Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e

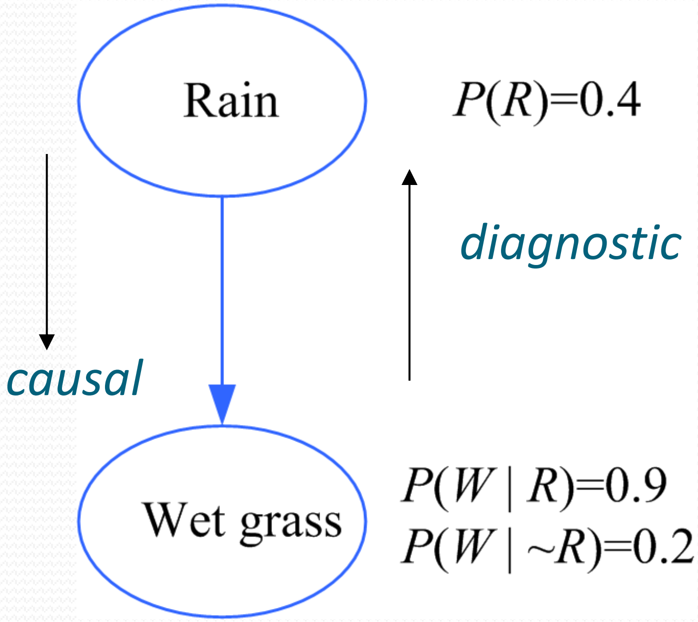
CHAPTER 16:

Graphical Models

Graphical Models

- Aka Bayesian networks, probabilistic networks
- Nodes are hypotheses (random vars) and the probabilities corresponds to our belief in the truth of the hypothesis
- Arcs are direct influences between hypotheses
- The structure is represented as a directed acyclic graph (DAG)
- The parameters are the conditional probabilities in the arcs (Pearl, 1988, 2000; Jensen, 1996; Lauritzen, 1996)

Causes and Bayes' Rule



Diagnostic inference:
Knowing that the grass is wet, what is the probability that rain is the cause?

$$
\begin{aligned}
P(R \mid W) &= \frac{P(W \mid R)P(R)}{P(W)} \\
&= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\end{aligned}
$$

