COMP 790-125: Goals for today

- Graphical Models
- Bayes Nets
- Conditional independence and d-separation
- Markov Nets
- Markov Blankets

For reference, Koller and Friedman PGM book.
Or, a chapter from another book
Graphical models – motivation

Graphical models enable compact representation of probability distributions.

Without any assumptions representing a distribution over \( n \) binary variables
\[
p(X_1, \ldots, X_n)
\]
requires room for \( 2^n - 1 \) real values.

However, if we were to factorize the distribution
\[
p(X_1, \ldots, X_n) = p(X_1)p(X_2|X_1)p(X_3|X_2)\ldots p(X_n|X_{n-1})
\]
the space requirements go down to \( n \) real values.

Graphical models capture assumptions which such factorization possible.
Conditional independence

Marginal independence you are familiar with

\[ p(X|Y, Z) = p(X) \]
\[ p(X, Y) = p(X)p(Y) \]

Conditional independence

\[ p(X|Y, Z) = p(X|Z) \]
\[ p(X, Y|Z) = p(X|Z)p(Y|Z) \]

For some random variables \( X, Y \) and \( Z \) with joint \( p(X, Y, Z) \) we say that \( X \) and \( Y \) are independent given \( Z \) if

\[ p(X = x|Z = z) = p(X = x|Y = y, Z = z) \]

for all \( x \in \text{dom } X, y \in \text{dom } Y, z \in \text{dom } Z \) such that \( p(z) > 0 \).

And shorthand for this relationship is

\[ X \perp Y|Z \]

Note that the relationship is symmetric.
Conditional independence between sets of variables

This generalizes to sets of variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$

$$\mathcal{X} \perp \mathcal{Y} | \mathcal{Z} \iff \forall X \in \mathcal{X}, Y \in \mathcal{Y} \quad X \perp Y | \mathcal{Z}$$

We can then also say that the distribution factors

$$p(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z})p(\mathcal{Y} | \mathcal{Z})p(\mathcal{Z})$$

We will also use notation $X_A$ where $A$ is a set of indices to denote a set of random variables, e.g. $A = \{1, 2, 5\}$ then $X_A = \{X_1, X_2, X_5\}$. 
Conditional independence properties

Symmetry

\[ X \perp Y | Z \implies Y \perp X | Z \]

Decomposition

\[ X \perp Y, W | Z \implies X \perp Y | Z \]

Weak union

\[ X \perp Y, W | Z \implies X \perp Y | Z, W \]

Contraction

\[ (X \perp Y | Z) \text{ and } (X \perp W | Z, Y) \implies X \perp Y, W | Z \]

Intersection

For positive distribution \( p(x, y, z, w) > 0, \forall x, y, z, w \):

\[ (X \perp Y | Z, W) \text{ and } (X \perp W | Z, Y) \implies X \perp Y, W | Z \]
Why do conditional independencies matter in practice?

We use models to make predictions

$$p(Y|X_1 = x_1, X_2 = x_2)$$

In this setting, we condition on the observed values $x_1, x_2$ to predict the hidden/unknown $Y$.

Hence, if $Y$ is independent of features $X_2$, we may make predictions without collecting that data.

There are other, algorithmic, consequences we will discuss later.
Naive Bayes model

Assume as set of random variables $X_1, \ldots, X_n$ and a categorical variable $C$. Naive Bayes assumes conditional independencies

$X_i \perp X_{[-i]} | C, \ i = 1, \ldots, n$

where $X_{[-i]} = \{X_j : j \neq i\}$.

In words, features are independent from each other given class.
Naive Bayes model

Given

\[ X_i \perp X_{[-i]} \mid C, \ i = 1, \ldots, n \]

We can factorize

\[ p(X_1, \ldots, X_n \mid C) = p(X_1 \mid C)p(X_2 \mid C) \ldots p(X_n \mid C) \]

Practical consequence is that we can model each feature separately.
A thought experiment: Naive Bayes failure mode

Naive Bayes model can exhibit an undesired effect if the conditional independence assumptions are violated.

