



*Lecture Slides for*

INTRODUCTION TO

# *Machine Learning*

ETHEM ALPAYDIN

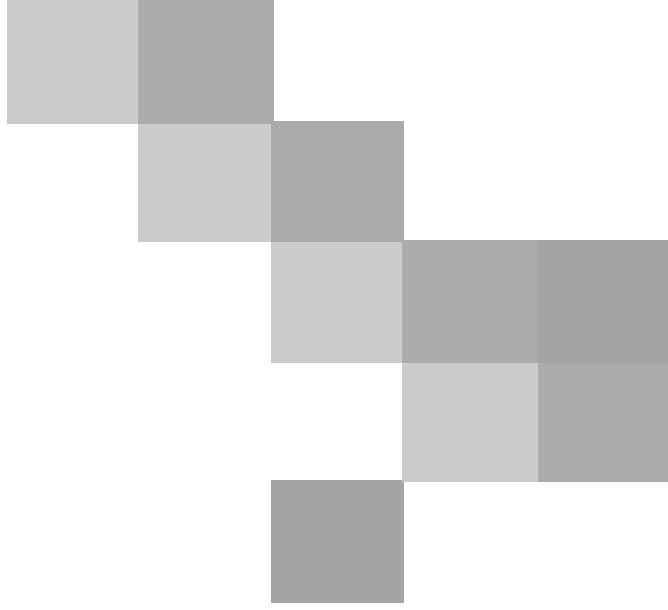
© The MIT Press, 2004

*alpaydin@boun.edu.tr*

*<http://www.cmpe.boun.edu.tr/~ethem/i2ml>*

CHAPTER 3:

*Bayesian Decision  
Theory*



# Probability and Inference

- The world  $\rightarrow$  unknown process  $\rightarrow$  data.
  - Because of our lack of knowledge about the process, we model it as a random process and use probability theory to analyse it.
- Result of tossing a coin is  $\in$  {Heads, Tails}
- Random var  $X \in \{1, 0\}$

Bernoulli:  $P\{X=1\} = p_o^X (1 - p_o)^{(1-X)}$  also  $P(X=1|p_o)$

$$P(X=1) = 1 - P(X=0) = p_o$$

- Sample:  $\mathbf{X} = \{x^t\}_{t=1}^N$
- Estimation:  $p_o = \# \{\text{Heads}\} / \#\{\text{Tosses}\} = \sum_t x^t / N$
- Prediction of next toss:
  - Heads if  $p_o > 1/2$ , Tails otherwise

# Classification

- Credit scoring: Inputs are income and savings.  
Output is low-risk vs high-risk
- Input:  $\mathbf{x} = [x_1, x_2]^T$ , Output:  $\mathbf{C} \in \{0, 1\}$
- Prediction:  
choose  $\begin{cases} \mathbf{C} = 1 & \text{if } P(\mathbf{C} = 1 \mid x_1, x_2) > 0.5 \\ \mathbf{C} = 0 & \text{otherwise} \end{cases}$   
or equivalently  
choose  $\begin{cases} \mathbf{C} = 1 & \text{if } P(\mathbf{C} = 1 \mid x_1, x_2) > P(\mathbf{C} = 0 \mid x_1, x_2) \\ \mathbf{C} = 0 & \text{otherwise} \end{cases}$
- Training == build model of this from data



## *Bayes' Rule: $K > 2$ Classes*

$$\begin{aligned} P(C_i | \mathbf{x}) &= \frac{p(\mathbf{x} | C_i)P(C_i)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x} | C_i)P(C_i)}{\sum_{k=1}^K p(\mathbf{x} | C_k)P(C_k)} \end{aligned}$$

$$P(C_i) \geq 0 \text{ and } \sum_{i=1}^K P(C_i) = 1$$

choose  $C_i$  if  $P(C_i | \mathbf{x}) = \max_k P(C_k | \mathbf{x})$   
Bayes' Classifier – choose class with the highest  
Posterior probability.