Conditional Independence

- *X* and *Y* are independent if

$$P(X,Y)=P(X)P(Y)$$

- *X* and *Y* are conditionally independent given *Z* if

$$P(X,Y|Z)=P(X|Z)P(Y|Z)$$

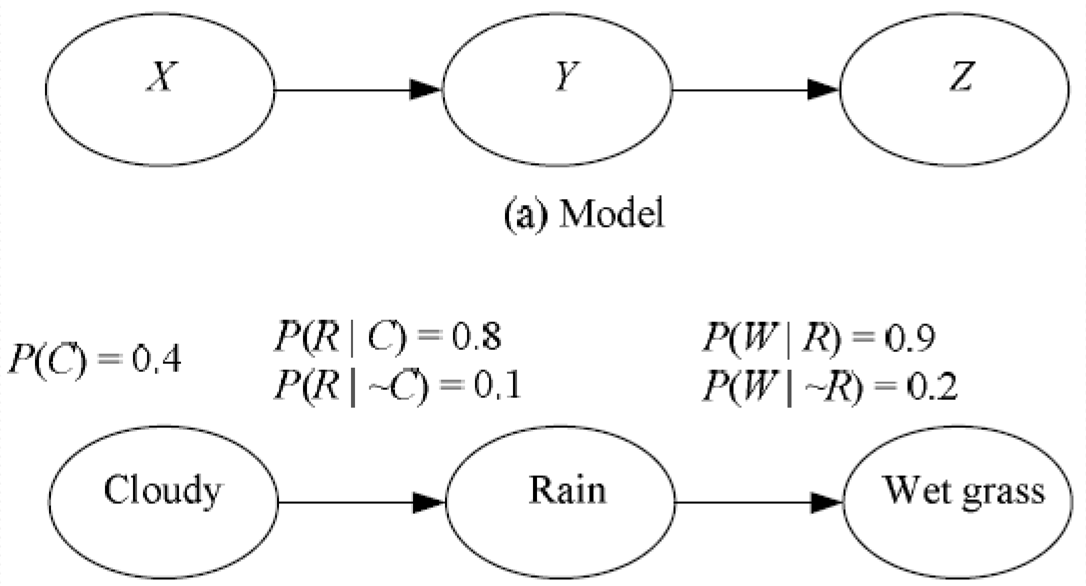
or

$$P(X|Y,Z)=P(X|Z)$$

- Three canonical cases: Head-to-tail, Tail-to-tail, head-to-head

Case 1: Head-to-Head

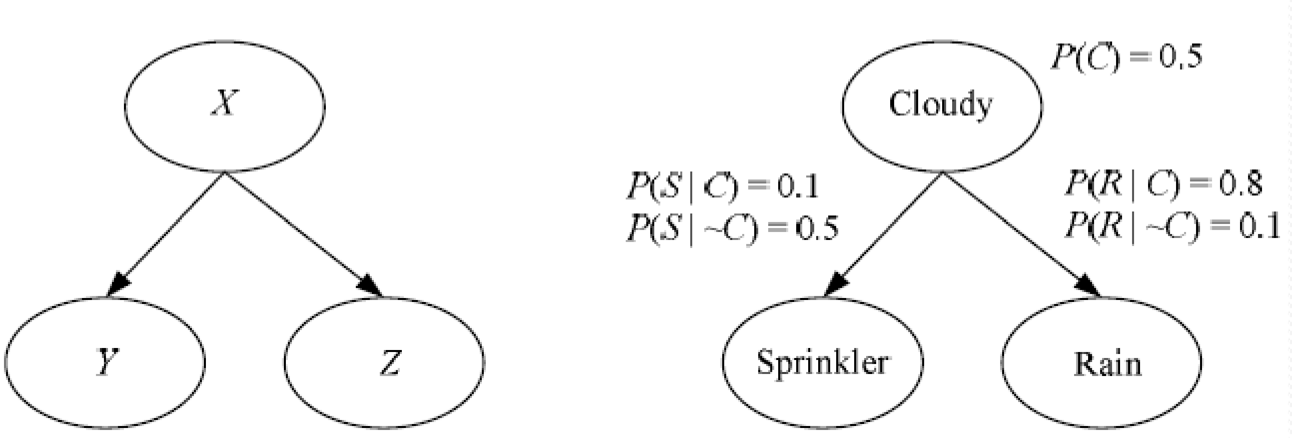
- $P(X,Y,Z)=P(X)P(Y|X)P(Z|Y)$



- $P(W|C)=P(W|R)P(R|C)+P(W|\sim R)P(\sim R|C)$

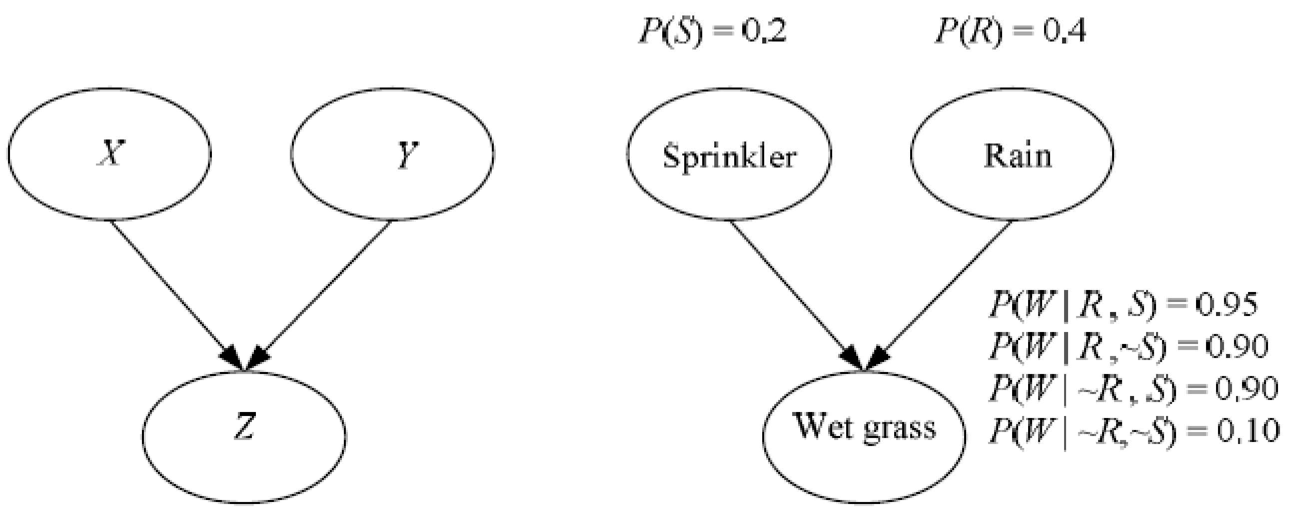
Case 2: Tail-to-Tail

- $P(X,Y,Z)=P(X)P(Y|X)P(Z|X)$

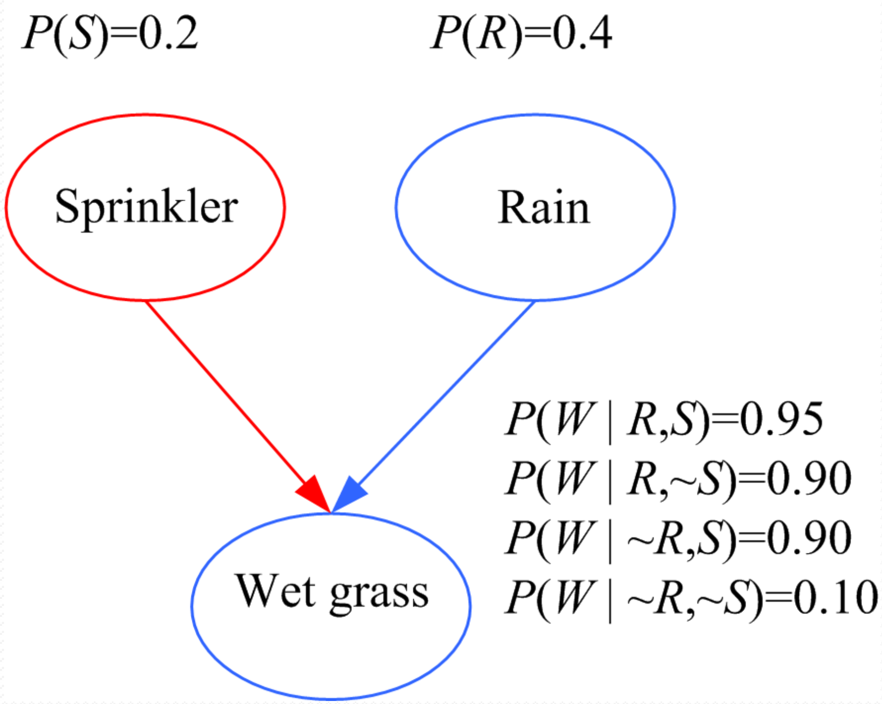


Case 3: Head-to-Head

- $P(X,Y,Z)=P(X)P(Y)P(Z|X,Y)$



