CSCE 970 Lecture 4: Introduction to Bayesian Networks

Stephen D. Scott

Introduction

- Shifting now from sequential data to single (non-sequential) fixed length feature vectors

- E.g. each vector represents a medical patient and the vector's components (features) correspond to results of particular medical tests

- Common problem: given a data set of training vectors, infer a model for the entire space of possible vectors

  – Will use this model to make predictions on new (previously unseen) instances

  – Similar to HMMs, except no sequential nature

Introduction
(cont'd)

- Many ways to approach this; we'll focus on developing probabilistic models via Bayesian networks
  - Model joint probability distributions by decomposing them into conditional probabilities
  - Algorithms can determine the probability of certain attribute values of a feature vector given others

Outline

- Preliminaries
- Naïve Bayes learning
- Introduction to Bayesian networks

Preliminaries
Probability

- Given a set $\Omega = \{e_1, \ldots, e_n\}$ of elements, a function $P(\cdot)$ that assigns a real number $P(E)$ to each event $E \subseteq \Omega$ is a probability function if

  1. $0 \leq P(\{e_i\}) \leq 1$ for all $i \in \{1, \ldots, n\}$
  2. $\sum_{i=1}^{n} P(\{e_i\}) = 1$
  3. For each event $E = \{e_{i_1}, e_{i_2}, \ldots, e_{i_k}\}$ such that $|E| \neq 1$,

$$P(E) = \sum_{j=1}^{k} P(\{e_{i_j}\})$$

- Given such a probability space, a random variable is a function on Ω

Preliminaries

Probability (Example 1.7)

- Let Ω contain all outcomes of a throw of a pair of fair dice:

$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(6,5),(6,6)\}$$

- Let RV X be the sum of each ordered pair and $Y =$ "odd" if both dice read odd numbers and "even" otherwise:

e	$X(e)$	$Y(e)$
$(1,1)$	2	odd
$(1,2)$	3	even
$\vdots$	$\vdots$	$\vdots$
$(6,6)$	12	even

- Then $X = 3$ represents event $\{(1,2),(2,1)\}$ and $P(X = 3) = 1/18$

- Uppercase letters ("X") represent RVs and lowercase ("x") represent specific values

Preliminaries

Joint Distributions

- In previous example, X ranged over the integers 2–12 and Y ranged over $\{$odd,even$\}$
  - Each value in each range had its own probability
- If we consider joint events (one from X's range, one from Y's) we get a joint probability distribution $P(x, y) = P(X = x, Y = y)$
- E.g. $x = 4$ and $y =$ odd represents the event $\{(1, 3), (3, 1)\}$ and $P(x, y) = 1/18$

Preliminaries
Marginal Probability

- If we have a handle on a joint distribution, we can sum across values of an RV to get the marginal probability distribution of another RV

- For two RVs X and Y,

$$P(X = x) = \sum_y P(X = x, Y = y)$$

- E.g.

$$\begin{aligned} P(X = 4) &= \sum_y P(X = 4, Y = y) \\ &= P(X = 4, Y = \text{odd}) + P(X = 4, Y = \text{even}) \\ &= 1/18 + 1/36 = 1/12 \end{aligned}$$

- Also see Example 1.15

Preliminaries

Conditional Probability

- Let E and F be events with $P(F) > 0$

- The conditional probability of E given F is

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

- E.g. if $x = 6$ and $y =$ even then

$$P(X = x) =$$

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

$$P(X = x \mid Y = y) =$$

Preliminaries

Bayes' Theorem

- An identity for conditional probabilities

- Given two events E and F with $P(E), P(F) > 0$

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)}$$

(Way to remember: the event named after the line goes in the denominator)

- E.g. When $x = 6$ and $y =$ even,

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} = \frac{(2/5)(5/36)}{27/36} = 2/27$$

Preliminaries

Independence of Events

- Two events E and F are independent if one of the following holds:

  1. $P(E \mid F) = P(E)$ and $P(E), P(F) \neq 0$
     (can switch roles of E and F for same result)

  2. $P(E) = 0$ or $P(F) = 0$

- E and F are independent iff $P(E \cap F) = P(E)P(F)$

- E.g. is the event $X = 6$ independent of $Y =$ even?

