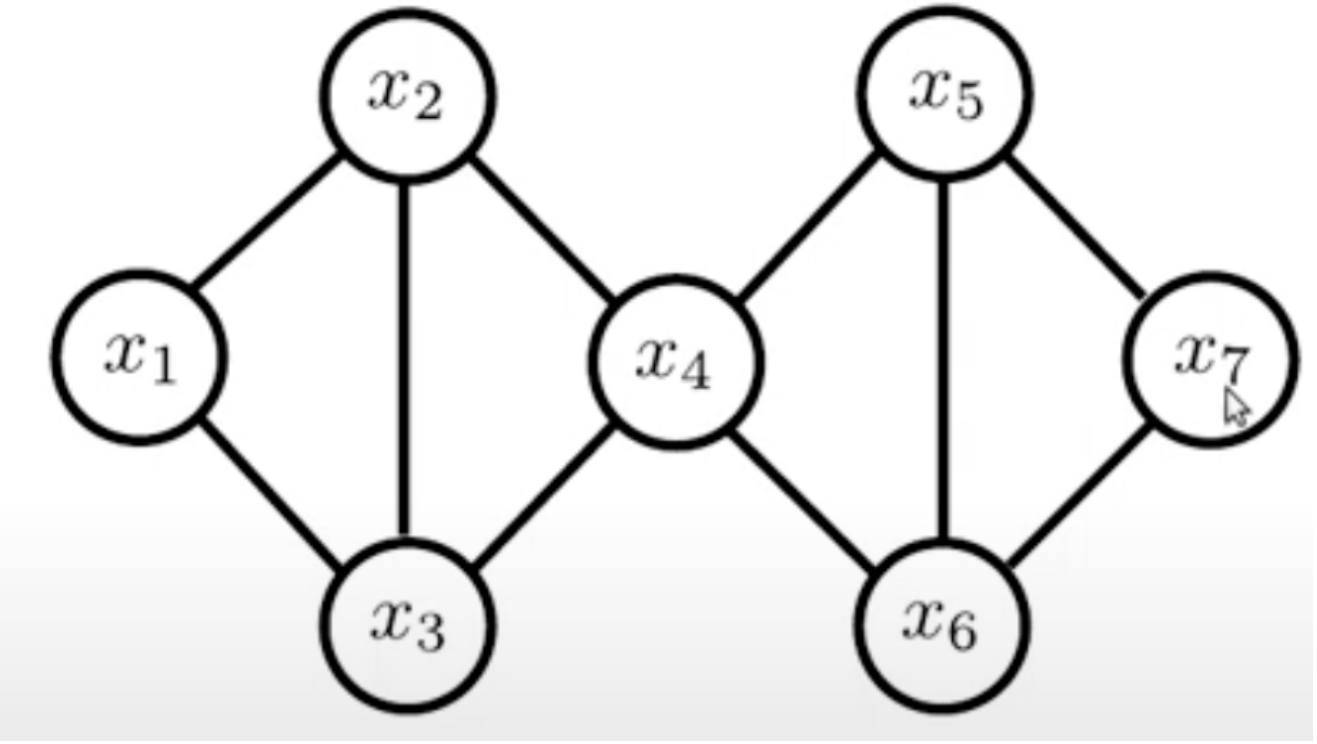
Global Markov Property

Separation

A subset S separates A from B if every path from a member of A to any member of B passes through S.

Conditional Independence (Global Markov Property)

For *disjoint* sets of variables (A, B, S) where S separates A from B, we have $A||B|S$ (A and B are conditionally independent given S).



Clique

A clique is a subset of vertices that are fully connected: (x_1, x_3), (x_2, x_4), (x_3, x_4), etc.
A maximum clique is a clique that cannot be extended by any other vertex: (x_1, x_2, x_3, x_4), (x_4, x_5, x_7, x_6), (x_4, x_5, x_6), (x_5, x_7, x_6)

Potential

A potential $\phi(x)$ is a non-negative function of the variable x. A **joint potential** $\phi(x_1, \ldots, x_D)$ is a non-negative function of the **set** of variables.

Local Markov Property

When conditioned on its neighbors, x becomes independent of the remaining variables of the graph:

$$p(x|X\setminus\{x\}) = p(x|ne(x))$$

The set of neighboring nodes $ne(x)$ is called **Markov blanket**.

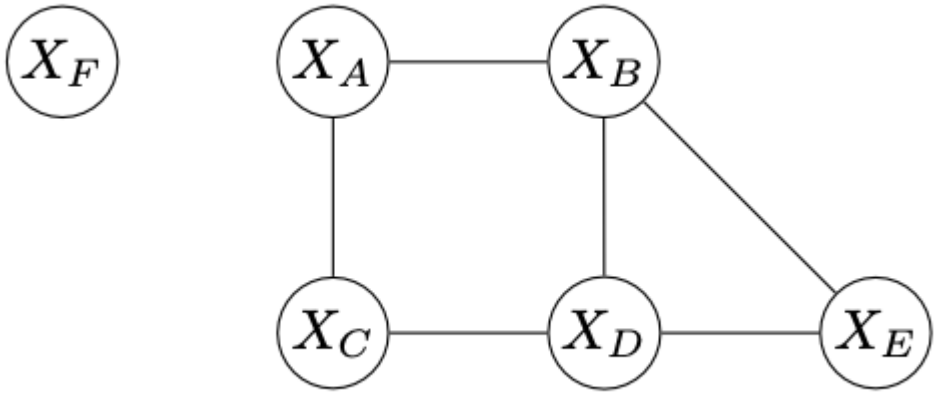
Markov Random Fields

For a set of variables $X = \{x_1, \dots, x_D\}$, a Markov Random Field is defined as a product of potentials over the **(maximal) cliques** $\{X_c\}_{c=1}^{C}$ of the undirected graph G:

