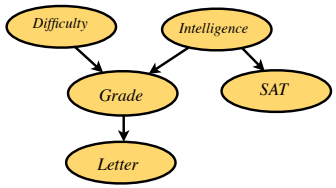
Comments on scholastic achievement example

For $P(I,D,G,L,S)$, suppose $I(Graph)$ tells us

$(D \perp I,S|\varnothing)$ $(L \perp I,D,S|G)$ $(S \perp D,G,L|I)$ $(G \perp S|I,D)$



Question from last time:

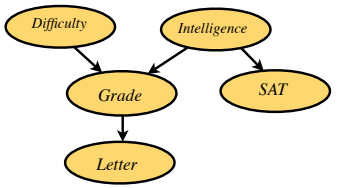
Why $(G \perp S|I,D)$ and not $(G \perp S|I)$?

1) We did not set out to write down all (derivable) statements
2) Can get $(G \perp S|I)$ from $(S \perp D,G,L|I)$

Comments on scholastic achievement example

For $P(I,D,G,L,S)$, suppose $I(Graph)$ tells us

$(D \perp I,S|\varnothing)$ $(L \perp I,D,S|G)$ $(S \perp D,G,L|I)$ $(G \perp S|I,D)$



Question from last time:

Why $(G \perp S|I,D)$ and not $(G \perp S|I)$?

1) We did not set out to write down all (derivable) statements
2) Can get $(G \perp S|I)$ from $(S \perp D,G,L|I)$

3) **We did not need it!**

More on the equivalence of the two interpretations of directed graphical models

Factorization semantics

Factors are p(node | parents)

Abstract semantics

$X_i \perp \text{NonDescendents}(X_i) \mid \text{Parents}(X_i)$

These are equivalent

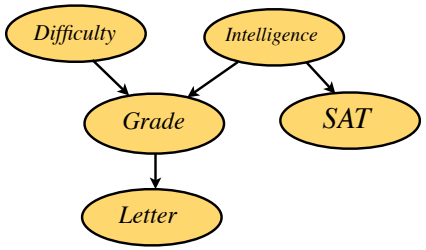
Proof of one direction by one example

Example going from I-map to a factorization

From Kollar and Friedman

For $P(I,D,G,L,S)$, suppose $I(Graph)$ tells us

$(D \perp I,S|\varnothing)$ $(L \perp I,D,S|G)$ $(S \perp D,G,L|I)$ $(G \perp S|I,D)$

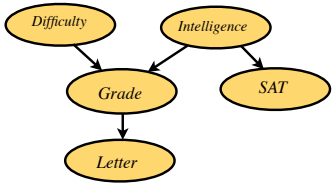


$$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$$

Notice that going left to right, the nodes we condition on are in lexigraphical order, which means that parents occur before children.

This means that for each factor, all variables conditioned on are either the parents, or non-descendents.

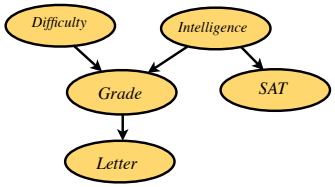
This menas that for each factor, we have a rule that gets rid of the non-descendents, leaving only the parents.



Example going from I-map to a factorization

$$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$$
$$(D \perp I|\varnothing) \Rightarrow P(D|I) = P(D)$$
$$(L \perp I,D,S|G) \Rightarrow P(L|I,D,G) = P(L|G)$$
$$(S \perp D,G,L|I) \Rightarrow P(S|I,D,G,L) = P(S|I)$$

So, $P(I,D,G,L,S) = P(I)P(D)P(G|I,D)P(L|G)P(S|I)$

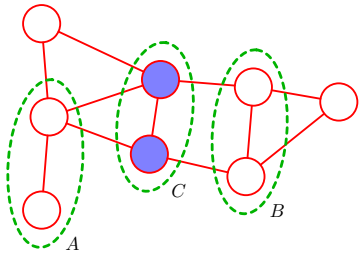


Undirected graphical models

- Also Markov Networks and Markov Random Fields
- Nodes represent (groups of) random variables
- Edges represent probabilistic relations between connected nodes.
- Uses
  - Models where undirected graphs clearly make more sense compared to directed graphs (e.g. spatial information)
  - Representation of inference for directed graphs

Undirected graphical models

- The analog to d-separation is simper



Here $(A \perp B|C)$ for all probability distributions represented by this graph.