Causal vs Diagnostic Inference



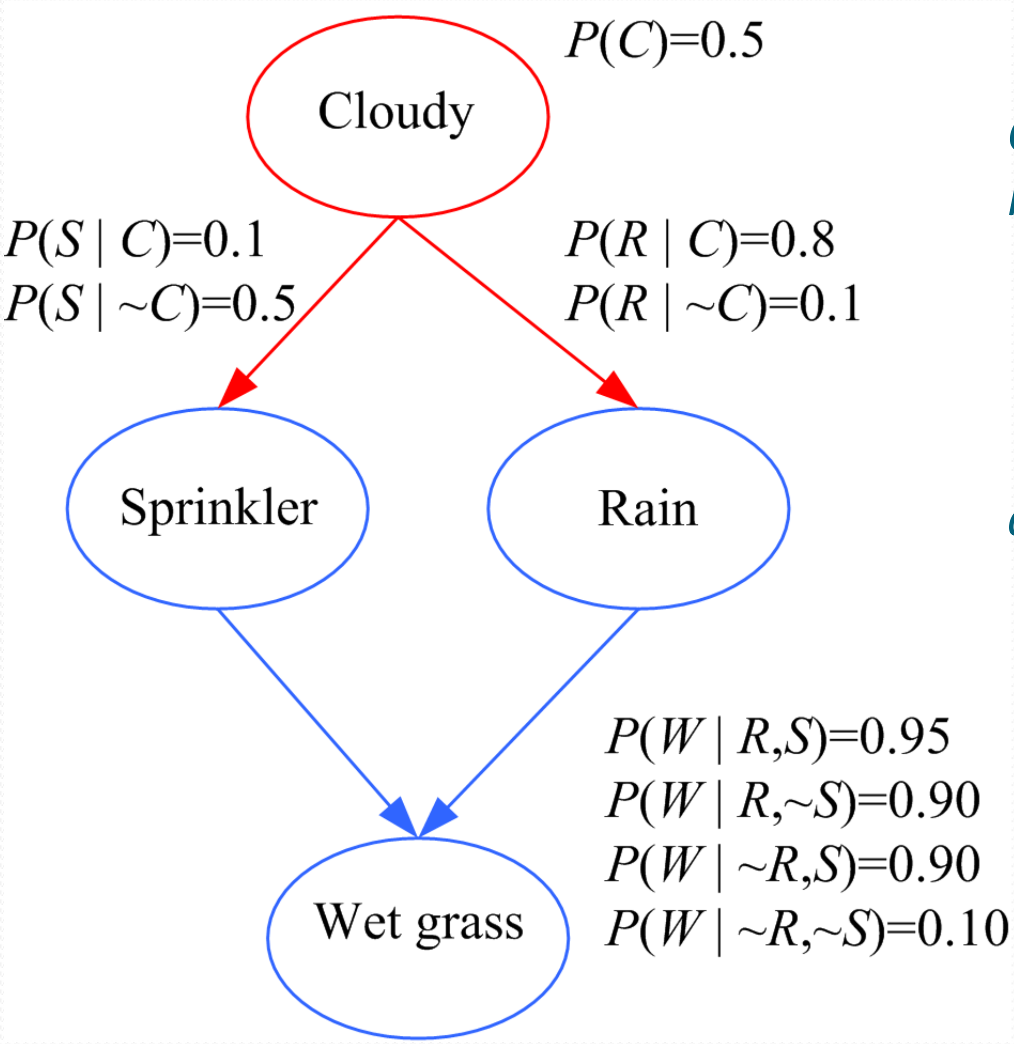
Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|{\sim}R,S)\,P({\sim}R|S)$$
$$= P(W|R,S)\,P(R) + P(W|{\sim}R,S)\,P({\sim}R)$$
$$= 0.95\ 0.4 + 0.9\ 0.6 = 0.92$$

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on? $P(S|W) = 0.35 > 0.2\ P(S)$

$P(S|R,W) = 0.21$ *Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

Causes



Causal inference:

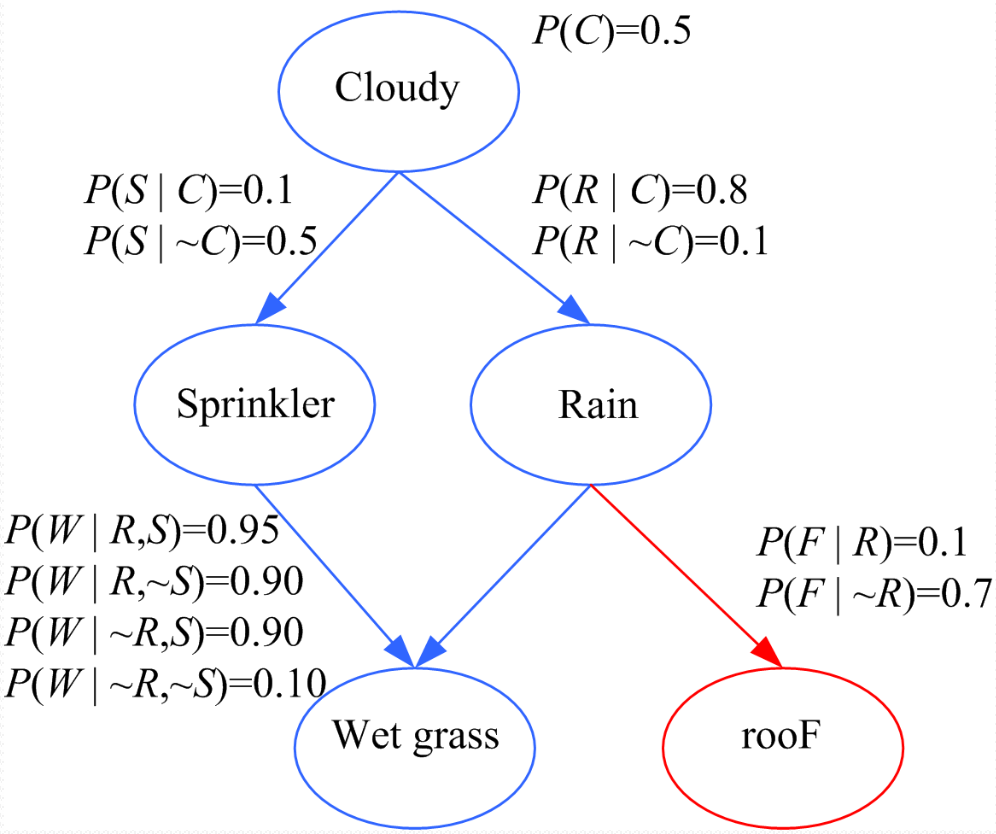
$P(W|C) = P(W|R,S)\, P(R,S|C) +$
$\quad P(W|{\sim}R,S)\, P({\sim}R,S|C) +$
$\quad P(W|R,{\sim}S)\, P(R,{\sim}S|C) +$
$\quad P(W|{\sim}R,{\sim}S)\, P({\sim}R,{\sim}S|C)$