Let's consider posterior you would use to predict class of a new sample:

$$p(C = c|X_1, \ldots, X_n) = \frac{p(C = c) \prod_i p(X_i|C = c)}{\sum_k p(C = k) \prod_i p(X_i|C = k)}$$

For simplicity, we assume uniform prior on classes $p(C) = \text{const.}$ and compare

- a model that uses a single feature $X_1$
- Naive Bayes model on $r$ copies of that feature $X_1 = X_2 = X_3 \ldots = X_r$

What happens to the posterior as $r$ grows larger?
Bayesian Networks offer a way to represent distributions.

A Directed Acyclic Graph (DAG) provides visual representation for the distribution.

A node in the graph corresponds to a random variable.

A directed edge indicates influence between random variables. Source node is called a parent and target a child.
Edges, paths

We will denote oriented edge by $X \to Y$ or $Y \leftarrow X$.

Un-oriented edge will be denoted by $X \rightarrow Y$.

A **path** is a sequence of nodes $X_1, \ldots, X_k$ such $X_i \rightarrow X_{i+1}$ or $X_i \rightarrow X_{i+1}$, for $i < k$.

For simplicity, we will assume directed path is a sequence of nodes $X_1, \ldots, X_k$ such that graph contains edges $X_i \rightarrow X_{i+1}$, for $i < k$. 
Parents of a node $X_i$, set of nodes $X_j$ such that graph contains edge $X_j \rightarrow X_i$, is denoted by $\text{Pa}_{X_i}$.

**Descendants** of a node are nodes that can be reached using a directed path starting at the node.

**Ancestors** of a node are nodes from which the node can be reached using a directed path.

**Non-descendants** are all nodes that are not descendants.
Bayesian Network

We can specify conditional probabilities – local models –

\[ p(X_i|\text{Pa}_{X_i}) \]

The joint distribution is then given by

\[ p(X_1, \ldots, X_n) = \prod_i p(X_i|\text{Pa}_{X_i}) \]

BN above:

\[ p(P, T, I, X, S) = p(P)p(T)p(I|P, T)p(X|I)p(S|T) \]
Bayesian network – conditional independencies

Graph specifies immediate dependencies.
Given a graph, conditional independencies

\[ X_i \perp \text{NonDescendants}_{X_i} | \text{Pa}_{X_i} \]

are called **local Markov assumptions**.
In words, variables in a Bayes nets are independent of their non-descendants given parents.

Think of a tree structured graph. How would you translate the above statement? What are descendants of a node in a tree?
Given random variables $X, Y, Z$ we can consider four different dependency structures.

**Figure 2.2** (a) An indirect causal effect; (b) an indirect evidential effect; (c) a common cause; (d) a common effect.
A graph might specify a dependence between variables: $X \rightarrow Y$ but the variables may still be independent. Explicitly

$$p(X|Y) = p(X)$$

is still a valid conditional probability.

A Bayes Net specifies independencies that **must** hold, but additional dependencies may hold as well.

We turn to question of which independencies, in addition to the local Markov assumptions, must hold.
Complete set of independence assumptions – d-separation

Given a graph on three nodes $X, Y, Z$ we can ask when can influence flow between variables.

We will call path active if the start and end nodes are not independent.

- **Causal Path** $X \rightarrow Z \rightarrow Y$ is active iff $Z$ is not given
- **Evidential Path** $X \leftarrow Z \leftarrow Y$ is active iff $Z$ is not given
- **Common cause** $X \leftarrow Z \rightarrow Y$ is active iff $Z$ is not given
- **Common effect** $X \rightarrow Z \leftarrow Y$ is active iff $Z$ or its descendants are given

Common effect structure is also called **v-structure**.
Complete set of independence assumptions – d-separation

A path $X_1, \ldots, X_n$ is active given $Z$ if

- for each v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, $X_i$ or one of its descendants is in $Z$.
- no other $X_i$ along the path is in $Z$.

