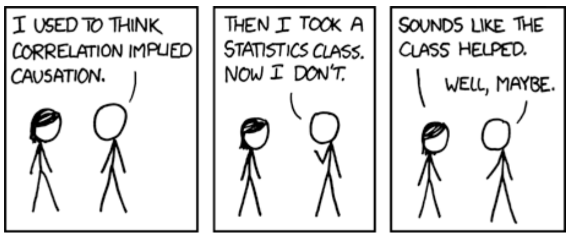
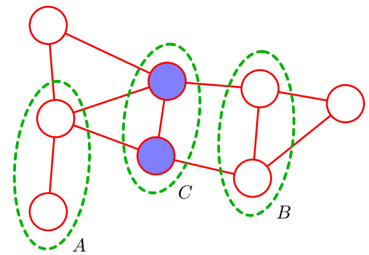
Announcements

- Today we will continue our discussion on Markov random fields.
- Much of this is from Bishop 8.3 (Misconception example is from K&F)



Undirected graphical models

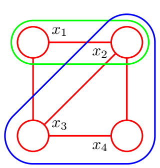
- The analog to d-separation is simper
  - Disjoint sets A and B are independent conditioned on C if all paths from nodes in A to nodes in B pass through C.
  - This defines the network semantics



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

Undirected graphical models

- We are headed to a factorization of the probability distribution in terms of functions over maximal cliques
  - A clique is fully connected subset of nodes from the graph
  - A maximal clique is a clique where no node in the graph can be added to it without it ceasing to be a clique.



All parwise linked nodes are cliques. For example $\{x_1, x_2\}$ is a clique (green). However, it is not a maximal clique. $\{x_2, x_3, x_4\}$ is a maximal clique (blue). If we add another node (only x_1 is left) we no longer have a clique.

Semantics of undirected graphical models

- For two nodes, x_i and x_j, not connected by a link,
  $$x_i \perp x_j | \mathbf{x}/\{i,j\}.$$
- So,
  $$p(\ldots, x_i, \ldots, x_j, \ldots) = p(x_i | \mathbf{x}/\{i,j\})\, p(x_j | \mathbf{x}/\{i,j\})\, p(\mathbf{x}/\{i,j\})$$
- This suggests that an appropriate factorization should not have factors with these (non directly linked) nodes together (so that it is consistent with the conditional independence)
- A group of nodes that are all (fully) connected cannot be factored by the above rule (and hence there is no simplification to be gained).

Factorization for undirected graphical models

Let C index maximal cliques. Then

$$p(\mathrm{x}) = \frac{1}{Z}\prod_{c}\psi_{C}(\mathrm{x}_{C})$$

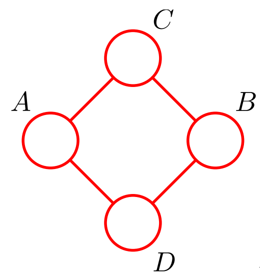
where Z $= \sum_{x}\prod_{c}\psi_{C}(\mathrm{x}_{C})$ (or $\int_{x}\prod_{c}\psi_{C}(\mathrm{x}_{C})$) is the partition function,

and $\psi_{C}(\mathrm{x}_{C})$ are the clique potentials.

If x_i and x_j do not share an edge, then they do not share cliques.

So $p(\mathrm{x}) = \dfrac{1}{Z}\displaystyle\prod_{c(i)}\psi_{C}(\mathrm{x}_{C})\prod_{c(j)}\psi_{C}(\mathrm{x}_{C})\prod_{c\notin c(i)\cup c(j)}\psi_{C}(\mathrm{x}_{C})$

Misconception example



$$p(A,B,C,D) \propto \psi(A,C)\psi(C,B)\psi(B,D)\psi(D,A)$$

Intuitively we have $(A \perp B | C,D)$ because the conditioning specifies C,D, and the factors with A have no B, and vice versa. Similarly, $(C \perp D | A,B)$.

However, let us derive a result to confirm this

Quick warm up

$$\sum_{i=1}^{3}\sum_{j=1}^{3}x_i a_j = x_1a_1 + x_1a_2 + x_1a_3 + x_2a_1 + x_2a_2 + x_2a_3 + x_3a_1 + x_3a_2 + x_3a_3$$

$$= (x_1 + x_2 + x_3)(a_1 + a_2 + a_3) \qquad \text{(gives all combinations } x_i \text{ of and } a_j\text{)}$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3}x_i a_j = \sum_{i=1}^{3}\left(x_i\sum_{j=1}^{3}a_j\right) \qquad \text{(distributive rule)}$$

$$= \left(\sum_{i=1}^{3}x_i\right)\left(\sum_{j=1}^{3}a_j\right) \qquad \left(\sum_{j=1}^{3}a_j \text{ is a constant pulled out a sum over } x\right)$$

$$= (x_1 + x_2 + x_3)(a_1 + a_2 + a_3)$$

$$\sum_{X,Y}\varphi(X,Z)\varphi(Y,Z) = \sum_{X}\varphi(X,Z)\sum_{Y}\varphi(Y,Z) \qquad \left(\text{also } \sum_{Y}\varphi(Y,Z)\sum_{X}\varphi(X,Z)\right)$$