and use the fact that

$P(R,S|C) = P(R|C)\, P(S|C)$

Diagnostic: $P(C|W) = ?$

Exploiting the Local Structure



$P(F \mid C) = ?$

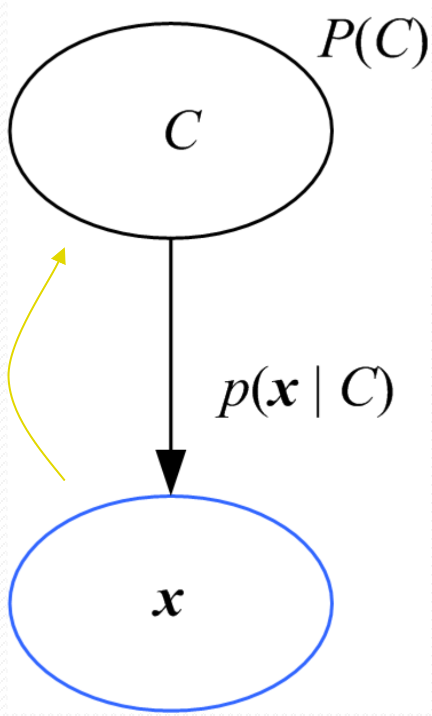
$$P(C,S,R,W,F)=P(C)P(S \mid C)P(R \mid C)P(W \mid S,R)P(F \mid R)$$

$$P(X_1,\ldots X_d)=\prod_{i=1}^{d} P(X_i \mid \text{parents}(X_i))$$

Classification

diagnostic

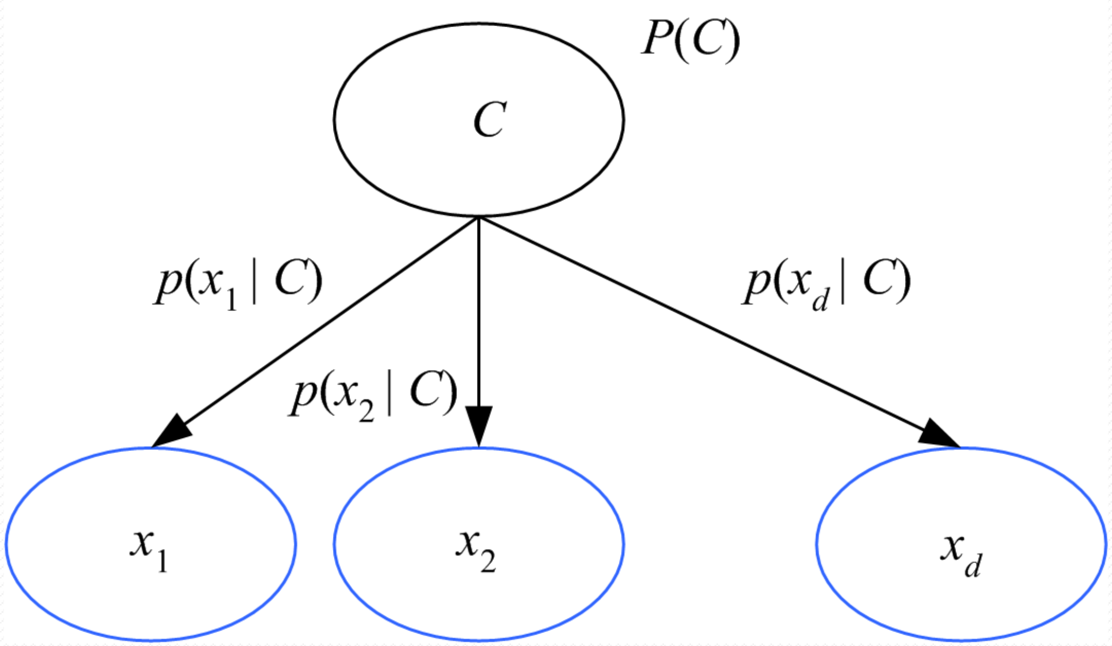
P (C | x)



Bayes' rule inverts the arc:

$$P(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C) P(C)}{p(\mathbf{x})}$$

