David Barber’s *Bayesian Reasoning and Machine Learning*:

- Introductory machine learning material, as well as graphical models
- Primary source of readings for first half of course (free online)
Probabilistic Graphical Models

Directed Graphical Model (Bayesian Network)

Factor Graph (Hypergraph)

Undirected Graphical Model (Markov Random Field)

Markov Properties

A more detailed view of factorization, useful for algorithms (next lecture)

Markov Properties

Factorization
Undirected Graphical Models: Markov Properties & Factorization
Undirected Graphical Models

Undirected Graphs

\( \mathcal{V} \rightarrow \) Set of \( N \) nodes or vertices, \( \{1, 2, \ldots, N\} \)

\( \mathcal{E} \rightarrow \) Set of undirected edges \((s,t)\), or equivalently \((t,s)\), linking pairs of nodes. The neighbors of a node are

\[ \Gamma(t) = \{ s \in \mathcal{V} \mid (s,t) \in \mathcal{E} \} \]

\( \mathcal{C} \rightarrow \) Set of cliques (fully connected subsets) of nodes. A maximal clique is not a strict subset of any other clique.

Undirected Graphical Model, or Markov Random Field (MRF)

\( x_s \rightarrow \) Random variable (continuous or discrete) associated with node \( s \).

Notation for variable groups: \( x_A = \{ x_s \mid s \in A \} \) for any \( A \subseteq \mathcal{V} \)
Set B separates A from C if all paths from A to C pass through B.

By definition, distribution is Markov if and only if for any B separating A and C:

\[ p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B) \]

\[ p(x_A \mid x_B, x_C) = p(x_A \mid x_B) \]

\[ p(x_C \mid x_B, x_A) = p(x_C \mid x_B) \]
Global & Local Markov Properties

Local Markov Property
• Given its neighbors, each node is independent of all other variables

\[ p(x \mid x_{V \setminus s}) = p(x \mid x_{\Gamma(s)}) \]

\[ \Gamma(s) = \{ t \in V \mid (s, t) \in E \} \]

• This local Markov property is a special case of the global Markov property

Global Markov Property
• Set B separates A from C if all paths from A to C pass through B
• By definition, distribution is Markov if and only if for any B separating A and C:

\[ p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B) \]

\[ p(x_A \mid x_B, x_C) = p(x_A \mid x_B) \quad p(x_C \mid x_B, x_A) = p(x_C \mid x_B) \]
Pairwise MRFs are Always Markov

**Random variable associated with node s.**

**Set of N nodes or vertices, \{1, 2, \ldots, N\}**

**Set of undirected edges (s,t) linking pairs of nodes.**

**Pairwise Markov Random Field:**

\[
p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t)
\]

\[
Z = \sum_{x} \prod_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t)
\]

\[
\phi_{st}(x_s, x_t) \geq 0 \quad \text{Arbitrary non-negative potential function.}
\]

\[
Z > 0 \quad \text{Normalization constant, or partition function. Defined as sum over all possible joint states.}
\]
Example: Nearest-Neighbor Spatial MRF

- Observed nodes: Features of 2D image (intensity, color, texture, …)
- Hidden nodes: Property of 3D world (depth, motion, object category, …)
Undirected MRFs are Always Markov

$\mathcal{X}_s \rightarrow$ Random variable associated with node $s$.

$\mathcal{V} \rightarrow$ Set of $N$ nodes or vertices, $\{1, 2, \ldots, N\}$

$\mathcal{E} \rightarrow$ Set of undirected edges $(s,t)$ linking pairs of nodes.

$\mathcal{C} \rightarrow$ Set of cliques (fully connected subsets) of nodes.

**Undirected Markov Random Field:**

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

$$Z = \sum_{x} \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

$\phi_c(x_c) \geq 0 \rightarrow$ Arbitrary non-negative potential function.

$Z > 0 \rightarrow$ Normalization constant, or partition function.

Defined as sum over all possible joint states.
Examples: Undirected Factorization

\[ p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c) \]

\[ Z = \sum_{x} \prod_{c \in \mathcal{C}} \phi_c(x_c) \]

For this graph, the maximal cliques are pairs of nodes, and undirected MRF becomes a pairwise MRF

Undirected graphs can represent arbitrary joint distributions, but in such cases the graph structure may be useless

Dependence order may vary across graph
Transformations of Undirected Models

\[ p(A, B, C) = \phi_{AC}(A, C')\phi_{BC}(B, C')/Z \]

Marginalising over \( C \) makes \( A \) and \( B \) (graphically) dependent.

\[ p(A, B | C) = p(A | C)p(B | C) \]

Conditioning on \( C \) makes \( A \) and \( B \) independent:
Why Does Factorization Imply Markov?

