CS340 Machine learning
Graphical models
Outline

• Undirected graphical models
• Directed graphical models
• Conditional independence
• Effects of node ordering
• Markov equivalence
• Bayesian modeling
Undirected graphical models

• A prob distribution factorizes wrt an undirected graph \( G \) if it can be written as

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

\[
Z = \sum_x \prod_{c \in C} \psi_c(x_c)
\]

• where \( C \) are the (maximal) cliques of \( G \), \( Z \) is the partition function and \( \psi(x_c) \geq 0 \) are potential functions.

Potential functions are like soft constraints. We will see examples later.
Example model

Alice and Bob more likely to agree (0,0 or 1,1) than disagree; more likely to both be right (0,0) than both be wrong (1,1)

Charles and Debbie more likely to disagree than agree

X=1 if student X has misconception about homework, else X=0

Source: Koller and Friedman p220
Inference

- Given a joint distribution, we can compute the marginals on any variables of interest

\[ p(b = 1) = \sum_{a=0}^{1} \sum_{c=0}^{1} \sum_{d=0}^{1} p(a, b = 1, c, d) = 0.18 \]

- And hence any conditionals of interest

\[ p(b = 1|c = 0) = \frac{p(b = 1, c = 0)}{p(c = 0)} = 0.06 \]
Graph separation

• We say $S$ separates $A$ and $B$ in $G$ if, when we remove edges connected to $S$, all paths from $A$ to $B$ are blocked.

• Hammersley-Clifford Theorem: if $p(x) > 0$ for all $x$, and $p$ factorizes over $G$, then graph separation iff conditional independence.

\[
A \perp_G B \mid S \iff A \perp_p B \mid S
\]

eg \{2,5\} separates 1 and 4
Markov properties

- **Global** \( A \perp B | S \)

- **Local** \( \alpha \perp V \setminus cl(\alpha) | bd(\alpha) \)

bd = boundary, 
cl = closure = boundary + node

A node is independent of the rest given its Markov blanket
Outline

- Undirected graphical models
- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling
Directed graphical models

- A prob distribution factorizes according to a DAG if it can be written as

\[
p(x) = \prod_{j=1}^{d} p(x_j | x_{\pi_j})
\]

where \(\pi_j\) are the parents of \(j\), and the nodes are ordered topologically (parents before children).

Each row of the conditional probability table (CPT) defines the distribution over the child’s values given its parents values. The model is locally normalized.

\[
p(x_{1:6}) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3)p(x_5|x_2, x_3)p(x_6|x_2, x_5)
\]
Example model

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

Source: Russell & Norvig
Example model

\[
p(C, S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R)
\]
Joint distribution

\[ p(C, S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R) \]

c s r w prob
0 0 0 0 0.200
0 0 0 1 0.000
0 0 1 0 0.005
0 0 1 1 0.045
0 1 0 0 0.020
0 1 0 1 0.180
0 1 1 0 0.001
0 1 1 1 0.050
1 0 0 0 0.090
1 0 0 1 0.000
1 0 1 0 0.036
1 0 1 1 0.324
1 1 0 0 0.001
1 1 0 1 0.009
1 1 1 0 0.000
1 1 1 1 0.040
Inference

- Prior that sprinkler is on

\[ p(S = 1) = \sum_{c=0}^{1} \sum_{r=0}^{1} \sum_{w=0}^{1} p(C = c, S = 1, R = r, W = w) = 0.3 \]

- Posterior that sprinkler is on given that grass is wet

\[ p(S = 1|W = 1) = \frac{p(S = 1, W = 1)}{p(W = 1)} = 0.43 \]

- Posterior that sprinkler is on given that grass is wet and it is raining

\[ p(S = 1|W = 1, R = 1) = \frac{p(S = 1, W = 1, R = 1)}{p(W = 1, R = 1)} = 0.19 \]

Explaining away!
Outline

- Undirected graphical models
- Directed graphical models
  - Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling
Conditional independence properties of DAGs

• For UGMs, independence $\equiv$ separation.
• For DGMs, independence $\equiv$ d-separation.
• Alternatively, we can convert a DGM to a UGM and use simple separation.
Moralization

- We can convert a DAG to an undirected graph by moralizing it, i.e., forcing unmarried parents who have a child to get connected, and then dropping all the arrows.
The ancestral graph of G wrt U is one in which we remove any node that is not in U or any ancestor of U, together with any edges in or out of such nodes.
Conditional independence in DAGs

• One can show that A is independent of B given S iff A \textit{d-separates} B given S, where d-separation is like graph separation but pays attention to edge orientation (cf Bayes ball). This is complex to define.

• A simpler definition is the following: A is independent of B given S iff A is separated from B given S in the moralization of the ancestral graph of G wrt A,B,S.
Example

• Is $1 \perp 4 \mid \{5,7\}$?
Chains and tents

1 \rightarrow 2 \rightarrow 3

1 \perp 3 \times \quad 1 \perp 3 \times

\leftarrow 2 \triangleright 3

1 \perp 2 \perp 3

1 \leftrightarrow 2 \leftrightarrow 3

1 \perp 2 \perp 3

1 \perp 2 \perp 3

1 \perp 2 \perp 3

1 \perp 2 \perp 3

1 \perp 2 \perp 3
V-structures

Explaining away couples parents of observed children or grand-children
Markov blankets for DAGs

• The Markov blanket of a node is the set that renders it independent of the rest of the graph.
• This is the parents, children and co-parents.