- Is the event $X = 10 \cup X = 12$ independent of $Y =$ odd?

Preliminaries

Conditional Independence of Events

- Can also have independence conditioned on other variables

- Events E and F are conditionally independent given G if $P(G) > 0$ and one of the following holds

  1. $P(E \mid F \cap G) = P(E \mid G)$ and $P(E \mid G), P(F \mid G) > 0$

  2. $P(E \mid G) = 0$ or $P(F \mid G) = 0$

Preliminaries

Conditional Independence of Events

Example

- Define third RV Z, defined as the product of the two dice results

$$P(X = 5 \mid Y = \text{even}) = \frac{4/36}{27/36} = 4/27 \neq 4/36 = P(X = 5)$$

$$P(X = 5 \mid Y = \text{even} \cap Z = 4) = \frac{2/36}{3/36} = 2/3 = P(X = 5 \mid Z = 4)$$

- Thus the event $X = 5$ is not independent of $Y =$ even, but is conditionally independent of it given $Z = 4$

Preliminaries

Independence of Random Variables

- Given probability space (Ω, P), two RVs A and B are independent (written $I_P(A, B)$) if, for all values a of A and b of B, the events $A = a$ and $B = b$ are independent

- I.e. for all values a and b, either $P(a) = 0$ or $P(b) = 0$ or $P(a \mid b) = P(a)$

- Generalizes to sets of RVs

Preliminaries

Independence of Random Variables

Example 1.16

$\Omega =$ set of all cards in a deck, P uniform

Variable	Values	Outcomes
R	$\{r1, r2\}$	royal/nonroyal cards
T	$\{t1, t2\}$	tens & jacks/not t & j
S	$\{s1, s2\}$	spades/nonspades

s	r	t	$P(r, t \mid s)$	$P(r, t)$
$s1$	$r1$	$t1$	$1/13$	$4/52 = 1/13$
$s1$	$r1$	$t2$	$2/13$	$8/52 = 2/13$
$s1$	$r2$	$t1$	$1/13$	$4/52 = 1/13$
$s1$	$r2$	$t2$	$9/13$	$36/52 = 9/13$
$s2$	$r1$	$t1$	$3/39 = 1/13$	$4/52 = 1/13$
$s2$	$r1$	$t2$	$6/39 = 2/13$	$8/52 = 2/13$
$s2$	$r2$	$t1$	$3/39 = 1/13$	$4/52 = 1/13$
$s2$	$r2$	$t2$	$27/39 = 9/13$	$36/52 = 9/13$

Thus $P(r, t \mid s) = P(r, t) \Rightarrow I_P(\{R, T\}, \{S\})$

Preliminaries

Conditional Independence of Random Variables

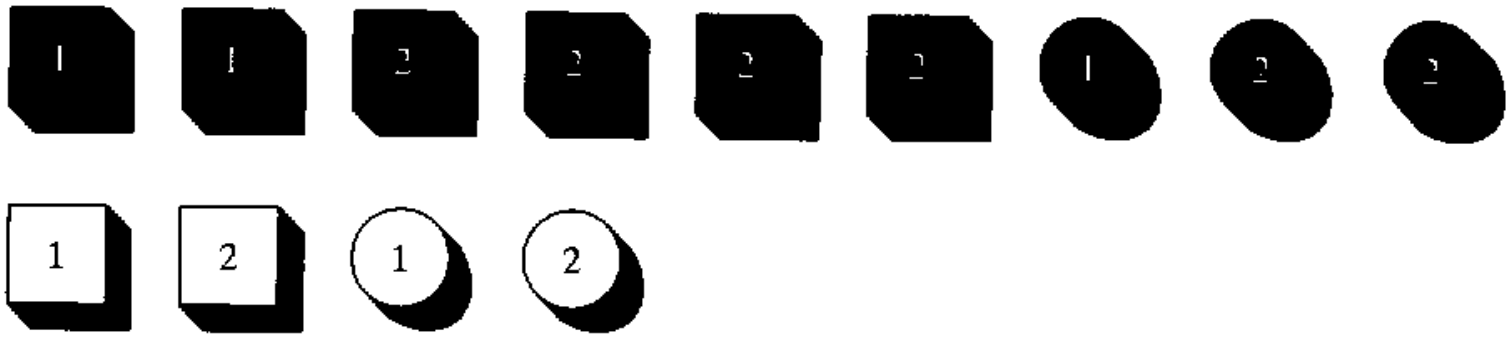
- Given probability space (Ω, P), two RVs A and B are conditionally independent given C (written $I_P(A, B \mid C)$) if, for all values a of A, b of B, and c of C, the events $A = a$ and $B = b$ are conditionally independent given event $C = c$