For sets of nodes $X, Y, Z$, we say $X$ and $Y$ are d-separated given $Z$, if there is no active path between any pair of nodes $X \in X$ and $Y \in Y$. 
Soundness and completeness

We say that a distribution factorizes according to Bayes Net if we can factorize joint according to the graph structure.

\[ p(X_1, ..., X_n) = \prod_i p(X_i | \text{Pa}_{X_i}). \]

**Soundness:** If a distribution factorizes according to a graph then independencies reported by d-separation are conditional dependencies of the distribution.

**Completeness:** Given a graph, if \( X \) and \( Y \) are not d-separated given \( Z \), then \( X \) and \( Y \) are dependent in some distribution \( P \) that factorizes according to the graph.

For proof, see the Koller and Friedman PGM book.
Undirected models

So far we looked at the graphical models specified in terms of a DAG (directed acyclic graph) and conditional probabilities. These were Bayes(ian belief) net(works).

We will come back to them, but we also want to familiarize ourselves with other common representations. Two types of undirected graphical models we will consider today:

- Markov random fields (markov nets)
- Factor graphs
The building blocks for Bayes Nets were conditional probabilities.

For MRFs the building blocks are potentials, e.g.

\[ \phi_i(X_{C_i}). \]

where potential operates on a subset of variables whose indices are in set \( C_i \).

Sets \( C_i \) correspond to **cliques** of the graph.

The set of random variables \( X_{C_i} \) is sometimes called scope of potential \( \phi_i \) (paralleling function scope).
**Gibbs distribution**: The full joint probability is then given as

\[ p(X) = \frac{1}{Z} \prod_i \phi_i(X_{C_i}). \]

Note that the subsets \( C_i \) for different \( i \) are not required to be disjoint.

The constant \( Z \) is called the partition function or normalization constant.

Just to tie things together a little bit suppose we had \( d \) potentials

\[ \phi_i(X_i) = \exp \left\{ -\frac{1}{2\sigma^2} (X_i^2 - 2X_i\mu + \mu^2) \right\} \]

then \( Z = (2\pi)^{d/2} \prod_i \sigma_i \) and plugging in this into Gibbs distribution above yields a Gaussian distribution.
Constructing a graph for an MRF

The graph is constructed by connecting each pair of nodes corresponding to random variables $X_i, X_j$ if $\exists k, i \in C_k, j \in C_j$.

For example an MRF

$$p(X, Y, Z) = \frac{1}{Z} \phi(X, Y)\psi(Y, Z)$$

and

$$Z = \sum_{X, Y, Z} \phi(X, Y)\psi(Y, Z)$$

corresponds to an undirected graph

[Diagram of an undirected graph with nodes X, Y, and Z connected by lines]
More examples of MRFs

This graphical model specifies a joint distribution

\[
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = \frac{1}{Z} \prod_i \phi(X_{C_i})
\]

\[
= \frac{1}{Z} \phi_1(X_1)\phi_2(X_2, X_1)\phi_3(X_3, X_2)\phi_4(X_4, X_2)
\]

\[
\phi_5(X_5, X_3)\phi_6(X_6, X_4)\phi_7(X_7, X_5, X_6)
\]
Another example of an MRF

\[ p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = \frac{1}{Z} \prod_i \phi(X_{C_i}) \]

\[ = \frac{1}{Z} \phi_1(X_1) \phi_2(X_2, X_1) \phi_3(X_3, X_1) \phi_4(X_4, X_2) \phi_5(X_5, X_2) \phi_6(X_6, X_3) \phi_7(X_7, X_3) \]
Differences from Bayes Nets

No worrying about orderings, conditional probabilities, acyclicity.

Conditional independence is much easier: \( X \perp Y | \mathcal{Z} \) if every path between \( X \) and \( Y \) contains a node corresponding to a variable in \( \mathcal{Z} \).