$$= \left(\sum_{X}\varphi(X,Z)\right)\left(\sum_{Y}\varphi(Y,Z)\right) \qquad \left(\sum_{Y}\varphi(Y,Z) \text{ is does not depend on } X\right)$$

$$p(X,Y,Z) = \varphi(X,Z)\varphi(Y,Z) \;\Leftrightarrow\; X \perp Y | Z \qquad \text{(algebra for } \Rightarrow\text{, other direction is easier)}$$

$$p(Y|Z)\,p(X|Z) = \frac{\sum_{X}\varphi(X,Z)\varphi(Y,Z)}{\sum_{X,Y}\varphi(X,Z)\varphi(Y,Z)}\frac{\sum_{Y}\varphi(X,Z)\varphi(Y,Z)}{\sum_{X,Y}\varphi(X,Z)\varphi(Y,Z)}$$

$$= \frac{\varphi(Y,Z)\sum_{X}\varphi(X,Z)}{\sum_{X}\varphi(X,Z)\sum_{Y}\varphi(Y,Z)}\frac{\varphi(X,Z)\sum_{Y}\varphi(Y,Z)}{\sum_{X}\varphi(X,Z)\sum_{Y}\varphi(Y,Z)}$$

$$= \frac{\varphi(Y,Z)\sum_{X}\varphi(X,Z)}{\left(\sum_{X}\varphi(X,Z)\right)\left(\sum_{Y}\varphi(Y,Z)\right)}\frac{\varphi(X,Z)\sum_{Y}\varphi(Y,Z)}{\left(\sum_{X}\varphi(X,Z)\right)\left(\sum_{Y}\varphi(Y,Z)\right)}$$

$$= \frac{\varphi(Y,Z)}{\left(\sum_{Y}\varphi(Y,Z)\right)}\frac{\varphi(X,Z)}{\left(\sum_{X}\varphi(X,Z)\right)} \qquad \text{(canceling green and red pairs)}$$

$$= \frac{\varphi(Y,Z)\varphi(X,Z)}{\sum_{X,Y}\varphi(Y,Z)\varphi(X,Z)}$$

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

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: $\Psi(x_1,x_2) = p(x_1)\,p(x_2|x_1)$

$\Psi(x_2,x_3) = p(x_3|x_2)$

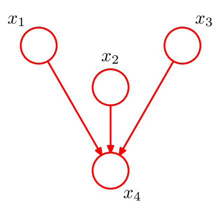
$\cdots$

$\Psi(x_{N-2},x_{N-1}) = p(x_{N-1}|x_{N-2})$

$\Psi(x_{N-1},x_N) = p(x_N|x_{N-1})$

From directed to undirected

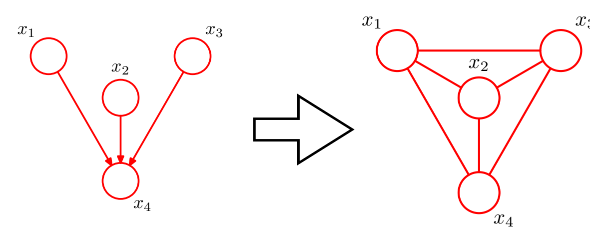
- Harder case (some nodes have multiple parents).
- Example:



- 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

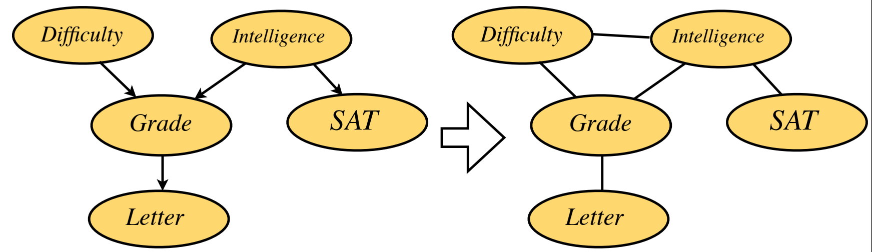
- Solution is to marry the parents.
- This makes the graph “moral”.
- Note that moralization looses conditional independence information.



From directed to undirected

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

Example of converting directed to undirected



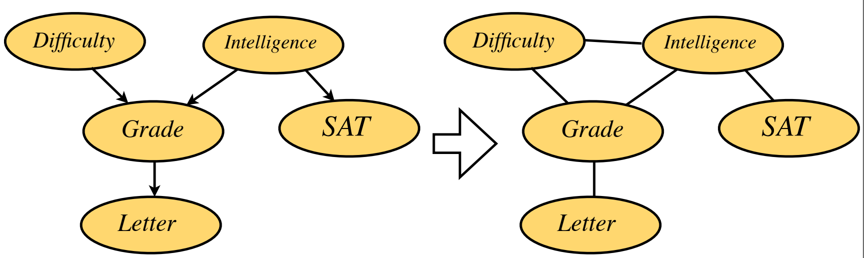
$$P(I,D,G,L,S) = P(I)P(D)P(G|I,D)P(L|G)P(S|I)$$

$$P = \psi(D,G,I)\psi(S,I)\psi(L,G)$$

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

(Is this unique?)