Semantics of undirected graphical models

Let C index maximal cliques. Then

$$p(\mathrm{x}) = \frac{1}{Z}\prod_c \psi_C(\mathrm{x}_C)$$

where $Z = \sum_x \prod_c \psi_C(\mathrm{x}_C)$ (or $\int_x \prod_c \psi_C(\mathrm{x}_C)$) is the partition function, and $\psi_C(\mathrm{x}_C)$ are the clique potentials.

We will assume that all $\psi_C(\mathrm{x}_C) > 0$.

Semantics of undirected graphical models (2)

In general, we leave the semantics of $\psi_C(\mathrm{x}_C)$ open, but for undirected graphs that come from directed graphs where each node has one parent, the semantics follows that for the directed graphs.

Since $\psi_C(\mathrm{x}_C) > 0$ we will often write $\psi_C(\mathrm{x}_C) = \exp\{-E(\mathrm{x}_C)\}$ where E() is the energy function.

Semantics of undirected graphical models (3)

Writing $\psi_C(\mathrm{x}_C) = \exp\{-E(\mathrm{x}_C)\}$ means that

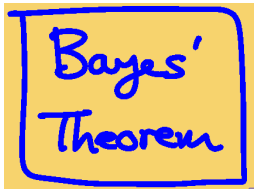
$$\begin{aligned} p(x) &= \frac{1}{Z}\prod_c \psi_x(\mathrm{x}_C) \\ &= \frac{1}{Z}\prod_c \exp\{-E(\mathrm{x}_C)\} \\ &= \frac{1}{Z}\exp\left\{\sum_c -E(\mathrm{x}_C)\right\} \\ &= \frac{1}{Z}\exp\{-E(x)\} \end{aligned}$$

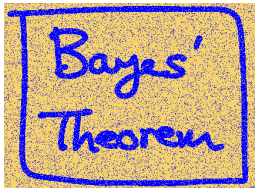
Where $E(x) = \sum_c -E(\mathrm{x}_C)$

Not done in detail in class in 2011. Extra details provided for notes.

Example of a Markov random field

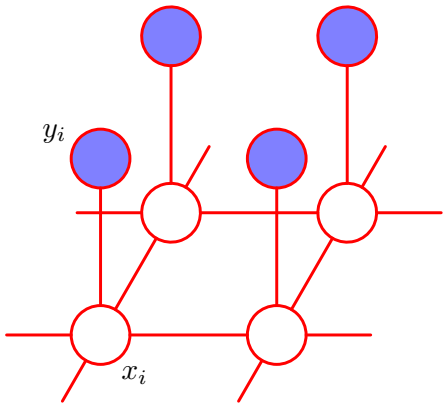
- Consider a binary image (pixels are either black or white).
- Pixels are represented by {-1,1}.
- Suppose the image have is an underlying accurate image where some of the bits have been flipped by a noise process.





Example of a Markov random field (2)

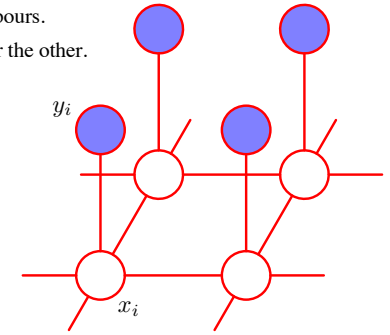
- Undirected graphical model.



Example of a Markov random field (2)

- For low energy (high probability)

$x_i = y_i$ most of the time (set by noise level)
$x_i = x_j$ most of the time if i and j are neighbours.
x_i could be biased to have one value or the other.



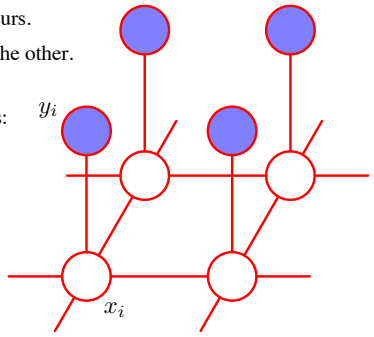
Example of a Markov random field (2)

- For low energy (high probability)

$x_i = y_i$ most of the time (set by noise level)
$x_i = x_j$ most of the time if i and j are neighbours.
x_i could be biased to have one value or the other.

