Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Bayesian Networks

Computing with Probabilities: Law of Total Probability

Law of Total Probability (aka "summing out" or marginalization)

$P(a) = \Sigma_b\ P(a, b)$

$= \Sigma_b\ P(a \mid b)\ P(b)$ where B is any random variable

Why is this useful?

given a joint distribution (e.g., P(a,b,c,d)) we can obtain any "marginal" probability (e.g., P(b)) by summing out the other variables, e.g.,

$P(b) = \Sigma_a \Sigma_c \Sigma_d\ P(a, b, c, d)$

Computing with Probabilities: Law of Total Probability

Example

- In a certain county
  - · 60% of registered voters are Republicans
  - · 30% are Democrats
  - · 10% are Independents.
- When those voters were asked about increasing military spending
  - · 40% of Republicans opposed it
  - · 65% of the Democrats opposed it
  - · 55% of the Independents opposed it.
- What is the probability that a randomly selected voter in this county opposes increased military spending?

Computing with Probabilities: Law of Total Probability

- $\Omega = \{\text{registered voters in the county}\}$
- $R = \{\text{registered republicans}\}$, $\Pr(R) = 0.6$
- $D = \{\text{registered democrats}\}$, $\Pr(D) = 0.3$
- $I = \{\text{registered independents}\}$, $\Pr(I) = 0.1$
- $B = \{\text{registered voters opposing increased military spending}\}$
- $\Pr(B|R) = 0.4$, $\Pr(B|D) = 0.65$, $\Pr(B|I) = 0.55$.

By the total probability theorem:

$\Pr(B)$

Computing with Probabilities: Law of Total Probability

- $\Omega = \{\text{registered voters in the county}\}$
- $R = \{\text{registered republicans}\}$, $\Pr(R) = 0.6$
- $D = \{\text{registered democrats}\}$, $\Pr(D) = 0.3$
- $I = \{\text{registered independents}\}$, $\Pr(I) = 0.1$
- $B = \{\text{registered voters opposing increased military spending}\}$
- $\Pr(B|R) = 0.4$, $\Pr(B|D) = 0.65$, $\Pr(B|I) = 0.55$.

By the total probability theorem:

$$\Pr(B) = \Pr(B|R)\Pr(R) + \Pr(B|D)\Pr(D) + \Pr(B|I)\Pr(I)$$
$$= (0.4 \cdot 0.6) + (0.65 \cdot 0.3) + (0.55 \cdot 0.1) = 0.49.$$

Computing with Probabilities: Law of Total Probability

Less obvious: we can also compute <u>any conditional probability of interest</u> given a joint distribution, e.g.,

$$P(c \mid b) = \Sigma_a \Sigma_d P(a, c, d \mid b)$$
$$= 1 / P(b) \;\; \Sigma_a \Sigma_d P(a, c, d, b)$$

where 1 / P(b) is just a normalization constant

Thus, the joint distribution contains the information we need to compute any probability of interest.

Computing with Probabilities: The Chain Rule or Factoring

We can always write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

(by definition of joint probability)

Repeatedly applying this idea, we can write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c,.. z)\, P(c| .. z)..P(z)$$

This factorization holds for any ordering of the variables

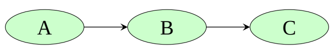
This is the chain rule for probabilities

Definition of Bayesian Networks

- A data structure that represents the dependence between variables
- Gives a concise specification of the *joint probability distribution*
- A Bayesian Network is a directed acyclic graph (DAG) in which the following holds:
  1. A set of random variables makes up the **nodes** in the network
  2. A set of **directed links** connects pairs of nodes
  3. Each node has a **conditional probability table** that quantifies the effects of its *parents* on the node
  4. The graph has not directed cycles (DAG)

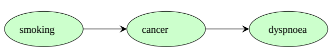
Conditional Independence – Causal Chains

- Causal chains give rise to conditional independence



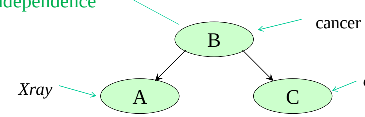
$$P(C \mid A \wedge B) = P(C \mid B)$$

- Example: "*Smoking causes cancer, which causes dyspnoea*"



Conditional Independence – Common Causes

- Common Causes (or ancestors) also give rise to conditional independence



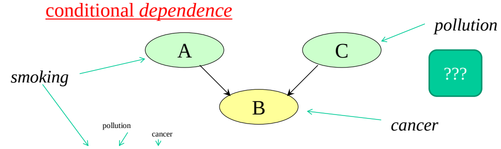
$$P(C \mid A \wedge B) = P(C \mid B) \;\; \boxed{\equiv \text{(A indep C) | B}}$$