- I.e. for all values a and b and c, either $P(a \mid c) = 0$ or $P(b \mid c) = 0$ or $P(a \mid b, c) = P(a \mid c)$

- Generalizes to sets of RVs

Preliminaries

Conditional Independence of Random Variables, Example 1.17



P is uniform

Var	Values	Outcomes
V	$\{v1, v2\}$	obj with "1"/"2"
S	$\{s1, s2\}$	square/round
C	$\{c1, c2\}$	black/white

c	s	v	$P(v \mid s, c)$	$P(v \mid c)$
$c1$	$s1$	$v1$	$1/3$	$3/9 = 1/3$
$c1$	$s1$	$v2$	$2/3$	$6/9 = 2/3$
$c1$	$s2$	$v1$	$1/3$	$3/9 = 1/3$
$c1$	$s2$	$v2$	$2/3$	$6/9 = 2/3$
$c2$	$s1$	$v1$	$1/2$	$2/4 = 1/2$
$c2$	$s1$	$v2$	$1/2$	$2/4 = 1/2$
$c2$	$s2$	$v1$	$1/2$	$2/4 = 1/2$
$c2$	$s2$	$v2$	$1/2$	$2/4 = 1/2$

Thus $P(v \mid s, c) = P(v \mid c) \Rightarrow I_P(\{V\}, \{S\} \mid \{C\})$

Basic Formulas for Probabilities

- Product Rule: probability $P(A \cap B)$ of conjunction of events A and B:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Sum Rule: probability of a disjunction of two events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Theorem of total probability: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

- If X takes on real values, then its expected value is

$$E(X) = \sum_{x} xP(x)$$

Naïve Bayes Classification

- Naïve Bayes classifiers are like Bayesian networks taken to the extreme in their conditional independence assumption

- Generally, the assumption is so unrealistic that NB is ineffective in predicting probabilities

- Still good at classification, however

- Successfully applied to text classification, diagnosis

Naïve Bayes Classification

(cont'd)

- Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2, \ldots, a_n \rangle$
- Most probable value of $f(x)$ is:

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\text{argmax}}\ P(v_j \mid a_1, a_2, \ldots, a_n) \\ &= \underset{v_j \in V}{\text{argmax}}\ \frac{P(a_1, a_2, \ldots, a_n \mid v_j)\, P(v_j)}{P(a_1, a_2, \ldots, a_n)} \\ &= \underset{v_j \in V}{\text{argmax}}\ P(a_1, a_2, \ldots, a_n \mid v_j)\, P(v_j) \end{aligned}$$

  (Second equality comes from where?)

- Thus all we have to do is model the joint distribution over the attributes conditioned on the labels
- Can we just frequency count our way out of this?

Naïve Bayes Classification
(cont'd)

- Problem with estimating probs from training data: estimating $P(v_j)$ easily done by counting, but there are exponentially (in n) many combs. of values of $a_1, \ldots, a_n$, so can't get estimates for most combs

- Naïve Bayes assumption:

$$P(a_1, a_2, \ldots, a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

  so naïve Bayes classifier:

$$v_{NB} = \underset{v_j \in V}{\text{argmax}}\ P(v_j) \prod_i P(a_i \mid v_j)$$

- Now have only polynomial number of probs to estimate

Naïve Bayes Algorithm

Naïve_Bayes_Learn

1. For each target value v_j

   (a) $\hat{P}(v_j) \leftarrow$ estimate $P(v_j) =$ fraction of exs with v_j

   (b) For each attribute value a_i of each attrib a

      i. $\hat{P}(a_i \mid v_j) \leftarrow$ estimate $P(a_i \mid v_j) =$ fraction of v_j-labeled exs with a_i