# Losses and Risks

- General ways of measuring performance/error
- Actions:  $\alpha_i$  (e.g. assign class  $C_i$  to some input)
- Loss of  $\alpha_i$  when the state is  $C_k : \lambda_{ik}$
- Expected risk (Duda and Hart, 1973) (for taking action  $\alpha_i$ )

$$R(\alpha_i | \mathbf{x}) = \sum_{k=1}^K \lambda_{ik} P(C_k | \mathbf{x})$$

choose  $\alpha_i$  if  $R(\alpha_i | \mathbf{x}) = \min_k R(\alpha_k | \mathbf{x})$

## *Losses and Risks: 0/1 Loss*

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{k=1}^K \lambda_{ik} P(C_k | \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k | \mathbf{x}) \\ &= 1 - P(C_i | \mathbf{x}) \end{aligned}$$

*For minimum risk, choose the most probable class*

## *Losses and Risks: Reject*

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K + 1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$R(\alpha_{K+1} | \mathbf{x}) = \sum_{k=1}^K \lambda P(C_k | \mathbf{x}) = \lambda$$

$$R(\alpha_i | \mathbf{x}) = \sum_{k \neq i} P(C_k | \mathbf{x}) = 1 - P(C_i | \mathbf{x})$$

choose  $C_i$  if  $P(C_i | \mathbf{x}) > P(C_k | \mathbf{x}) \forall k \neq i$  and  $P(C_i | \mathbf{x}) > 1 - \lambda$   
reject otherwise

# Discriminant Functions

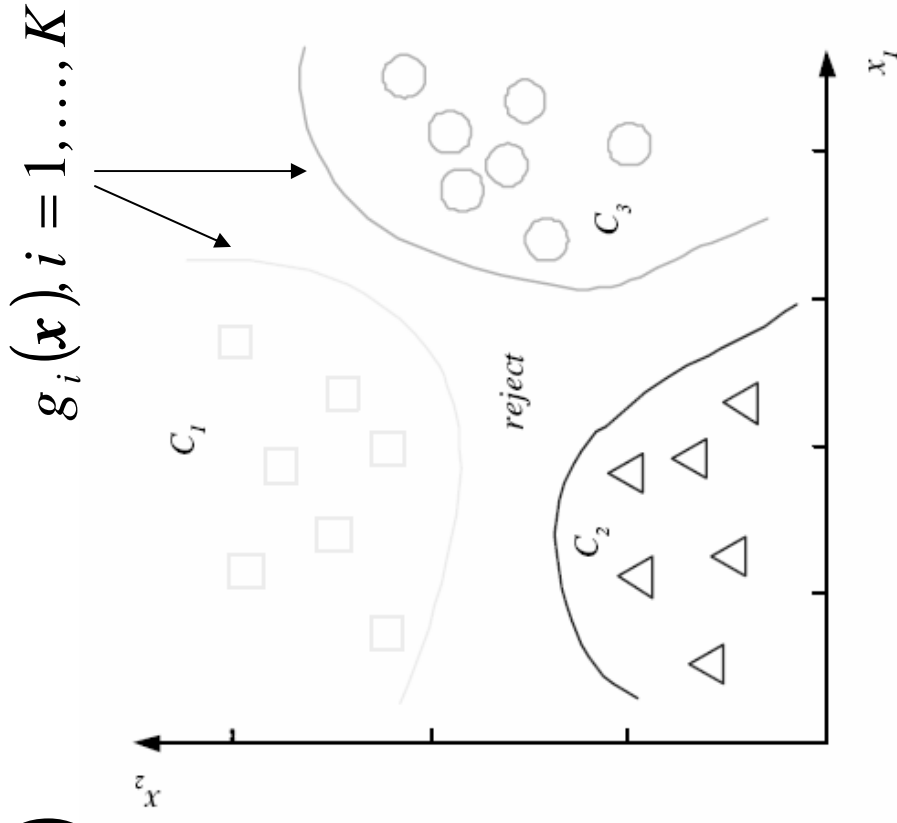
choose  $C_i$  if  $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i | \mathbf{x}) \\ P(C_i | \mathbf{x}) \\ p(\mathbf{x} | C_i)P(C_i) \end{cases}$$