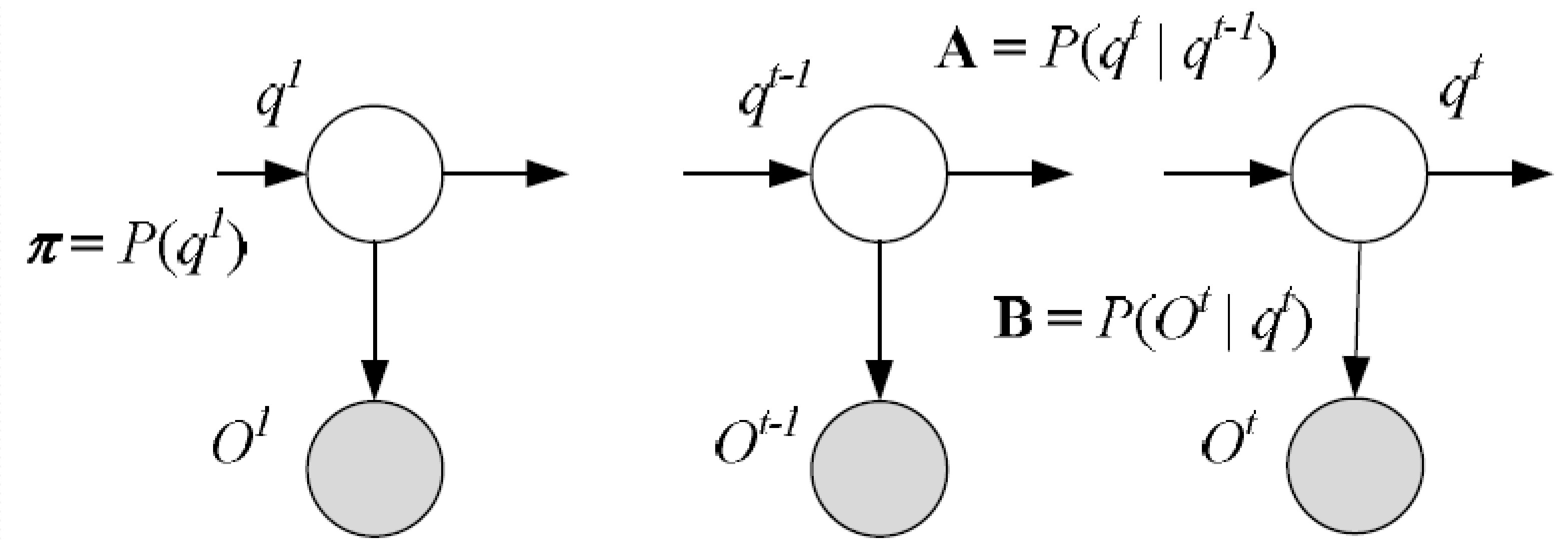
Naive Bayes' Classifier

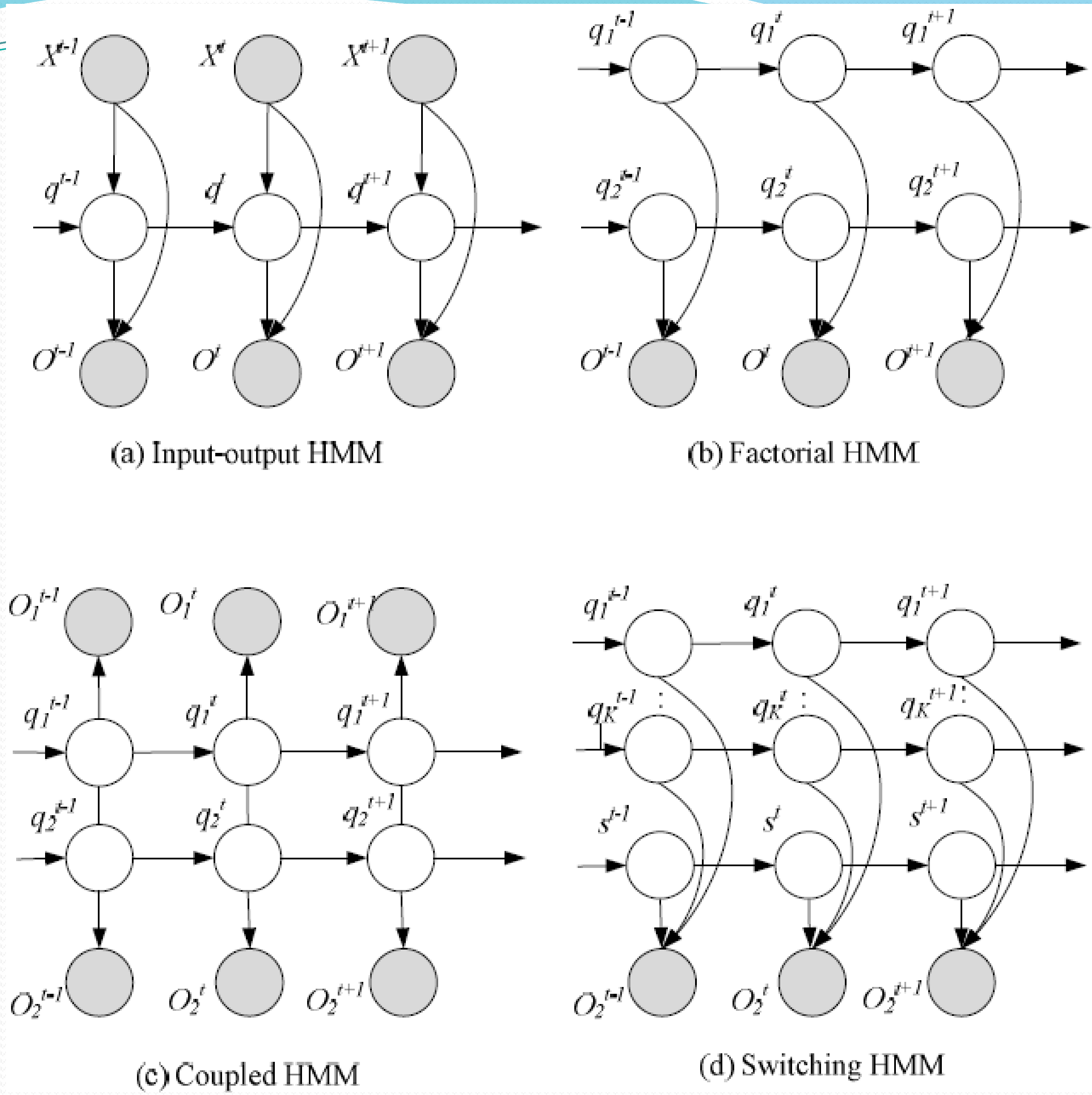


Given C, x_j are independent:

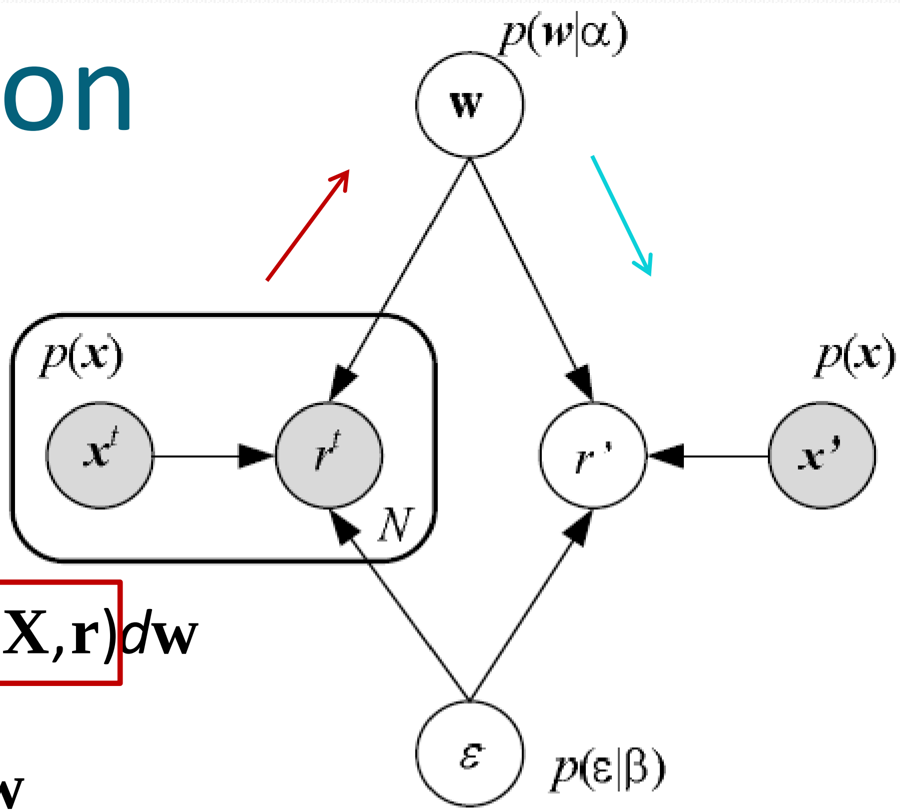
$$p(\mathbf{x}|C) = p(x_1|C)\,p(x_2|C)\,\ldots\,p(x_d|C)$$

Hidden Markov Model as a Graphical Model




(a) Input-output HMM
(b) Factorial HMM
(c) Coupled HMM
(d) Switching HMM

Linear Regression



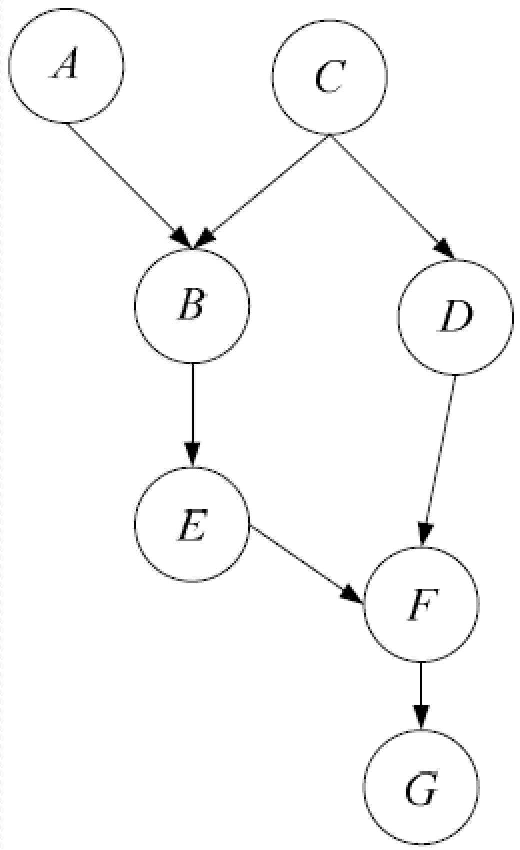
$$p(r'|\mathbf{x}',\mathbf{r},\mathbf{X}) = \int \boxed{p(r'|\mathbf{x}',\mathbf{w})}\,\boxed{p(\mathbf{w}\,|\,\mathbf{X},\mathbf{r})}\,d\mathbf{w}$$

$$= \int p(r'|\mathbf{x}',\mathbf{w})\frac{p(\mathbf{r}\,|\,\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{r})}d\mathbf{w}$$

$$\propto \int p(r'|\mathbf{x}',\mathbf{w})\prod_t p(r^t\,|\,\mathbf{x}^t,\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