Classify_New_Instance(x)

$$v_{NB} = \underset{v_j \in V}{\text{argmax}}\ \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i \mid v_j)$$

Naïve Bayes Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example to classify: $\langle Outlk = sun, Temp = cool, Humid = high, Wind = strong \rangle$

Assign label $v_{NB} = \text{argmax}_{v_j \in V} \; P(v_j) \prod_i P(a_i \mid v_j)$

$$P(y) \cdot P(sun \mid y) \cdot P(cool \mid y) \cdot P(high \mid y) \cdot P(strong \mid y)$$
$$= (9/14) \cdot (2/9) \cdot (3/9) \cdot (3/9) \cdot (3/9) = 0.0053$$

$$P(n) \; P(sun \mid n) \; P(cool \mid n) \; P(high \mid n) \; P(strong \mid n)$$
$$= (5/14) \cdot (3/5) \cdot (1/5) \cdot (4/5) \cdot (3/5) = 0.0206$$

So $v_{NB} = n$

Naïve Bayes

Subtleties

- Conditional independence assumption is often violated, i.e.

$$P(a_1, a_2, \ldots, a_n \mid v_j) \neq \prod_i P(a_i \mid v_j)$$

  . . . but it works surprisingly well anyway. Note don't need estimated posteriors $\hat{P}(v_j \mid x)$ to be correct; need only that

$$\underset{v_j \in V}{\text{argmax}}\ \hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j) = \underset{v_j \in V}{\text{argmax}}\ P(v_j) P(a_1, \ldots, a_n \mid v_j)$$

- Sufficient conditions given in [Domingos & Pazzani, 1996]

- But not really trustworthy for probability estimates!

Bayesian Belief Networks

- Sometimes naïve Bayes assumption of conditional independence too restrictive

- But inferring probabilities is intractable without some such assumptions

- Bayesian belief networks (also called Bayes Nets) describe conditional independence among subsets of variables

- Allows combining prior knowledge about dependencies among variables with observed training data

Bayesian Belief Networks

Directed Acyclic Graphs

- A graph $G = (V, E)$ consists of a set of vertices V, which are connected to each other with edges from a set E

- A directed graph is a graph in which each edge (x, y) is an ordered pair, with direction from its head x to its tail y

  - x is y's parent

- A directed acyclic graph (DAG) G is a directed graph where there is no path from a node to itself

  - If there's a path from x to y, then y is a descendent of x and x is an ancestor of y

Bayesian Belief Networks

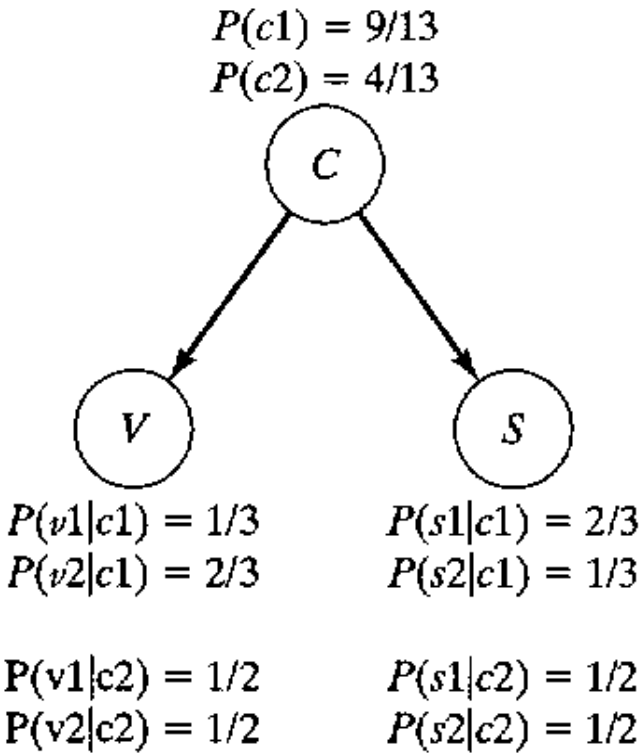
The Markov Property

- Consider a joint probability distribution P and a DAG $G = (V, E)$. (G, P) satisfies the Markov condition if for each RV $X \in V$, the set $\{X\}$ is conditionally independent of the set of its nondescendents given the set of its parents, i.e. if PA_X is the set of parents and ND_X nondescendents, then $I_P(\{X\}, ND_X \mid PA_X)$

