Probabilistic Graphical Models

CMSC 691
UMBC
Two Problems for Graphical Models

\[ p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_C \psi_C(x_c) \]

Finding the normalizer  Computing the marginals
Two Problems for Graphical Models

\[ p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_c \psi_c(x_c) \]

Finding the normalizer \hspace{2cm} Computing the marginals

\[ Z = \sum_x \prod_c \psi_c(x_c) \]
Two Problems for Graphical Models

Finding the normalizer

$$\mathcal{P}(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_{c} \psi_c(x_c)$$

$$Z = \sum_{x} \prod_{c} \psi_c(x_c)$$

Computing the marginals

$$Z_n(v) = \sum_{x:x_n=v} \prod_{c} \psi_c(x_c)$$

Example: 3 variables, fix the 2\textsuperscript{nd} dimension

$$Z_2(v) = \sum_{x_1} \sum_{x_3} \prod_{c} \psi_c(x = (x_1, v, x_3))$$
Two Problems for Graphical Models

Finding the normalizer

\[ p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_C \psi_C(x_c) \]

\[ Z = \sum_x \prod_C \psi_C(x_c) \]

Q: Why are these difficult?
A: Many different combinations

Computing the marginals

\[ Z_n(v) = \sum_{x:x_n=v} \prod_C \psi_C(x_c) \]

Example: 3 variables, fix the 2nd dimension

\[ Z_2(v) = \sum_{x_1} \sum_{x_3} \prod_C \psi_C(x = (x_1, v, x_3)) \]
A graph $G$ that represents a probability distribution over random variables $X_1, \ldots, X_N$
Probabilistic Graphical Models

A graph $G$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Graph $G = (\text{vertices } V, \text{ edges } E)$

Distribution $p(X_1, \ldots, X_N)$
Probabilistic Graphical Models

A graph $G$ that represents a probability distribution over random variables $X_1, ..., X_N$

Graph $G = (\text{vertices } V, \text{ edges } E)$

Distribution $p(X_1, ..., X_N)$

Vertices $\leftrightarrow$ random variables

Edges show dependencies among random variables
Probabilistic Graphical Models

A graph $G$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Graph $G = (\text{vertices } V, \text{ edges } E)$
Distribution $p(X_1, \ldots, X_N)$

Vertices $\leftrightarrow$ random variables
Edges show dependencies among random variables

Two main flavors: \textit{directed} graphical models and \textit{undirected} graphical models
Outline

Directed Graphical Models

Undirected Graphical Models

Factor Graphs
Directed Graphical Models

A directed (acyclic) graph $G=(V,E)$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Joint probability factorizes into factors of $X_i$ conditioned on the parents of $X_i$
Directed Graphical Models

A directed (acyclic) graph $G=(V,E)$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Joint probability factorizes into factors of $X_i$ conditioned on the parents of $X_i$

Benefit: read the independence properties are transparent
Directed Graphical Models

A directed (acyclic) graph $G=(V,E)$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Joint probability factorizes into factors of $X_i$ conditioned on the parents of $X_i$

A graph/joint distribution that follows this is a **Bayesian network**
Bayesian Networks: Directed Acyclic Graphs

\[ p(x_1, x_2, x_3, \ldots, x_N) = \prod_{i} p(x_i \mid \pi(x_i)) \]

“parents of”
topological sort
Bayesian Networks: Directed Acyclic Graphs

\[ p(x_1, x_2, x_3, \ldots, x_N) = \prod_{i} p(x_i \mid \pi(x_i)) \]

\[ p(x_1, x_2, x_3, x_4, x_5) = ??? \]
Bayesian Networks: Directed Acyclic Graphs

\[ p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_3)p(x_2|x_1, x_3)p(x_4|x_2, x_3)p(x_5|x_2, x_4) \]
Bayesian Networks: Directed Acyclic Graphs

\[ p(x_1, x_2, x_3, \ldots, x_N) = \prod_{i} p(x_i \mid \pi(x_i)) \]

exact inference in general DAGs is NP-hard

inference in trees can be exact
Directed Graphical Model Notation

Unshaded nodes are unobserved (latent) R.V.s

Shaded nodes are observed R.V.s
Variables $X$ & $Y$ are conditionally independent given $Z$ if all (undirected) paths from (any variable in) $X$ to (any variable in) $Y$ are d-separated by $Z$.