Example of converting directed to undirected



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

$$P = \psi(D,G,I)\psi(S,I)\psi(L,G)$$

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

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

Energy function encoding

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

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 (as we have just done).

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.

Energy function encoding (2)

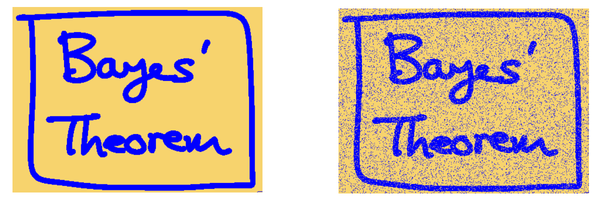
Writing $\psi_C\left(\mathrm{x}_C\right)=\exp\left\{-E\left(\mathrm{x}_C\right)\right\}$ means that

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

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

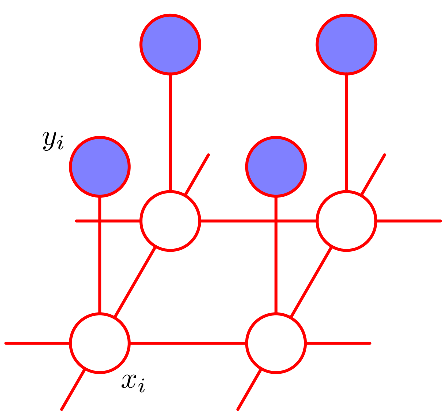
Example of a Markov random field

- Consider a binary image (pixels are either black or white).
  - Pixels are represented by {-1,1}.
- Neighboring pixels tend to have the same color
- 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)

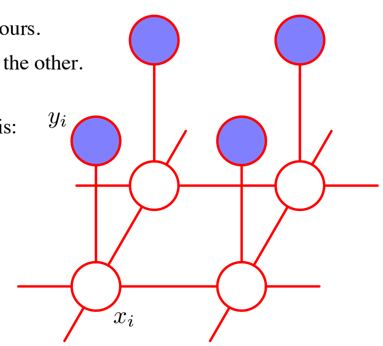
- For low energy (high probability)

$x_i = y_i$ 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.

A simple energy function for the entire grid is:

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

Because values are 1 and -1, being the same makes the sums bigger, being different makes them smaller.



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.

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)

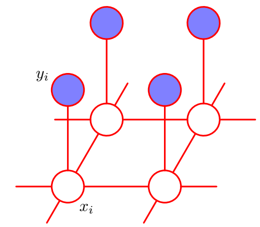
For unique $\{x_i, x_{j \in neighbor(i)}\}$ max clique, $\mathrm{E}(x_i, x_j) = -\beta \bullet x_i \bullet x_j \quad (\beta > 0)$

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)

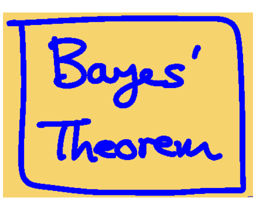
- Notice in the previous analysis we assigned arguably symmetric cliques different potentials
  - Left boundary x_i might get different potentials than right boundary x_i.
  - Some x_{ij} get a factor for the bias, other do not.
- Notice that exact assignment to clique potentials may not matter
- We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

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

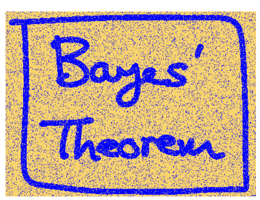


Example of a Markov random field (5)

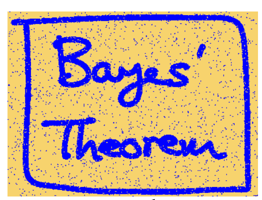
- Finding a low energy (high probability) state using ICM (iterated conditional modes).
  - Initialize x_i to y_i.
  - For each i, change x_i if energy decreases.
  - Repeat until energy no longer can be decreased.
- Converges to a local minimum because we only decrease.



original

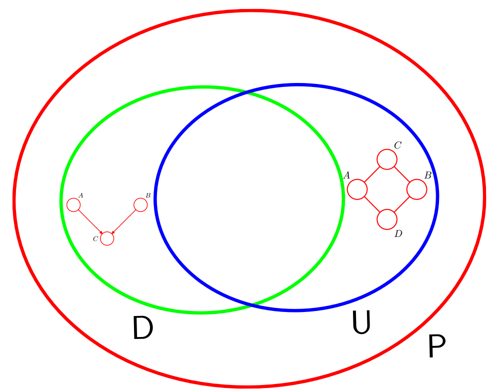


with noise



result

Directed and undirected perfect maps



D is subset of distributions in P that are perfectly represented by directed graphs; similarly U for undirected graphs.