- If (G, P) satisifes the Markov condition, then (G, P) is a Bayesian network

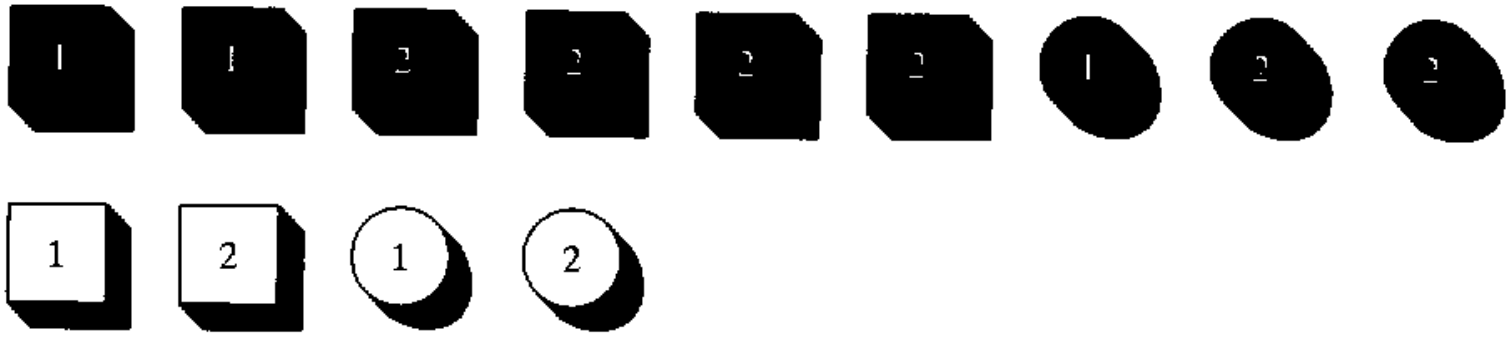
Bayesian Belief Networks

Each node in the DAG corresponds to a RV, and has a probability distribution on that RV conditioned on its parents



Bayesian Belief Networks

Example 1.29



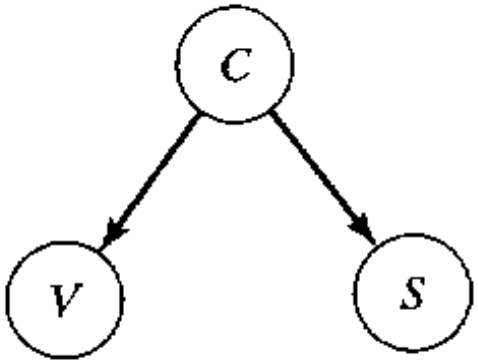
P is uniform

Var	Values	Outcomes
V	$\{v1, v2\}$	obj with “1”/“2”
S	$\{s1, s2\}$	square/round
C	$\{c1, c2\}$	black/white

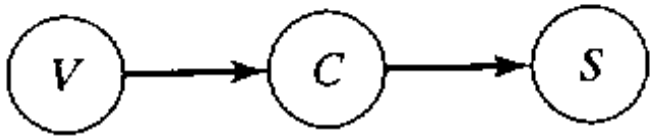
We already showed that $I_P(\{V\}, \{S\} \mid \{C\})$. Which of the following DAGs make a Bayes net with P?

Bayesian Belief Networks

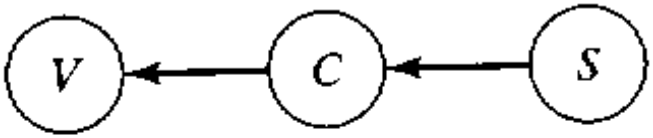
Example 1.29 (cont'd)



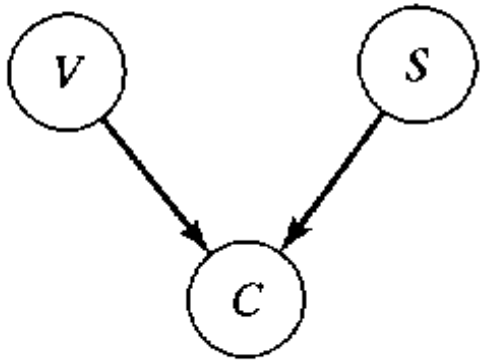
(a)



(b)



(c)



(d)

Bayesian Belief Networks

Example 1.29 (cont'd)

(a) V's conditional probability distribution depends on only C. When C is known, then V's distribution depends on no other variables (similarly for S)

(b) V's distribution depends on nothing; S depends on only C. When C is fixed, then S depends on nothing.