(ok for 0/1 loss)

$K$  decision regions  $R_1, \dots, R_K$

$$R_i = \{\mathbf{x} | g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$



# $K=2$ Classes

- Dichotomizer ( $K=2$ ) vs Polychotomizer ( $K>2$ )

- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$

choose  $\begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$

- *Log odds:*

$$\log \frac{P(C_1 | \mathbf{x})}{P(C_2 | \mathbf{x})}$$

# Utility Theory

- (given) Prob of state  $k$  given evidence  $\mathbf{x}$ :  $P(S_k|\mathbf{x})$
- (and) Utility of  $\alpha_i$  when state is  $k$ :  $U_{ik}$
- (then) Expected utility:  
$$EU(\alpha_i | \mathbf{x}) = \sum_k U_{ik} P(S_k | \mathbf{x})$$
  
Choose  $\alpha_j$  if  $EU(\alpha_i | \mathbf{x}) = \max_j EU(\alpha_j | \mathbf{x})$
- i.e a rational choice – maximize expected utility (usually equivalent to minimizing expected risk).

# *Value of Information*

- (e.g. when to incorporate new features – do a blood test  $v$ 's take pulse)
- Expected utility using  $\mathbf{x}$  only

$$EU(\mathbf{x}) = \max_i \sum_k U_{ik} P(S_k | \mathbf{x})$$

- Expected utility using  $\mathbf{x}$  and new feature  $z$

$$EU(\mathbf{x}, z) = \max_i \sum_k U_{ik} P(S_k | \mathbf{x}, z)$$

- $z$  is useful if  $EU(\mathbf{x}, z) > EU(\mathbf{x})$
- Difference between two is value of new information minus extra “complexity” of adding  $z$



# *Bayesian Networks*

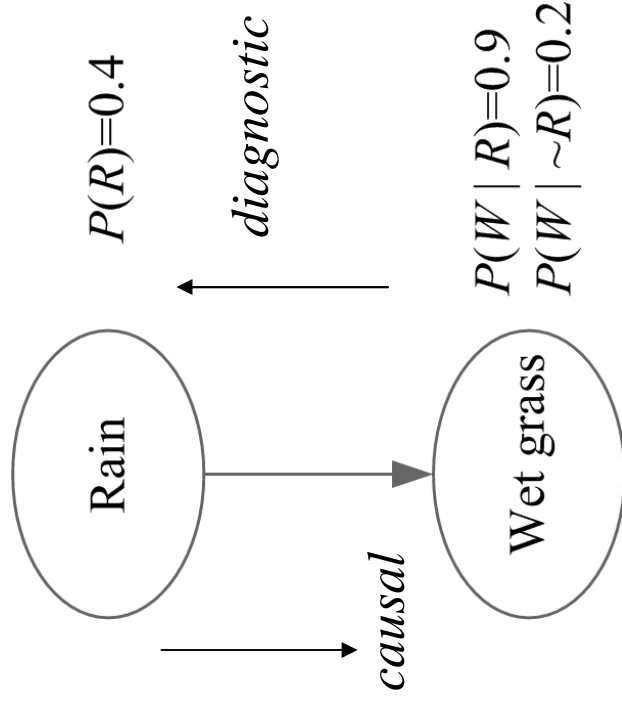
- Aka graphical models, probabilistic networks
- Nodes are hypotheses (random vars) and the prob corresponds to our belief in the truth of the hypothesis
- Arcs intuitively represent direct influences between hypotheses
  - Much easier for a domain expert to specify these influences than specific probabilities.
- The structure is represented as a directed acyclic graph (DAG)
- The parameters are the conditional probs in the arcs
- (Pearl, 1988, 2000; Jensen, 1996; Lauritzen, 1996)

- Used to capture uncertain knowledge in a natural and efficient way.
  - ...and to make inferences about data.
- A Bayesian network is a complete representation of the domain – i.e. a simplification of the joint distribution

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

- In general we can write the joint distribution as:
- Then the absence of arcs indicates conditional independence assumptions (i.e some node is not **directly** influenced by some other node).