Denoting variables by integers for compact formulas:

\[ p(1, 2, 3, 4, 5, 6, 7) = \frac{1}{Z} \phi(1, 2, 3)\phi(2, 3, 4)\phi(4, 5, 6)\phi(5, 6, 7) \]

\[ p(1, 2, 3, 4, 5, 6, 7) \propto \phi(1, 2, 3)\phi(2, 3, 4)\phi(4, 5, 6)\phi(5, 6, 7) \]

*Normalization constant irrelevant for testing factorization*
Why Does Factorization Imply Markov?

Denoting variables by integers for compact formulas:

\[ p(1, 2, 3, 4, 5, 6, 7) = \frac{1}{Z} \phi(1, 2, 3)\phi(2, 3, 4)\phi(4, 5, 6)\phi(5, 6, 7) \]

\[ p(1, 2, 3, 4, 5, 6, 7) \propto \phi(1, 2, 3)\phi(2, 3, 4)\phi(4, 5, 6)\phi(5, 6, 7) \]

\[ p(1, 7|4) \propto \sum_{2,3,5,6} p(1, 2, 3, 4, 5, 6, 7) \]

\[ \propto \sum_{2,3,5,6} \phi(1, 2, 3)\phi(2, 3, 4)\phi(4, 5, 6)\phi(5, 6, 7) \]

\[ = \left\{ \sum_{2,3} \phi(1, 2, 3)\phi(2, 3, 4) \right\} \left\{ \sum_{5,6} \phi(4, 5, 6)\phi(5, 6, 7) \right\} \]

\[ \propto p(1|4)p(7|4) \]
Why Does Factorization Imply Markov?

\[ p(1, 2, 3, 4, 5, 6, 7) \propto \phi(1, 2, 3)\phi(2, 3, 4)\phi(4, 5, 6)\phi(5, 6, 7) \]

\[ p(1, 7|4) \propto \left\{ \sum_{2,3} \phi(1, 2, 3)\phi(2, 3, 4) \right\} \left\{ \sum_{5,6} \phi(4, 5, 6)\phi(5, 6, 7) \right\} \propto p(1|4)p(7|4) \]

General Undirected Markov Random Fields:

\[ p(x) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c) \]

If B separates A from C, we have:

- Some potentials may depend on \( x_A, x_B \)
- Some potentials may depend on \( x_B, x_C \)
- But there are no potentials linking \( x_A, x_C \)
Hammersley-Clifford Theorem

\[ p(x) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c) \quad Z = \sum_x \prod_{c \in C} \phi_c(x_c) \]

Hammersley-Clifford Theorem (1971)

- A distribution defined as a normalized product of non-negative clique potential functions is always Markov with respect to the corresponding graph.
- If a strictly positive distribution, for which \( p(x) > 0 \) for all \( x \), is Markov then it can always be factorized into corresponding clique potentials.
- This factorization into clique potentials is not unique.
- There do exist (strange) degenerate Markov distributions which do not factorize, but theorem guarantees that if we learn factorized potentials, they will be Markov.
Directed Graphical Models: Factorization & Markov Properties

This model has 37 variables and 504 parameters, created by hand using knowledge elicitation (probabilistic expert system).
Directed Graphical Models

Chain rule implies that any joint distribution equals:

\[ p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1) \cdots p(x_N \mid x_{N-1}, \ldots, x_1) \]

Directed graphical model implies a restricted factorization:

\[
p(x) = \prod_{s=1}^{N} p(x_s \mid x_{\Gamma(s)})
\]

\( \Gamma(s) \rightarrow \) set of parents of node \( s \), possibly empty

Valid for any directed acyclic graph (DAG):

equivalent to dropping conditional dependencies in chain rule

\[
p(x_{1:5}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3)p(x_5 \mid x_1, x_2, x_3, x_4)
= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2, x_3)p(x_5 \mid x_3)
\]
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Choosing an ordering
Without loss of generality, we can write

\[ p(A, R, E, B) = p(A|R, E, B)p(R, E, B) \]
\[ = p(A|R, E, B)p(R|E, B)p(E, B) \]
\[ = p(A|R, E, B)p(R|E, B)p(E|B)p(B) \]

Assumptions:
- The alarm is not directly influenced by any report on the radio,
  \( p(A|R, E, B) = p(A|E, B) \)
- The radio broadcast is not directly influenced by the burglar variable,
  \( p(R|E, B) = p(R|E) \)
- Burglaries don’t directly ‘cause’ earthquakes, \( p(E|B) = p(E) \)

Therefore

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]

\[ p(B = 1) = 0.01 \]

\[ p(E = 1) = 0.000001 \]

Why don’t we all use expert systems? Accurate assessment of such probabilities can be hard.
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]

\[ p(B = 1) = 0.01 \]

\[ p(A|B, E) \]

\[ p(E = 1) = 0.000001 \]

\[ p(R|E) \]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Radio = 1</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Inference: Probability of Burglary given Alarm

\[ p(B = 1|A = 1) = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)} \]

\[ = \frac{\sum_{E,R} p(A = 1|B = 1, E)p(B = 1)p(E)p(R|E)}{\sum_{B,E,R} p(A = 1|B, E)p(B)p(E)p(R|E)} \approx 0.99 \]
Example: The Alarm Network

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[ p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B) \]

\[ p(B = 1) = 0.01 \quad p(E = 1) = 0.000001 \]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Radio broadcast is not directly influenced by the Burglar variable.