(c) Same as (b).

Bayesian Belief Networks

Example 1.29 (cont'd)

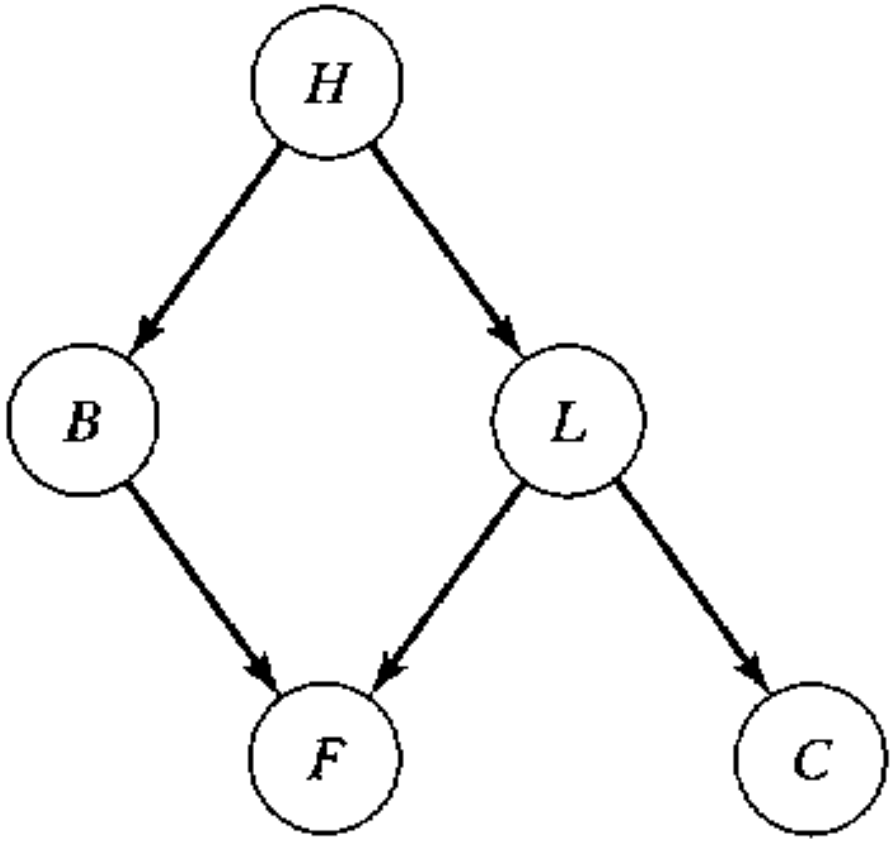
(d) When C is unknown, then V and S are independent, and C's distribution depends on V and S. But say that e.g. $V \in \{0, 1\}$ indicates whether a car's battery is dead or alive, $S \in \{0, 1\}$ indicates if a car's tank is empty or full, and $C \in \{0, 1\}$ indicates whether a gas guage reads empty or full.

- V and S are independent if C unknown
- Knowing C suddenly relates V and S since e.g. $V = 0$ influences the probability that $S = 0$
- We'll discuss this more later

Bayesian Belief Networks

Team Exercise

What are the conditional independencies in a distribution P if (G, P) is a Bayes net with the following graph G?