# *Causes and Bayes' Rule*



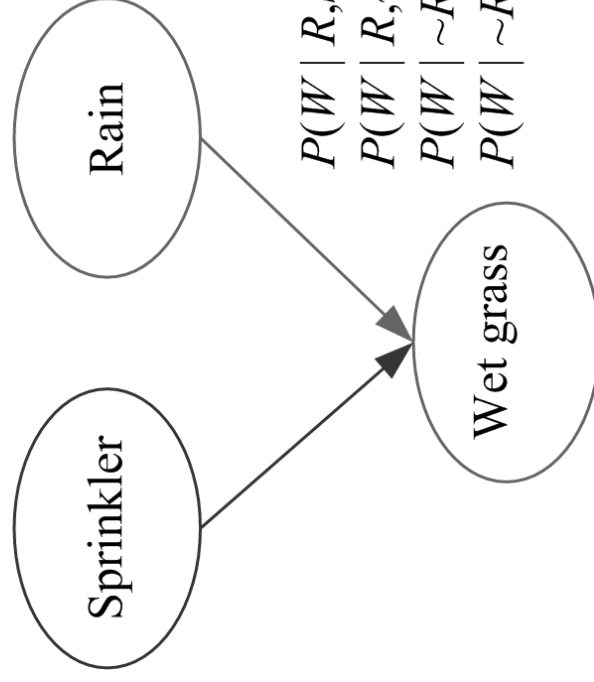
*Diagnostic inference:*  
*Knowing that the grass is wet,*  
*what is the probability that rain is*  
*the cause?*