d-Separation

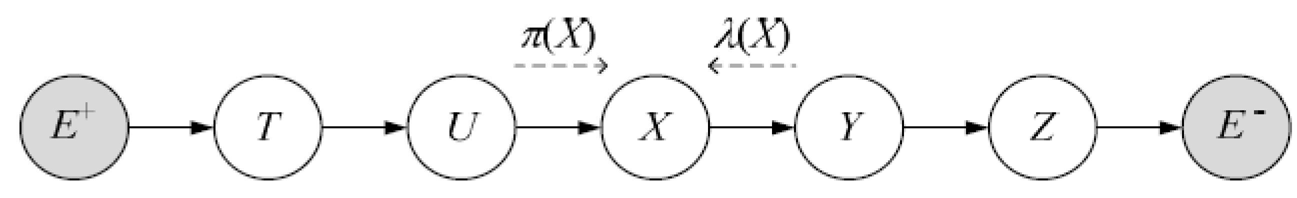
- A path from node *A* to node *B* is blocked if
  a) The directions of edges on the path meet head-to-tail (case 1) or tail-to-tail (case 2) and the node is in *C*, or
  b) The directions of edges meet head-to-head (case 3) and neither that node nor any of its descendants is in *C*.
- If all paths are blocked, *A* and *B* are d-separated (conditionally independent) given *C*.



BCDF is blocked given *C*.
BEFG is blocked by *F*.
BEFD is blocked unless *F* (or *G*) is given.

Belief Propagation (Pearl, 1988)

- Chain:



$$P(X \mid E) = \frac{P(E \mid X)P(X)}{P(E)} = \frac{P(E^+, E^- \mid X)P(X)}{P(E)}$$

$$= \frac{P(E^+ \mid X)P(E^- \mid X)P(X)}{P(E)} = \alpha\pi(X)\lambda(X)$$

$$\pi(X) = \sum_U P(X \mid U)\pi(U)$$

$$\lambda(X) = \sum_Y P(Y \mid X)\lambda(Y)$$

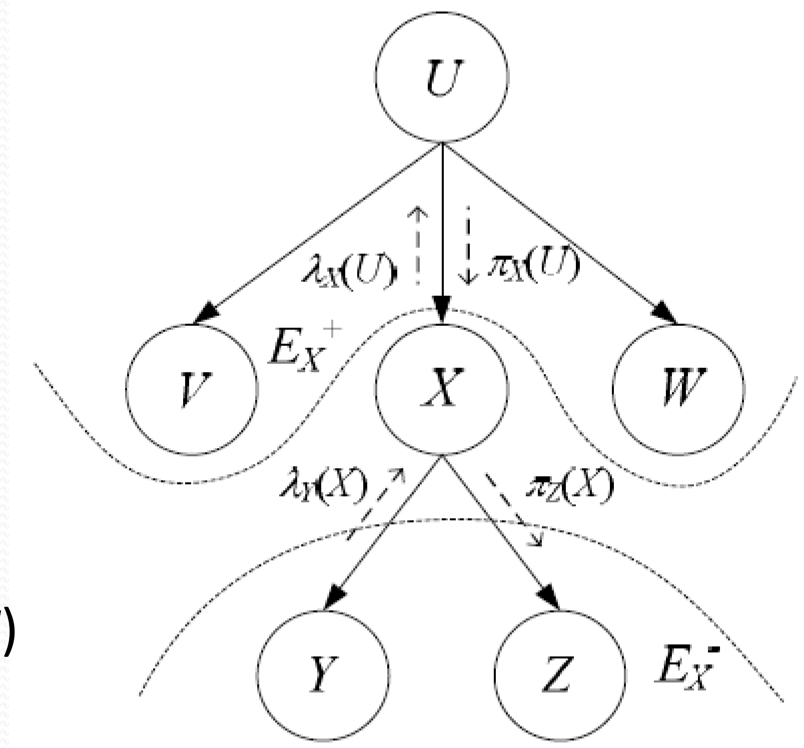
Trees

$$\lambda(X) = P(E_X^- \mid X) = \lambda_Y(X)\lambda_Z(X)$$

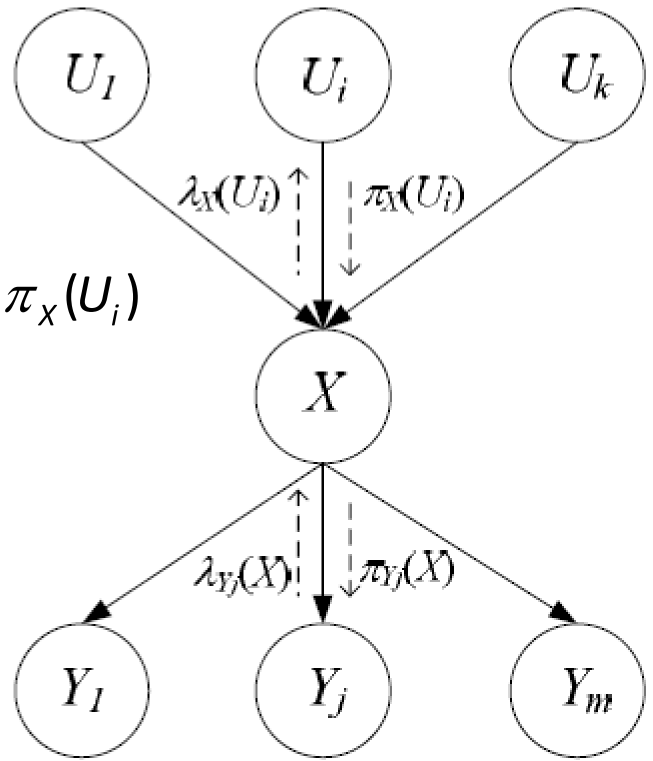
$$\lambda_X(U) = \sum_X \lambda(X)P(X \mid U)$$