Bayesian Belief Networks

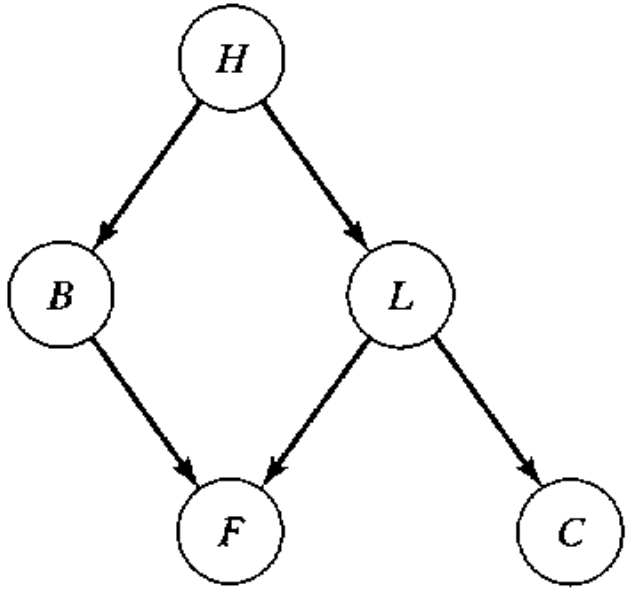
Factorization of a Joint Distribution

- We already discussed the problems with directly estimating a joint distribution
  - Exponential number of combinations of attribute values makes it impossible to get enough training data to estimate the distribution
  - Also, the need to sum over all combinations of values makes marginalizing intractable
- Markov condition simplifies this problem by allowing factorization of the joint distribution

Bayesian Belief Networks

Factorization of a Joint Distribution

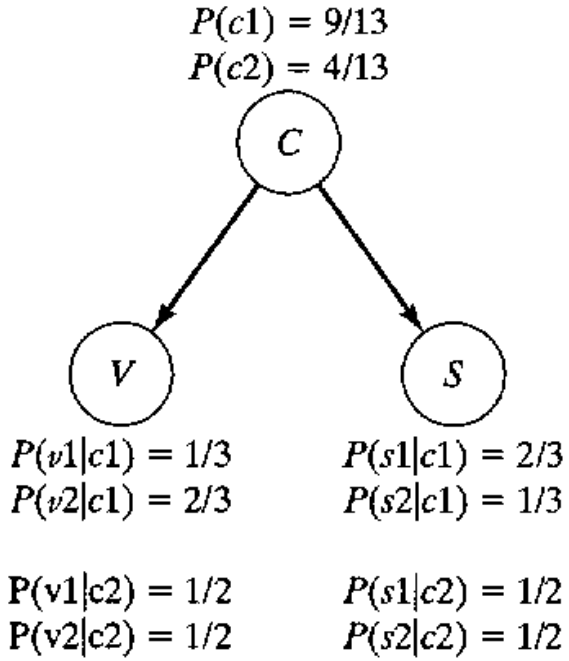
- Theorem 1.4: If (G, P) satisfies the Markov condition, then P equals the product of its conditional distributions of allnodes given values of their parents (when they exist)
- E.g. $P(f, c, b, \ell, h) = P(f \mid b, \ell)P(c \mid \ell)P(b \mid h)P(\ell \mid h)P(h)$



Can estimate each conditional probability separately

Bayesian Belief Networks

Factorization of a Joint Distribution (example)



$$P(v, s, c) = P(v \mid c)P(s \mid c)P(c)$$

Earlier we showed $P(v1, s1, c1) = 2/13$. Factorization yields

$$P(v1 \mid c1)P(s1 \mid c1)P(c1) = (1/3)(2/3)(9/13) = 2/13$$

Also works for DAGs (b) and (c)

Bayesian Belief Networks

Generalization of Naïve Bayes

Now it's obvious how Bayes nets generalize naïve Bayes.

How?

Bayesian Belief Networks

Starting with the DAG

- The process also works in reverse
  - Start with a DAG $G = (V, E)$ where each node in V is a RV with a discrete conditional distribution
  - Then the joint distribution P that comes from multiplying the conditional distributions satisfies the Markov condition with G
- This is how we'll typically work: define local conditional distributions with a DAG and then analyze the resultant joint distribution
- Also works with some continuous distributions, e.g. Gaussian

Bayesian Belief Networks

Starting with the DAG (example)

- $H =$ smoking history, $B =$ bronchitis, $L =$ lung cancer, $F =$ fatigue, $C =$ chest X-ray result

- Scientific studies and experts' opinions define conditional distribs