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

# Causal vs Diagnostic Inference

$$P(S)=0.2$$

$$P(R)=0.4$$



$$P(W | R, S)=0.95$$

$$P(W | R, \sim S)=0.90$$

$$P(W | \sim R, S)=0.90$$

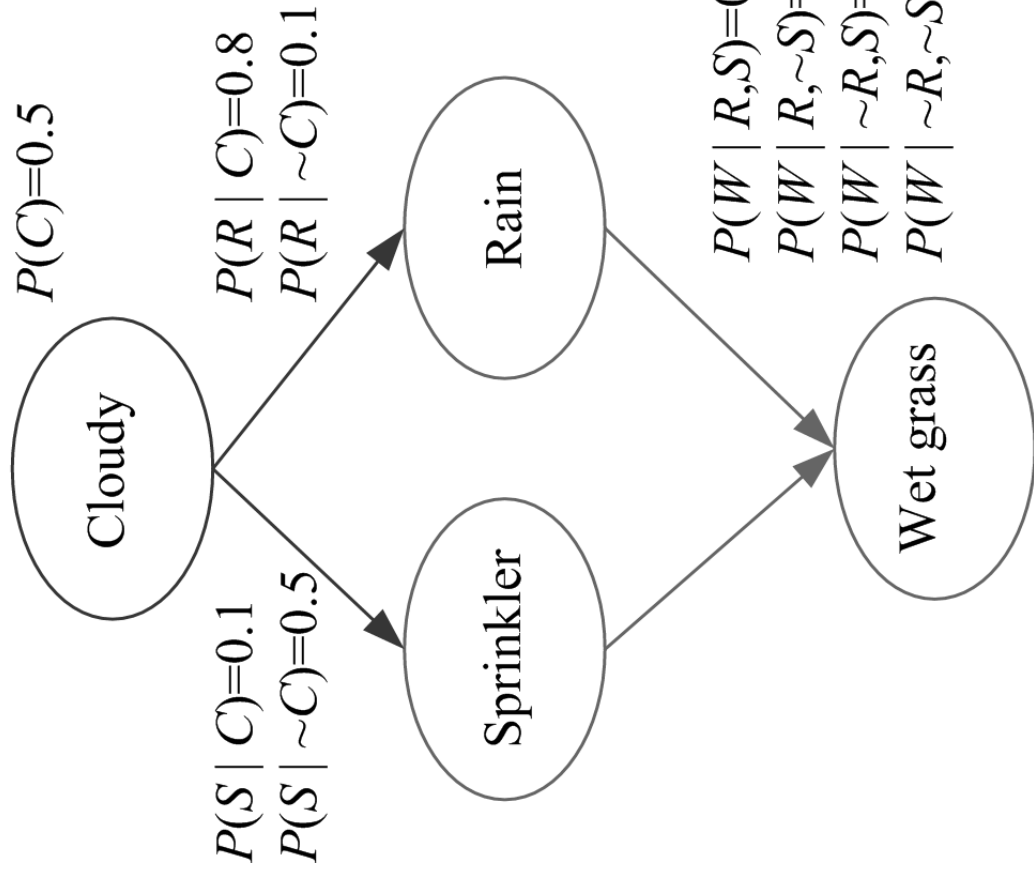
$$P(W | \sim R, \sim S)=0.10$$

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

$$\begin{aligned} P(W|S) &= P(W|R,S) P(R|S) + \\ &\quad P(W|\sim R,S) P(\sim R|S) \\ &= P(W|R,S) P(R) + \\ &\quad P(W|\sim