$$\pi(X) = P(X \mid E_X^+) = \sum_U P(X \mid U)\pi_X(U)$$

$$\pi_y(X) = \alpha\lambda_Z(X)\pi(X)$$



Polytrees



$$\pi(X) = P(X \mid E_X^+) = \sum_{U_1}\sum_{U_2}\cdots\sum_{U_k} P(X \mid U_1, U_2, \cdots, U_k) \prod_{i=1}^{k} \pi_X(U_i)$$

$$\pi_{y_j}(X) = \alpha \prod_{s \neq j} \lambda_{Y_s}(X)\pi(X)$$

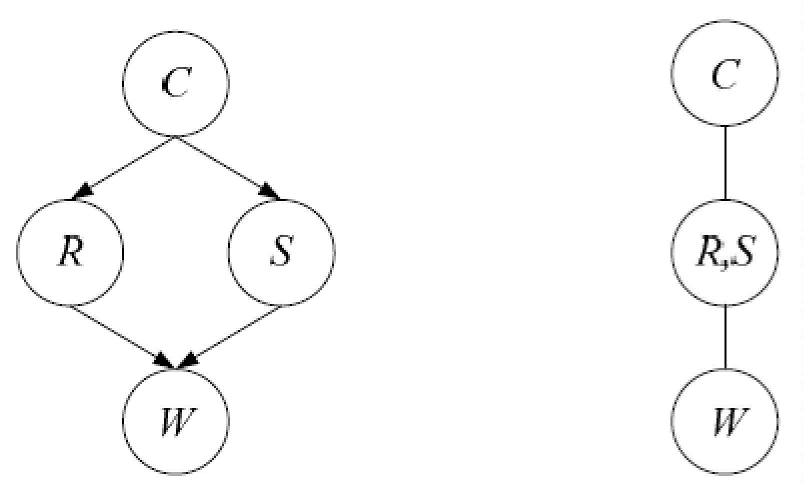
$$\lambda_X(U_i) = \beta \sum_X \lambda(X) \sum_{U_{r \neq i}} P(X \mid U_1, U_2, \cdots, U_k) \prod_{r \neq i} \pi_X(U_r)$$

$$\lambda(X) = \prod_{j=1}^{m} \lambda_{Y_j}(X)$$

How can we model $P(X|U_1,U_2,...,U_k)$ cheaply?

Junction Trees

- If *X* does not separate E^+ and E^-, we convert it into a junction tree and then apply the polytree algorithm



Tree of moralized, clique nodes

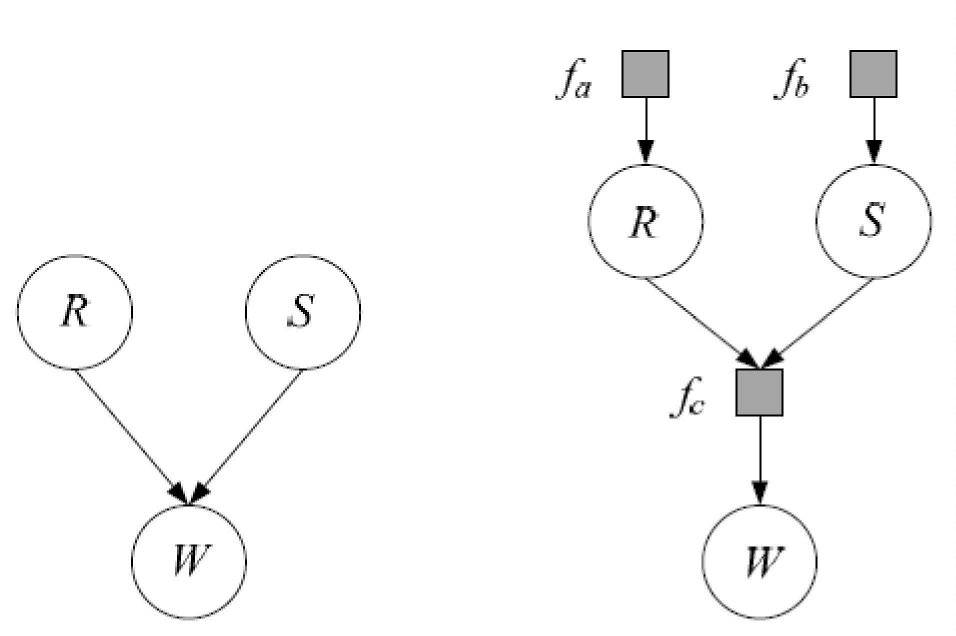
Undirected Graphs: Markov Random Fields

- In a Markov random field, dependencies are symmetric, for example, pixels in an image
- In an undirected graph, *A* and *B* are independent if removing *C* makes them unconnected.
- Potential function $\psi_c(X_c)$ shows how favorable is the particular configuration *X* over the clique *C*
- The joint is defined in terms of the clique potentials

$$p(X) = \frac{1}{Z}\prod_{C}\psi_C(X_C) \text{ where normalizer } Z = \sum_{X}\prod_{C}\psi_C(X_C)$$

Factor Graphs

- Define new factor nodes and write the joint in terms of them



$$p(X) = \frac{1}{Z} \prod_S f_S(X_S)$$

Learning a Graphical Model

- Learning the conditional probabilities, either as tables (for discrete case with small number of parents), or as parametric functions
- Learning the structure of the graph: Doing a state-space search over a score function that uses both goodness of fit to data and some measure of complexity

Influence Diagrams

decision node

chance node

utility node