Example: "*Cancer is a common cause of the two symptoms: a positive Xray and dyspnoea*"

One has dyspnoea (C) because of cancer (B)
so he does not need an Xray test

Conditional Dependence – Common Effects

- Common effects (or their descendants) give rise to <u>conditional *dependence*</u>



$$P(A \mid C \wedge B) \neq P(A)P(C) \equiv \neg((A \text{ indep } C) \mid B)$$

- Example: "*Cancer is a common effect of pollution and smoking*"
  Given cancer, smoking "explains away" pollution

We know that you smoke and have cancer, we do not need to assume that your cancer was caused by pollution

Belief Network Example

- Burglar alarm at home ‹
  - Fairly reliable at detecting a burglary
  - Responds at times to minor earthquakes
- Two neighbors, on hearing alarm, calls police
  - John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too. ‹
  - Mary likes loud music and sometimes misses the alarm altogether

Bayesian Networks

- A Bayesian network specifies a joint distribution in a structured form
- Represent dependence/independence via a directed graph
  - Nodes = random variables
  - Edges = direct dependence
- Structure of the graph ⇔ Conditional independence relations

In general,

$p(X_1, X_2, ...X_N) = \Pi\, p(X_i \mid parents(X_i))$

The full joint distribution The graph-structured approximation

- Requires that graph is acyclic (no directed cycles)
- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

What Independence does a Bayes Net Model?

- In order for a Bayesian network to model a probability distribution, the following must be true by definition:

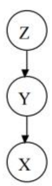
  Each variable is conditionally independent of all its non-descendants in the graph given the value of all its parents.

- This implies

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

What Independence does a Bayes Net Model?

- Example:



Given Y, does learning the value of Z tell us nothing new about X?

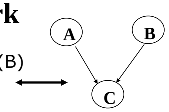
I.e., is $P(X|Y, Z)$ equal to $P(X \mid Y)$?

Yes. Since we know the value of all of X's parents (namely, Y), and Z is not a descendant of X, X is conditionally independent of Z.

Also, since independence is symmetric, $P(Z|Y, X) = P(Z|Y)$.

Example of a simple Bayesian network

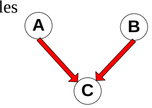
p(A,B,C) = p(C|A,B)p(A)p(B)



- Probability model has simple factored form
- Directed edges => direct dependence
- Absence of an edge => conditional independence
- Also known as belief networks, graphical models, causal networks
- Other formulations, e.g., undirected graphical models

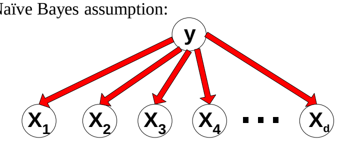
Bayesian Belief Networks

- Provides graphical representation of probabilistic relationships among a set of random variables
- Consists of:
  - A directed acyclic graph (dag)
    - Node corresponds to a variable
    - Arc corresponds to dependence relationship between a pair of variables
  - A probability table associating each node to its immediate parent



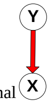
Conditional Independence

- Naïve Bayes assumption:

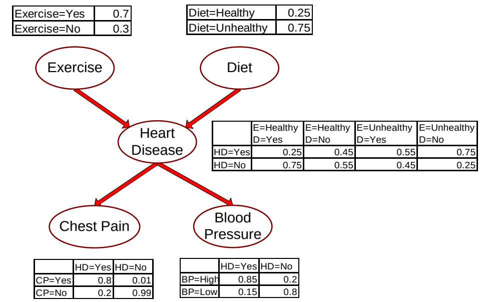


Probability Tables

- If X does not have any parents, table contains prior probability P(X)
- If X has only one parent (Y), table contains conditional probability P(X|Y)
- If X has multiple parents ($Y_1, Y_2, \ldots, Y_k$), table contains conditional probability $P(X|Y_1, Y_2, \ldots, Y_k)$



Example of Bayesian Belief Network



Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75

	E=Healthy D=Yes	E=Healthy D=No	E=Unhealthy D=Yes	E=Unhealthy D=No
HD=Yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Example of Inferencing using BBN