And immediately from this if we let \text{neighbors}(X) denote the set of neighbors of \( X \) then

\[
X \perp Y | \text{neighbors}(X)
\]

for \( Y \notin X \cup \text{neighbors}(X) \). The set of neighbors is also called Markov blanket.
Soundness and completeness analogouess of MRFs

**Soundness:** If a Gibbs distribution factorizes according to a graph, then all the local Markov properties given by the graph hold for the distribution.

**Hammersley-Clifford theorem:** Let $P$ be positive distribution. If all the independence constraints given by a graph hold for the distribution $P$ than that distribution is a Gibbs distribution over the graph.
MRF applications

Very popular in computer vision.

Can be seen as “energy” models where the energy of configuration is given by

\[ E(X) = - \sum \log \phi_i(X_{C_i}) \]

Gaussian MRFs have several nice properties that enable learning the model structure (the graph). We will come back to this.
Factor graphs: Another undirected representation

These are the easiest to get off the ground.

\[
p(A, B, C) = \frac{1}{Z} f(A, B) g(B, C)
\]

Two types of nodes: variable (one for each random variable) and factor nodes (one for each potential).
Joint probability is given in the same form as in MRF

\[ p(X) = \prod \phi_i(X_{C_i}) \]

but the scope and the different factors are easier to read off of the graph.
Conditional independence for factor graphs

Neighbors as in MRF, if two variables appear in scope of the same potential.

Reading this off of the graph is much easier, since scope is obvious based on connections to factor node.

Conditional independence is easy again: \( X \perp Y | \mathcal{Z} \) if every path between \( X \) and \( Y \) contains a node corresponding to variable \( \mathcal{Z} \).

And in general

\[
X \perp Y | \text{neighbors}(X)
\]

for \( Y \notin X \cup \text{neighbors}(X) \).
More examples of Factor Graphs

This graphical model specifies a joint distribution

\[
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = \frac{1}{Z} \prod_i \phi(X_{C_i})
\]

\[
= \frac{1}{Z} f_1(X_1)f_2(X_2, X_1)f_3(X_3, X_2)f_4(X_4, X_2)
\]

\[
f_5(X_5, X_3)f_6(X_6, X_4)f_7(X_7, X_5, X_6)
\]
Another example of a Factor Graph

\[ p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = \frac{1}{Z} \prod_i \phi(X_{C_i}) \]

\[ = \frac{1}{Z} f_1(X_1) f_2(X_2, X_1) f_3(X_3, X_1) f_4(X_4, X_2) f_5(X_5, X_2) f_6(X_6, X_3) f_7(X_7, X_3) \]
In some cases we may have a particular model structure that is repeatedly reused.

- Same parameter reused across data instances
- Same structure reused across instances
- Any other repetitive regularity
Plate notation: IID

\[ p(X_1, X_2, \ldots, X_N | \theta) = p(X_1 | \theta)p(X_2 | \theta) \ldots p(X_N | \theta) = \prod_i p(X_i | \theta) \]

Random variables \( X_i \) are independently and identically distributed.
An early example of a comp bio graphical model is a phylogeny

Each of the random variables is a nucleotide. The conditional probabilities $p(X_{i,n}|X_{j,n})$ are specified by a single $4 \times 4$ mutation matrix $\theta$ shared across $N$ positions in aligned genome sequences.
Graphical model for MoG

\begin{align*}
p(h_n = k) &= \pi_k \\
p(x_n|h_n = k) &= \mathcal{N}(x_n|\mu_k, \Psi_k)
\end{align*}

Mixture of Gaussians model for $N$ data points and $K$ classes (mixture components)
Graphical model for MoPWM

\[ \pi_c \rightarrow h_n \rightarrow x_n \]

\[ \begin{align*}
  p(h_n = k) &= \pi_k \\
  p(x_n | h_n = k) &= \prod_j \prod_v \theta_k^{[x_n,j=v]} 
\end{align*} \]

Mixture of PWMs model for \( N \) data points and \( K \) classes (mixture components)