X & Y are d-separated if for all paths $P$, one of the following is true:

1. $P$ has a chain with an observed middle node.
   
   ![Chain Diagram]

2. $P$ has a fork with an observed parent node.
   
   ![Fork Diagram]

3. $P$ includes a “v-structure” or “collider” with all unobserved descendants.
   
   ![Collider Diagram]
D-Separation: Testing for Conditional Independence

Variables $X$ & $Y$ are conditionally independent given $Z$ if all (undirected) paths from (any variable in) $X$ to (any variable in) $Y$ are $d$-separated by $Z$

- **$d$-separation**
  - $X$ & $Y$ are $d$-separated if for all paths $P$, one of the following is true:
    - **observing $Z$ blocks the path from $X$ to $Y$**
    - **not observing $Z$ blocks the path from $X$ to $Y$**

**Examples:****
- **P has a chain with an observed middle node**
  - $X \rightarrow Z \rightarrow Y$
  - Observing $Z$ blocks the path from $X$ to $Y$

- **P has a fork with an observed parent node**
  - $Z$ is observed as a parent of both $X$ and $Y$

- **P includes a “v-structure” or “collider” with all unobserved descendants**
  - $X \rightarrow Z \leftarrow Y$
D-Separation: Testing for Conditional Independence

Variables X & Y are conditionally independent given Z if all (undirected) paths from (any variable in) X to (any variable in) Y are \textit{d-separated} by Z.

\begin{align*}
\text{observing Z blocks} & \text{ the path from X to Y} \\
\text{not observing Z blocks} & \text{ the path from X to Y}
\end{align*}

\[ p(x, y, z) = p(x)p(y)p(z|x, y) \]
\[ p(x, y) = \sum_z p(x)p(y)p(z|x, y) = p(x)p(y) \]
Markov Blanket

the set of nodes needed to form the complete conditional for a variable $x_i$

$$p(x_i | x_j \neq i) = \frac{p(x_1, ..., x_N)}{\int p(x_1, ..., x_N) dx_i}$$

factorization of graph

$$= \frac{\prod_k p(x_k | \pi(x_k))}{\int \prod_k p(x_k | \pi(x_k)) dx_i}$$

factor out terms not dependent on $x_i$

$$= \frac{\prod_{k: k = i \text{ or } i \in \pi(x_k)} p(x_k | \pi(x_k))}{\int \prod_{k: k = i \text{ or } i \in \pi(x_k)} p(x_k | \pi(x_k)) dx_i}$$

Markov blanket of a node $x$ is its parents, children, and children's parents

(in this example, shading does not show observed/latent)
Outline

Directed Graphical Models

Undirected Graphical Models

Factor Graphs
Undirected Graphical Models

An *undirected* graph $G=(V,E)$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Joint probability factorizes based on cliques in the graph
Undirected Graphical Models

An undirected graph $G=(V,E)$ that represents a probability distribution over random variables $X_1, \ldots, X_N$

Joint probability factorizes based on cliques in the graph

Common name: Markov Random Fields
Undirected Graphical Models

An *undirected* graph $G=(V,E)$ that represents a probability distribution over random variables $X_1, ..., X_N$

Joint probability factorizes based on cliques in the graph

Common name: **Markov Random Fields**

Undirected graphs can have an alternative formulation as **Factor Graphs**
Markov Random Fields: Undirected Graphs

\[ p(x_1, x_2, x_3, \ldots, x_N) \]
Markov Random Fields: Undirected Graphs

**clique**: subset of nodes, where nodes are pairwise connected

**maximal clique**: a clique that cannot add a node and remain a clique

\[ p(x_1, x_2, x_3, \ldots, x_N) \]
Markov Random Fields: Undirected Graphs

**clique**: subset of nodes, where nodes are pairwise connected.

**maximal clique**: a clique that cannot add a node and remain a clique.

\[
p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_C \psi_C(x_C)
\]