- Given: X = (E=No, D=Yes, CP=Yes, BP=High)
  - Compute P(HD|E,D,CP,BP)?
- P(HD=Yes| E=No,D=Yes) = 0.55
  P(CP=Yes| HD=Yes) = 0.8
  P(BP=High| HD=Yes) = 0.85
  - P(HD=Yes|E=No,D=Yes,CP=Yes,BP=High) $\propto$
- P(HD=No| E=No,D=Yes) = 0.45
  P(CP=Yes| HD=No) = 0.01
  P(BP=High| HD=No) = 0.2
  - P(HD=No|E=No,D=Yes,CP=Yes,BP=High) $\propto$

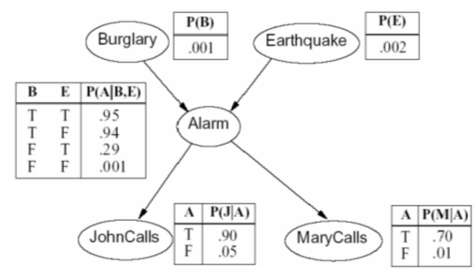
Classify X as Yes

Example of Inferencing using BBN

- Given: X = (E=No, D=Yes, CP=Yes, BP=High)
  - Compute P(HD|E,D,CP,BP)?
- P(HD=Yes| E=No,D=Yes) = 0.55
  P(CP=Yes| HD=Yes) = 0.8
  P(BP=High| HD=Yes) = 0.85
  - P(HD=Yes|E=No,D=Yes,CP=Yes,BP=High) $\propto 0.55 \times 0.8 \times 0.85 = 0.374$
- P(HD=No| E=No,D=Yes) = 0.45
  P(CP=Yes| HD=No) = 0.01
  P(BP=High| HD=No) = 0.2
  - P(HD=No|E=No,D=Yes,CP=Yes,BP=High) $\propto 0.45 \times 0.01 \times 0.2 = 0.0009$

Classify X as Yes

Belief Network Example



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

The joint probability distribution

- Probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both Mary and John call:

P(J ∧ M ∧ A ∧ ~B ∧ ~E)
= ?

The joint probability distribution

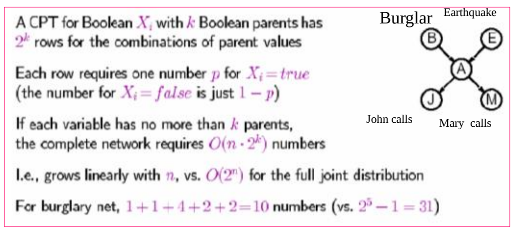
- Probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both Mary and John call:

P(J ∧ M ∧ A ∧ ~B ∧ ~E)
= P(J | A) P(M | A) P(A | ¬B ∧ ¬E) P(¬B) P(¬E)
= 0.9 X 0.7 X 0.001 X 0.999 X 0.998
= 0.00062

Incremental Network Construction

- 1. Choose the set of relevant variables Xi that describe the domain
- 2. Choose an ordering for the variables (very important step)
- 3. While there are variables left:
  a) Pick a variable X and add a node for it
  b) Set Parents(X) to some minimal set of existing nodes such that the conditional independence property is satisfied
  c) Define the conditional prob table for X.

Compactness of Bayes Net


A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

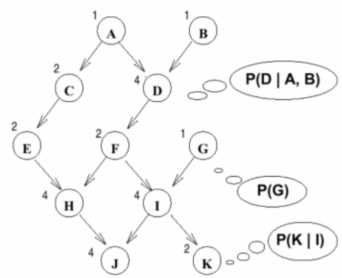
Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)

If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Example (Binary valued Variables)



- A couple CPTS are "shown"
- Explicit joint requires 2^{11} -1 =2047 parmtrs
- BN requires only 27 parmtrs (the number of entries for each CPT is listed)

Causal Intuitions

- The BN can be constructed using an arbitrary ordering of the variables.
- However, some orderings will yield BN's with very large parent sets. This requires exponential space, and exponential time to perform inference.
- Empirically, and conceptually, a good way to construct a BN is to use an ordering based on causality. This often yields a more natural and compact BN.

Causal Intuitions

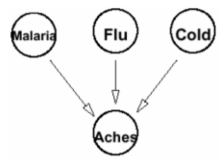
- Malaria, the flu and a cold all “cause” aches. So use the ordering that causes come before effects
  Malaria, Flu, Cold, Aches

Pr(A,C, F, M) = Pr(A|M,F,C) Pr(C|M,F) Pr(F|M) Pr(M)

•Each of these disease affects the probability of aches, so the first conditional probability does not change.