A simple energy function for the entire grid is:

$$E(\mathbf{x},\mathbf{y}) = h\sum x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_y$$



Example of a Markov random field (3)

$x_i = y_i$ most of the time (set by noise level)
$x_i = x_j$ most of the time if i and j are neighbours.
x_i could be biased to have one value or the other.

Additional details glossed over in class provided in notes.

For each $\{x_i, y_i\}$ maximum clique, $\mathrm{E}(x_i, y_i) = -\eta \bullet x_i \bullet y_i \quad (\eta > 0)$
(high probablity corresponds to low energy)

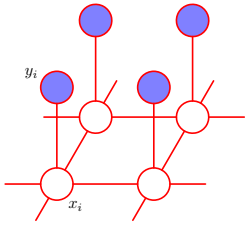
For unique $\{x_i, x_{j \in neighbor(i)}\}$ max clique, $\mathrm{E}(x_i, x_j) = -\beta \bullet x_i \bullet x_j \quad (\beta > 0)$

For a subset of the above cliques, one for each i, add in a term $h \bullet x_i$.

Example of a Markov random field (4)

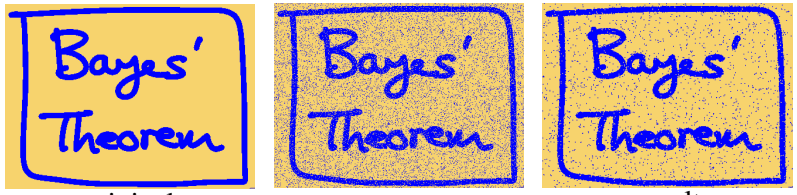
- Notice in the previous analysis we assigned arguably symmetric cliques different potentials
  - Left boundary x_i might get different potentials than right boundary x_i.
  - Some x_{ij} get a factor for the bias, other do not.
- Notice that exact assignment to clique potentials may not matter
- We can jump readily quickly to the overall picture, hence:

$$E(\mathbf{x},\mathbf{y}) = h\sum x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_y$$



Example of a Markov random field (3)

- Finding a low energy (high probability) state using ICM (iterated conditional modes).
  - Initialize x_i to y_i.
  - For each i, change x_i if energy decreases.
  - Repeat until energy no longer can be decreased.
- Converges to a local minimum because we only decrease.



original with noise result

From directed to undirected

- Easy case (all nodes have at most one parent).
- Example:

- Becomes:

From directed to undirected

- Convert:

$$p(\mathrm{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ \cdots\ p(x_{N-1}|x_{N-2})p(x_N|x_{N-1})$$

- To:

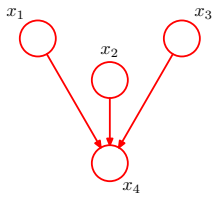
$$p(\mathrm{x}) = \Psi(x_1,x_2)\Psi(x_2,x_3)\ \ldots\ \Psi(x_{N-2},x_{N-1})\Psi(x_{N-1},x_N)$$

- Inspection suggests:

$$\Psi(x_1,x_2) = p(x_1)p(x_2|x_1)$$
$$\Psi(x_2,x_3) = p(x_3|x_2)$$
$$\cdots$$
$$\Psi(x_{N-2},x_{N-1}) = p(x_{N-1}|x_{N-2})$$
$$\Psi(x_{N-1},x_N) = p(x_N|x_{N-1})$$

From directed to undirected

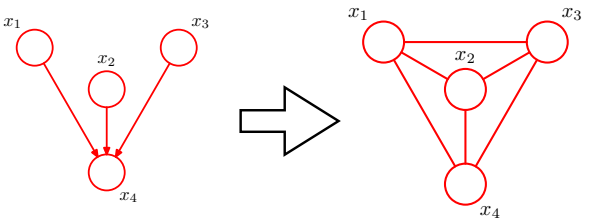
- Harder case (some nodes have multiple parents).
- Example:



- Because this implies conditioning on three variables, the potentials for the clique are a function of four variables.
- These nodes need to be part of a clique (but they are not).

From directed to undirected

- Solution is to marry the parents.
- This makes the graph “moral”.
- Note that moralization looses conditional independence information.



From directed to undirected

- Complete algorithm
  - Make the graph moral.
  - Initialize maximal clique potential to one.
  - Multiply each factor in p() into an appropriate clique potential.
  - Note that Z=1