R,S) P(\sim R) \\ &= 0.95 \cdot 0.4 + 0.9 \cdot 0.6 = 0.92 \end{aligned}$$

*Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on?  $P(S|W) = 0.35 > 0.2 P(S)$*   
*Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

# Bayesian Networks: Causes



*Causal inference:*

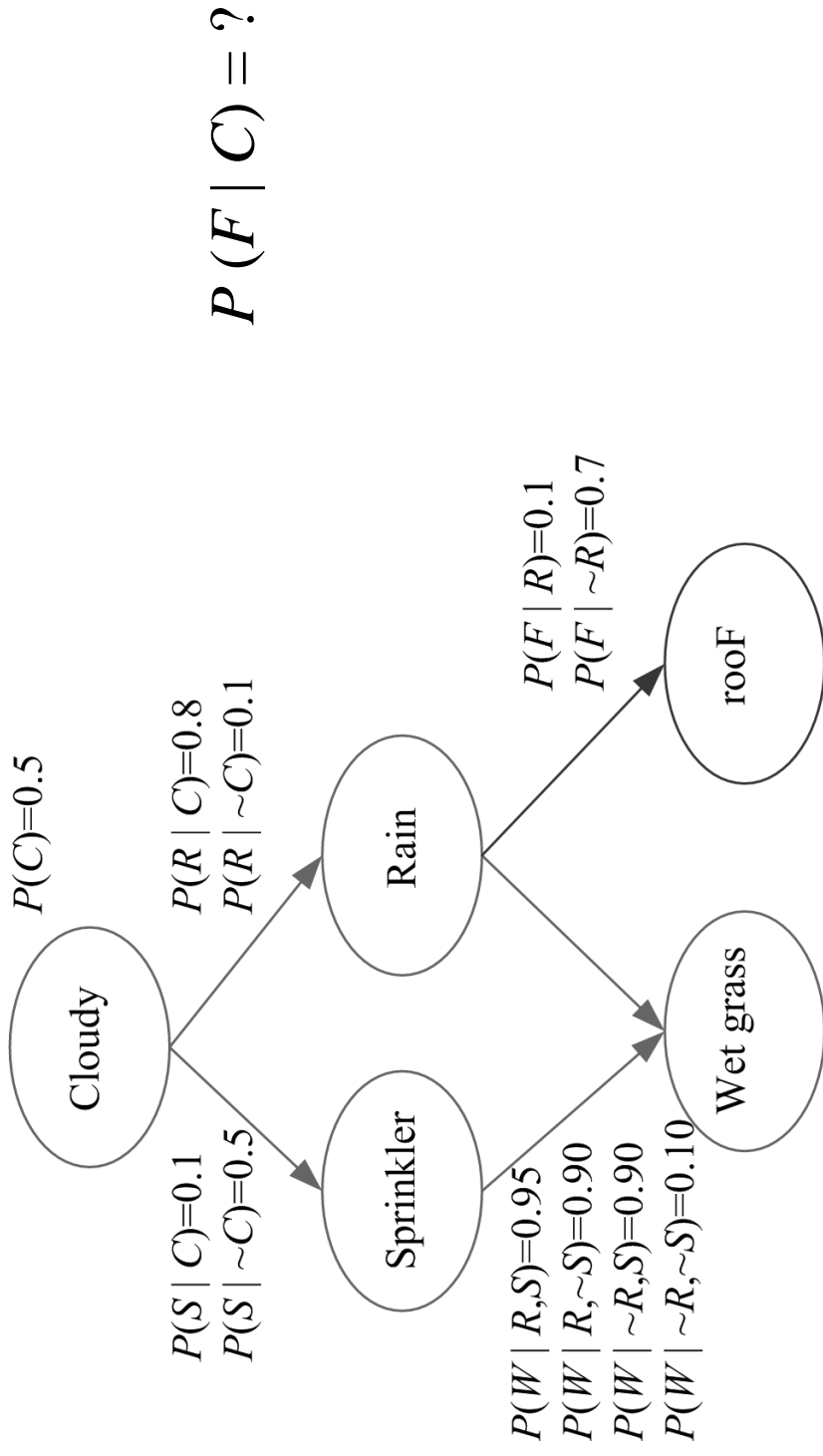
$$P(W|C) = P(W|R,S) P(R,S|C) + P(W|\sim R,S) P(\sim R,S|C) + P(W|R,\sim S) P(R,\sim S|C) + P(W|\sim R,\sim S) P(\sim R,\sim S|C)$$