The radio broadcast is not directly influenced by the Burglar variable.

Additional Evidence: The radio broadcasts an earthquake warning:

\[ p(R = 1) \]

Initial Evidence: The alarm is sounding

\[ p(A = 1) \]

Assumptions:

Without loss of generality, we can write

Choosing an ordering

Earthquake 'explains away' to an extent the fact that the alarm is ringing.

Therefore

Burglaries don't directly 'cause' earthquakes,

This probability drops dramatically when she hears that there has been an earthquake.

A similar calculation gives

\[ p(B = 0 | A = 1, R = 1) \approx 0.99 \]

Inference: Probability of Burglary given Alarm and Earthquake

\[ p(B = 1 | A = 1, R = 1) \approx 0.01 \]

“Explaining Away”
Real Alarm Networks: ICU Monitoring

Beinlich et al., 1989

Aleks, Russell, et al., 2008
Conditional Independence in Directed Models

All belief networks with three nodes and two links:

\[ A \perp B \mid C \]

\[ A \not\perp B \mid C \]

In (a), (b) and (c), \( A, B \) are conditionally independent given \( C \).

(a) \[ p(A, B \mid C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A \mid C)p(B \mid C)p(C)}{p(C)} = p(A \mid C)p(B \mid C) \]

(b) \[ p(A, B \mid C) = \frac{p(A)p(C \mid A)p(B \mid C)}{p(C)} = \frac{p(A, C)p(B \mid C)}{p(C)} = p(A \mid C)p(B \mid C) \]

(c) \[ p(A, B \mid C) = \frac{p(A \mid C)p(C \mid B)p(B)}{p(C)} = \frac{p(A \mid C)p(B \mid C)}{p(C)} = p(A \mid C)p(B \mid C) \]

In (d) the variables \( A, B \) are conditionally dependent given \( C \),
\[ p(A, B \mid C) \propto p(C \mid A, B)p(A)p(B). \]

Marginally independent but conditionally dependent!
In (a), (b) and (c), the variables $A, B$ are marginally dependent.

In (d) the variables $A, B$ are marginally independent.

$$p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B)$$
A collider contains two or more incoming arrows along a chosen path.

Summary of two previous slides:

If $C$ has more than one incoming link, then $A \perp B$ and $A \not\perp B \mid C$. In this case $C$ is called **collider**.

If $C$ has at most one incoming link, then $A \perp B \mid C$ and $A \not\perp B$. In this case $C$ is called **non-collider**.
The Connection Graph

All paths in the connection graph need to be blocked to obtain $A \perp D \mid B, C$:

- Non-collider in the conditioning set blocks a path
- Collider outside the conditioning set blocks a path

$B \perp C \mid A$ and $B \nLeftarrow C \mid A$
Deducing Directed Independence

Given three sets of nodes $\mathcal{X}, \mathcal{Y}, \mathcal{C}$, if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are blocked by $\mathcal{C}$, then $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{C}$.

A path $\mathcal{P}$ is blocked by $\mathcal{C}$ if at least one of the following conditions is satisfied:

1. there is a collider $\bullet$ in the path $\mathcal{P}$ such that neither the collider nor any of its descendants is in the conditioning set $\mathcal{C}$.

2. there is a non-collider in the path $\mathcal{P}$ that is in the conditioning set $\mathcal{C}$.

$d$-connected/separated
We use the phrase ‘$d$-connected’ if there is a path from $\mathcal{X}$ to $\mathcal{Y}$ in the ‘connection’ graph – otherwise the variable sets are ‘$d$-separated’. Note that $d$-separation implies that $\mathcal{X} \perp \mathcal{Y} | \mathcal{Z}$, but $d$-connection does not necessarily imply conditional dependence.
Markov: Directed versus Undirected

\[
\begin{align*}
X & \perp Y \mid \{W, Z\} \\
W & \perp Z \mid \{X, Y\}
\end{align*}
\]

Can represent one, but not both simultaneously, of these conditional independencies in a single directed model.

Graph separation implies that we cannot represent unconditional independence, but conditional dependence, in an undirected model.