\[
p(X_i | X_{-i}) = \frac{p(X_i, X_{-i})}{\sum_x p(X_i, X_{-i})} = \frac{p(X_i, U_{1:n}, Y_{1:m}, Z_{1:m}, R)}{\sum_x p(x, U_{1:n}, Y_{1:m}, Z_{1:m}, R)} = \frac{p(X_i | U_{1:n})[\prod_j p(Y_j | X_i, Z_j)]P(U_{1:n}, Z_{1:m}, R)}{\sum_x p(X_i = x | U_{1:n})[\prod_j p(Y_j | X_i = x, Z_j)]P(U_{1:n}, Z_{1:m}, R)} = \frac{p(X_i | U_{1:n})[\prod_j p(Y_j | X_i, Z_j)]}{\sum_x p(X_i = x | U_{1:n})[\prod_j p(Y_j | X_i = x, Z_j)]}
\]

Useful for Gibbs sampling
Local directed Markov property

- A node is independent of its non-descendants given its parents
Ordered directed Markov property

- A node is independent of its predecessors (in some total ordering) given its parents.
Equivalence

- Thm: the following are all equivalent for DAG G
  - $P$ factorizes according to $G$
  - $P$ obeys the global Markov property wrt $G$
  - $P$ obeys the local Markov property wrt $G$
  - $P$ obeys the directed Markov property wrt $G$
Outline

- Undirected graphical models
- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling
Example model

• Suppose the true distribution is

\[ p(B, E, A, J, M) = p(B)p(E)p(A|B, E)p(J|A)p(M|A) \]
Choosing the “wrong” ordering

- If we choose the order MJABE, we get a more densely connected network, otherwise this will make independence statements that are not true.
- E.g., in the original model we have \( E \perp M \mid A, \ E \perp J \mid A, \ E \not\perp B \mid A \)
  so we must connect E to B,A but not M,J

Source: Russell & Norvig
A worse ordering

• If we pick the order MJEBA, the graph becomes fully connected, and thus makes no independence statements (and therefore includes the true distribution).
Outline

• Undirected graphical models
• Directed graphical models
• Conditional independence
• Effects of node ordering
  • Markov equivalence
• Bayesian modeling
Markov equivalence

- The following 3 graphs all assert the same set of conditional independencies, namely $X \text{ indep } Y \mid Z$; hence they are equivalent.

This v-structure is not equivalent.
Markov equivalence

• Thm: 2 DAGs are Markov equivalent iff they have the same undirected skeleton and the same set of v-structures

\[ G_1 \cong G_2 ? \]
\[ G_1 \cong G_3 ? \]
PDAGs

• We can uniquely represent each equivalence class using a partially directed acyclic graph (aka essential graph).

• This uses undirected edges if they are reversible, and directed edges if they are compelled.
Outline

• Undirected graphical models
• Directed graphical models
• Conditional independence
• Effects of node ordering
• Markov equivalence
→• Bayesian modeling
• If we treat the parameters as random variables, we can add them as nodes to the graph.
• Here we assume global parameter independence.
Repetitive structure

- If we have iid samples, the variables get replicated but the parameters are tied / shared
For shorthand, we use plates

\[ p(D, \theta) = p(\theta_c)p(\theta_s)p(\theta_r)p(\theta_w) \times \prod_{i=1}^{n} p(c_i|\theta_c)p(s_i|c_i, \theta_s)p(r_i|c_i, \theta_r)p(w_i|s_i, r_i, \theta_w) \]
Factored prior, likelihood, posterior

- Since the parameters are independent in the prior, and the likelihood is factorized, they are also independent in the posterior

\[
p(\theta|D) \propto p(\theta)p(D|\theta) = p(\theta_c) \prod_i p(c_i|\theta_c) \times p(\theta_s) \prod_i p(s_i|c_i, \theta_s) \times p(\theta_r) \prod_i p(r_i|c_i, \theta_r) \times p(\theta_w) \prod_i p(w_i|s_i, r_i, \theta_s)
\]
Local parameter independence

- In the case of CPTs, we assume each row of the table is an independent multinomial.
Hyperparameters

- The hyperparameters are often fixed constants, hence shaded

\[ p(D, \theta | \alpha) = \prod_j p(\theta_j | \alpha_j) \prod_i p(x_{ij} | x_i, \pi_j, \theta_j) \]
Posterior over parameters factorizes

\[ p(\theta_R|D) = \prod_{k=0}^{1} p(\theta_R|C=k) \prod_{i=1}^{n} I(c_i = k)p(r_i|\theta_R|C=k) = \prod_{k} Dir(\theta_R|C=k|\alpha_R|C=k)Mu(n_R,C=k|\theta_R|C=k,n) \]

\[ p(\theta|D) = \prod_{j=1}^{d} \prod_{k \in \text{Pa}(j)} Dir(\theta_{jk}|\alpha_{jk} + n_{jk}) \]
Naïve Bayes classifier
Example: Binary features

\[ p(D, \pi, \theta | \alpha, a, b) \]

\[ = p(\pi | \alpha) \prod_i p(y_i | \pi) \prod_c \left[ \prod_j \prod_{i: y_i = c} p(x_{ij} | \theta_{jc}) \right] p(\theta_{jc}) \]

\[ = Dir(\pi | \alpha) Mu(n | \pi) \prod_c \prod_j Bin(n_{jc1} | \theta_{jc}, n_{jc}) Beta(\theta_{jc} | a_{jc}, b_{jc}) \]

\[ = Dir(\pi | \alpha + n) \prod_c \prod_j Beta(\theta_{jc} | a_{jc} + n_{jc1}, b_{jc} + n_{jc0}) \]

\[ n_{jc1} = \sum_i I(y_i = c)I(x_{ij} = 1) \]

\[ n_{jc0} = \sum_i I(y_i = c)I(x_{ij} = 0) \]

\[ n_{jc} = n_c = \sum_i I(y_i = c) \]

\[ n = (n_1, \ldots, n_C) \]