*and use the fact that*

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

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


# Bayesian Nets: Local structure



$$P(F | C) = ?$$

$$P(C, S, R, W, F) = P(C)P(S | C)P(R | C)P(W | S, R)P(F | R)$$

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

- 
- A Bayesian network can be far more compact than the full joint distribution
    - Example of a locally structured/sparse system.
  - If we assume in a Bayesian network that each node is influenced by  $k$  others, and we have a total of  $n$  variables (all Boolean):
    - Each conditional probability table will be  $2^k$  numbers, and total network will require  $n2^k$  numbers.
    - Joint distribution will need  $2^n$  numbers.
    - E.g (for  $n=30$ ,  $k=5$ ): 960 v's over a billion!



## *Bayesian Networks: Inference*

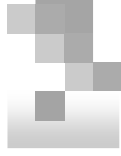
$$P(C, S, R, W, F) = P(C) P(S|C) P(R|C) P(W|R, S) P(F|R)$$

$$P(C, F) = \sum_S \sum_R \sum_W P(C, S, R, W, F)$$

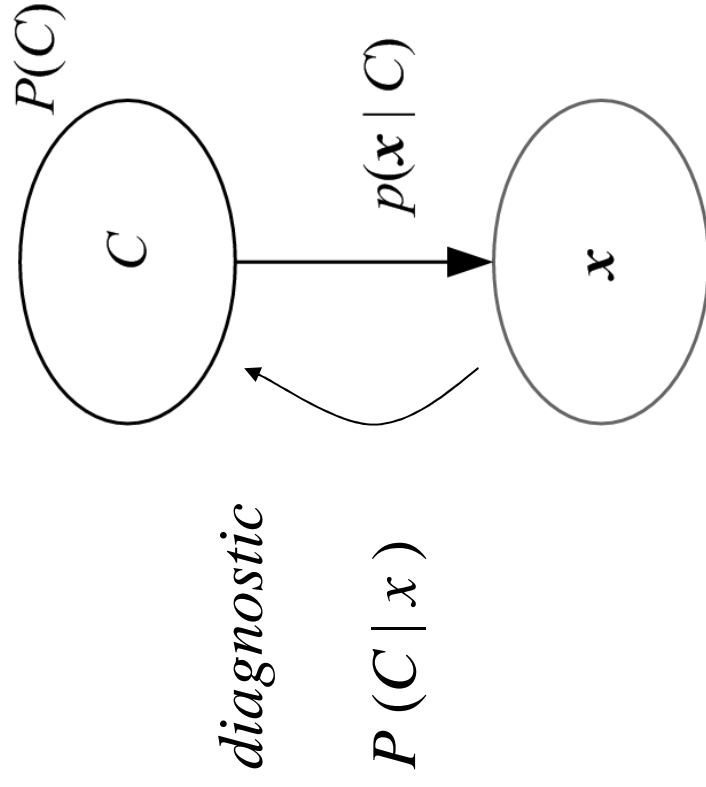
$$P(F|C) = P(C, F) / P(C) \quad \textit{Not efficient!}$$

Belief propagation (Pearl, 1988)

Junction trees (Lauritzen and Spiegelhalter, 1988)



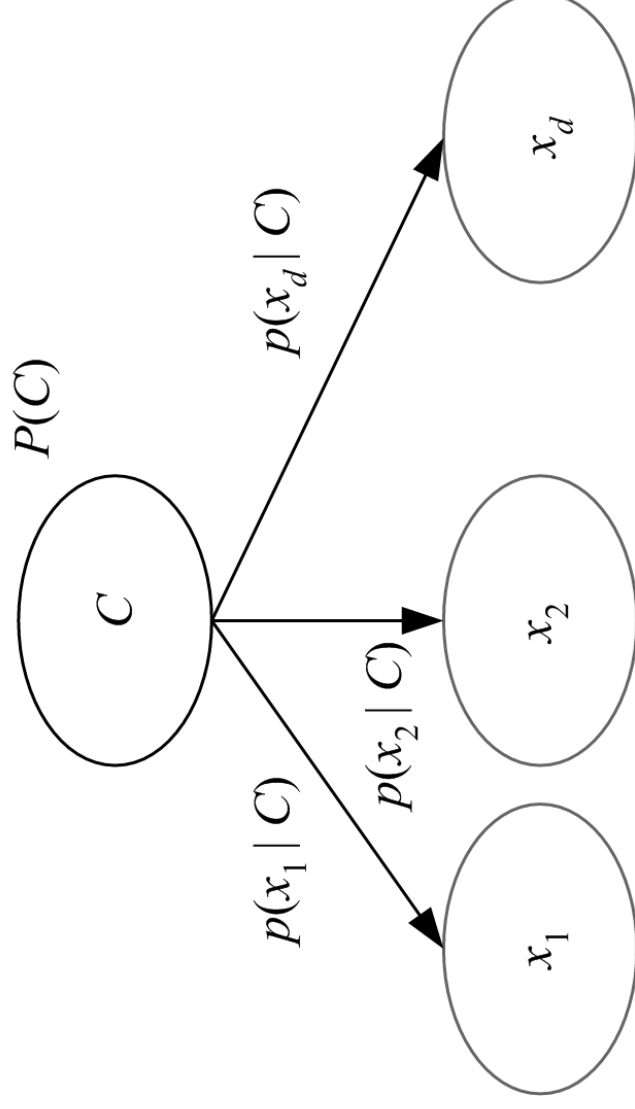
# Bayesian Networks: Classification



Bayes' rule inverts the arc:

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

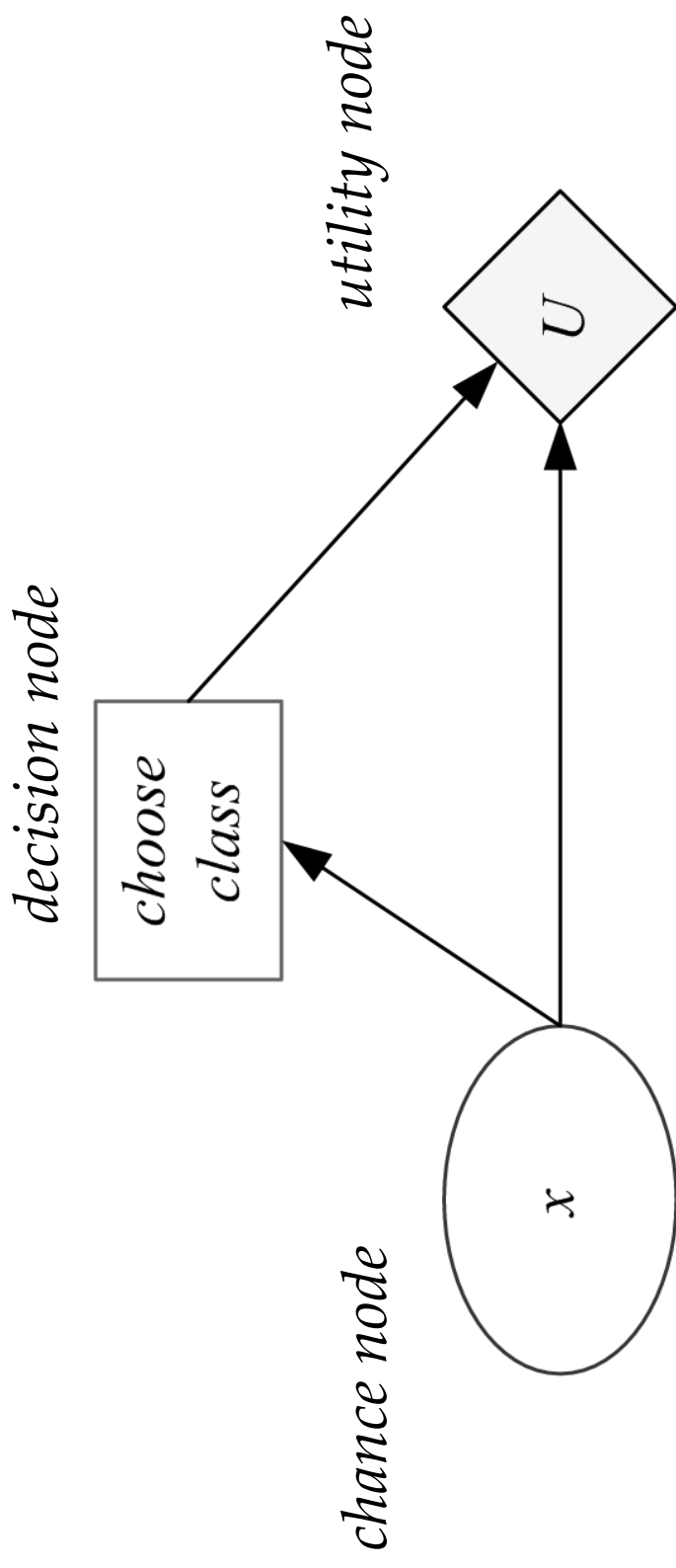
# Naive Bayes' Classifier



Given  $C$ ,  $x_j$  are independent:

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

# *Influence Diagrams*



# Association Rules

- Association rule:  $X \rightarrow Y$
- Support ( $X \rightarrow Y$ ):

$$P(X, Y) = \frac{\# \{ \text{customers who bought } X \text{ and } Y \}}{\# \{ \text{customers} \}}$$

- Confidence ( $X \rightarrow Y$ ):

$$P(Y | X) = \frac{P(X, Y)}{P(X)}$$
$$= \frac{\# \{ \text{customers who bought } X \text{ and } Y \}}{\# \{ \text{customers who bought } X \}}$$

Apriori algorithm (Agrawal et al., 1996)