$$p(X) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c(X_c)$$

MRFs are undirected graphical models. Hence, our model now consists of two parts:
the undirected graph G and so called (maximum) cliques potentials ϕ_A
A clique in an undirected graph G is a fully connected subset of notes. Hence, single nodes are also considered cliques.

A clique is **maximal** if there is no clique that strictly contains it, i.e. we cannot add another node to the clique which is fully connected to all others.
Suppose we are given the following graph:



Then, our maximum cliques are $\{X_A, X_C\}, \{X_A, X_B\}, \{X_C, X_D\}, \{X_B, X_D, X_E\}, \{X_F\}$
So, in this graph, $\{X_B, X_D\}$ is not a maximum clique because we **can** add X_E to it and all nodes in the new clique would still be connected to each other.

MRF Joint Distribution

$$p(x_1, \dots, x_N, z_1, \dots, z_N) = p(x_1, \dots, x_N|z_1, \dots, z_N)p(z_1, \dots, z_N)$$

where x_i and z_i are categorical variables.
MRF inference would seek the posterior of hidden variables:

$$p(z_1,\dots,z_N|x_1,\dots,x_N) = \frac{p(x_1,\dots,x_N|z_1,\dots,z_N) * p(z_1,\dots,z_N)}{\sum_{z_1}\dots\sum_{z_N} p(x_1,\dots,x_N|z_1,\dots,z_N) * p(z_1,\dots z_N)}$$

The evidence below is intractable to compute. So, how do we simplify it?

MRF Assumptions(Constraints)

1. Each observation only depends on a single hidden variable(same as HMM):

$$p(x_1,\dots,x_N|z_1,\dots,z_N) = \prod_{i=1}^{N} p(x_i|z_i)$$

2. a) Each hidden variable only depends on its neighbors: $z_i||z_j|N(z_i)$
   2.b) Neighborhood of a hidden variable z_i split in pairs such that the prior $p(z_1,\dots,z_N)$ decomposes into

$$p(z_1,\dots,z_N) = \prod_{z_i\in N(z_i)} p(z_i,z_j)$$

For example, the following undirected graph factorizes into this (2.a):

```
![[Screen Shot 2024-07-04 at 13.48.50.png]]
```

Assumption 2.b implies that we model the dependencies with cliques ≤ 2

Then, considering all these assumptions, our factorized model for inference simplifies into:

$$p(z_1,\dots,z_N|x_1,\dots,x_N) = \frac{\prod_{i=1}^{N} p(x_i|z_i)\prod_{z_i\in N(z_i)} p(z_i,z_j)}{\sum_{z_i}\dots\sum z_N \prod_{i=1}^{N} p(x_i|z_i)\prod_{z_i\in N(z_i)} p(z_i,z_j)}$$

The inference task for MRF is then to find the maximum likelihood:
$z* = argmaxp(z_1,\dots,z_N|x_1,\dots,x_N)$. This inference is tractable.
Learning for MRF includes: finding distributions $p(x_i|z_i)$ and $p(z_i,z_j)$. However, learning is **intractable** for most problems.
Instead of estimating $p(x_i|z_i)$ and $p(z_i,z_j)$, use potentials.

Potentials

We can write our joint probability distribution in terms of the maximum cliques $\{\phi_A\}$:

$$p(x_1,\dots,x_N) = \frac{1}{Z}\prod_{A}\phi_A(x_A)$$

Note that we now need a normalization constant Z which we did not need for Bayesian networks. The requirement for clique potentials is to be positive for any x_A, and are thus often modeled by an energy function $\phi_A(x_A) = exp(f(x_A))$
For the very first example we have, joint probability distribution can now be written as:

$$p(x_A, \dots, x_E) = \frac{1}{Z}\phi_{A,B}(x_A, x_B)\phi_{A,C}(x_A, x_C)\phi_{C,D}(x_C, x_D)\phi_{B,D,E}(x_B, x_D, x_E)\phi_F(x_F)$$

where $Z = \sum_{x_A}\sum_{x_B}\dots\sum_{x_F}\phi_{A,B}(x_A, x_B)\phi_{A,C}(x_A, x_C)\dots\phi_F(x_F)$
When maximizing the conditional likelihood, we don't care about Z variable

Hammersley-Clifford Theorem

A probability distribution that has a strictly positive mass or density satisfies the **Markov properties** with respect to an undirected graph G if and only if it is a Gibbs random field, i.e its density can be **factorized** over the (maximal) cliques of the graph.