- global normalization
- maximal cliques
- potential function (not necessarily a probability!)
- variables part of the clique C
Markov Random Fields: Undirected Graphs

**clique**: subset of nodes, where nodes are pairwise connected

**maximal clique**: a clique that cannot add a node and remain a clique

\[
p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_c \psi_C(x_c)
\]

- global normalization
- maximal cliques
- potential function (not necessarily a probability!)

variables part of the clique C
Markov Random Fields: Undirected Graphs

**clique**: subset of nodes, where nodes are pairwise connected

**maximal clique**: a clique that cannot add a node and remain a clique

$p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_C \psi_C(x_c)$

Q: What restrictions should we place on the potentials $\psi_C$?
Markov Random Fields: Undirected Graphs

**clique**: subset of nodes, where nodes are pairwise connected

**maximal clique**: a clique that cannot add a node and remain a clique

\[
p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_{C} \psi_C(x_C)
\]

Q: What restrictions should we place on the potentials \( \psi_C \)?

A: \( \psi_C \geq 0 \) (or \( \psi_C > 0 \))
Terminology: Potential Functions

\[ p(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{Z} \prod_C \psi_C(x_C) \]

- energy function (for clique C)

\[ \psi_C(x_C) = \exp \left( -E(x_C) \right) \]

Boltzmann distribution

(get the total energy of a configuration by summing the individual energy functions)
Ambiguity in Undirected Model Notation

$p(x, y, z) \propto \psi(x, y, z)$

$p(x, y, z) \propto \psi_1(x, y)\psi_2(y, z)\psi_3(x, z)$
Outline

Directed Graphical Models

Undirected Graphical Models

Factor Graphs
MRFs as Factor Graphs

Undirected graphs: \( G=(V,E) \) that represents \( p(X_1, \ldots, X_N) \)

Factor graph of \( p \): Bipartite graph of evidence nodes \( X \), factor nodes \( F \), and edges \( T \)

Evidence nodes \( X \) are the random variables

Factor nodes \( F \) take values associated with the potential functions

Edges show what variables are used in which factors
MRFs as Factor Graphs

Undirected graphs: 
\( G = (V, E) \) that represents 
\( p(X_1, ..., X_N) \)

Factor graph of \( p \): 
Bipartite graph of evidence nodes \( X \), factor nodes \( F \), and edges \( T \)
MRFs as Factor Graphs

Undirected graphs: $G= (V,E)$ that represents $p(X_1, ..., X_N)$

Factor graph of $p$: Bipartite graph of evidence nodes $X$, factor nodes $F$, and edges $T$

Evidence nodes $X$ are the random variables
MRFs as Factor Graphs

Undirected graphs: $G=(V,E)$ that represents $p(X_1, \ldots, X_N)$

Factor graph of $p$: Bipartite graph of evidence nodes $X$, factor nodes $F$, and edges $T$

**Evidence nodes $X$** are the random variables

**Factor nodes $F$** take values associated with the potential functions
MRFs as Factor Graphs

Undirected graphs: $G = (V, E)$ that represents $p(X_1, \ldots, X_N)$

Factor graph of $p$: Bipartite graph of evidence nodes $X$, factor nodes $F$, and edges $T$

**Evidence nodes** $X$ are the random variables

**Factor nodes** $F$ take values associated with the *potential functions*

**Edges** show what variables are used in which factors
Different Factor Graph Notation for the Same Graph

- Triangle graph
- Undirected connections

- Directed graph
- Directed connections

- Directed graph
- Square connections

---
Directed vs. Undirected Models: Moralization

\[ x_1 \rightarrow x_2 \rightarrow x_4 \]
\[ x_3 \rightarrow x_4 \]
Directed vs. Undirected Models: Moralization

\[ p(x_1, ..., x_4) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \]
Directed vs. Undirected Models: Moralization

\[ p(x_1, ..., x_4) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \]

parents of nodes in a directed graph must be connected in an undirected graph
Example: Linear Chain

Directed (e.g., hidden Markov model [HMM]; generative)
Example: Linear Chain

Directed (e.g., hidden Markov model [HMM]; generative)

Directed (e.g., maximum entropy Markov model [MEMM]; conditional)
Example: Linear Chain

**Directed** (e.g., hidden Markov model [HMM]; generative)

$z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$w_1 \quad w_2 \quad w_3 \quad w_4$

**Directed** (e.g., maximum entropy Markov model [MEMM]; conditional)

$z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$w_1 \quad w_2 \quad w_3 \quad w_4$

**Undirected** (e.g., conditional random field [CRF])

$z_1 \leftrightarrow z_2 \leftrightarrow z_3 \leftrightarrow z_4$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$w_1 \quad w_2 \quad w_3 \quad w_4$
Example: Linear Chain

Directed (e.g., hidden Markov model [HMM]; generative)

Undirected as factor graph (e.g., conditional random field [CRF])

Directed (e.g., maximum entropy Markov model [MEMM]; conditional)
Example: Linear Chain Conditional Random Field

Widely used in applications like part-of-speech tagging

President Obama told Congress ...
Example: Linear Chain Conditional Random Field

Widely used in applications like part-of-speech tagging:

President Obama told Congress ...

and named entity recognition:

President Obama told Congress ...
Linear Chain CRFs for Part of Speech Tagging

A linear chain CRF is a conditional probabilistic model of the sequence of tags $z_1, z_2, \ldots, z_N$ conditioned on the *entire* input sequence $x_{1:N}$.
Linear Chain CRFs for Part of Speech Tagging

\[ p(\clubsuit | \diamondsuit) \]

A linear chain CRF is a **conditional probabilistic model** of the sequence of tags \( z_1, z_2, ..., z_N \) conditioned on the *entire* input sequence \( x_{1:N} \).
A linear chain CRF is a conditional probabilistic model of the sequence of tags $z_1, z_2, ..., z_N$ conditioned on the entire input sequence $x_{1:N}$.
A linear chain CRF is a conditional probabilistic model of the sequence of tags $z_1, z_2, ..., z_N$ conditioned on the entire input sequence $x_{1:N}$.
Linear Chain CRFs for Part of Speech Tagging

\[ p(z_1, z_2, \ldots, z_N \mid x_{1:N}) \]
Linear Chain CRFs for Part of Speech Tagging

\[
p(z_1, z_2, \ldots, z_N | x_{1:N}) \propto \prod_{i=1}^{N} \exp(\langle \theta^{(f)}, f_i(z_i) \rangle + \langle \theta^{(g)}, g_i(z_i, z_{i+1}) \rangle)
\]
Linear Chain CRFs for Part of Speech Tagging

\( g_j \): inter-tag features
(can depend on any/all input words \( x_{1:N} \))
Linear Chain CRFs for Part of Speech Tagging

$g_j$: inter-tag features (can depend on any/all input words $x_{1:N}$)

$\mathbf{f}_i$: solo tag features (can depend on any/all input words $x_{1:N}$)
Linear Chain CRFs for Part of Speech Tagging

\[ g_j: \text{inter-tag features} \]
\[ \text{(can depend on any/all input words} \ x_{1:N} \text{)} \]

\[ f_i: \text{solo tag features} \]
\[ \text{(can depend on any/all input words} \ x_{1:N} \text{)} \]

Feature design, just like in maxent models!
Linear Chain CRFs for Part of Speech Tagging

\( g_j \): inter-tag features
(can depend on any/all input words \( x_{1:N} \))

\( f_i \): solo tag features
(can depend on any/all input words \( x_{1:N} \))

Example:

\[
g_{j,N\rightarrow V}(z_j, z_{j+1}) = 1 \text{ (if } z_j = N \text{ & } z_{j+1} = V) \text{ else 0}
g_{j,told,N\rightarrow V}(z_j, z_{j+1}) = 1 \text{ (if } z_j = N \text{ & } z_{j+1} = V \text{ & } x_j = told) \text{ else 0}
\]
Outline

Directed Graphical Models

Undirected Graphical Models

Factor Graphs