•It is reasonable to assume that these diseases are independent of each other: having or not having one does not change the probability of having the others.
  So Pr(C|M,F) = Pr(C)   Pr(F|M) = Pr(F)

Causal Intuitions



- This yields a fairly simple Bayes net.
- Only need one big CPT, involving the family of “Aches”.

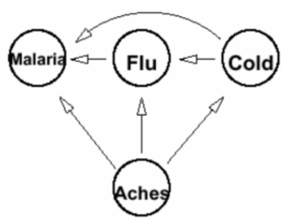
Causal Intuitions

- Suppose we build the BN for distribution P using the opposite ordering
  - •i.e., we use ordering Aches, Cold, Flu, Malaria
    Pr(M, F, C, A) = Pr(M|A,C,F) Pr(F|A,C) Pr(C|A) Pr(A)
  - We can’t reduce Pr(M|A,C,F). Probability of Malaria is clearly affected by knowing aches. What about knowing aches and Cold, or aches and Cold and Flu?
    - •Probability of Malaria is affected by both of these additional pieces of knowledge

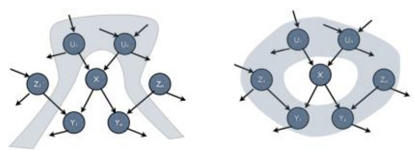
- Similarly, we can’t reduce Pr(F|A,C).
- Pr(C|A) ≠ Pr(C)

Causal Intuitions

- Obtain a much more complex Bayes net. In fact, we obtain no savings over explicitly representing the full joint distribution (i.e., representing the probability of every atomic event).



Conditional Independence Relations in Bayesian Networks

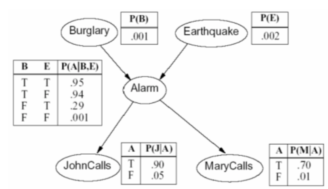


A node, X, is *conditionally independent* of its non-descendants, Z_i, given its parents, U_i.

A node, X, is *conditionally independent* of all other nodes in the network given its *Markov blanket*: its parents, U_i, children, Y_i, and children's parents, Z_i.

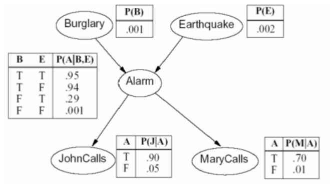
(a)   (b)

Example-(a)



JohnCalls is independent of Burglary and Earthquake given the value of Alarm.

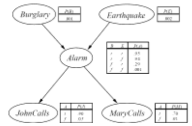
Example-(b)



Burglary is independent of JohnCalls and MaryCalls, given the values of Alarm and Earthquake. .

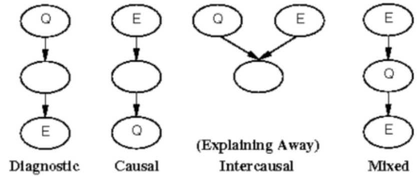
Inference (Reasoning) in Bayesian Networks

- Consider answering a query in a Bayesian Network
  - Q = set of query variables
  - e = evidence (set of instantiated variable-value pairs)
  - Inference = computation of conditional distribution P(Q | e)
- Examples
  - P(burglary | alarm)
  - P(earthquake | JCalls, MCalls)
  - P(JCalls, MCalls | burglary, earthquake)



- Can we use the structure of the Bayesian Network
  to answer such queries efficiently? Answer = yes
  - Generally speaking, complexity is inversely proportional to sparsity of graph

Types of Inference



Samples Inferences

- **Diagnostic (evidential, abductive)**: From effect to cause.
  - P(Burglary | JohnCalls) =
  - P(Burglary | JohnCalls ∧ MaryCalls) =
  - P(Alarm | JohnCalls ∧ MaryCalls) =
  - P(Earthquake | JohnCalls ∧ MaryCalls) =
- **Causal (predictive)**: From cause to effect
  - P(JohnCalls | Burglary) =
  - P(MaryCalls | Burglary) =
- **Intercausal (explaining away)**: Between causes of a common effect.
  - P(Burglary | Alarm) =
  - P(Burglary | Alarm ∧ Earthquake) =
- **Mixed**: Two or more of the above combined
  - (diagnostic and causal) P(Alarm | JohnCalls ∧ ¬Earthquake) =
  - (diagnostic and intercausal) P(Burglary | JohnCalls ∧ ¬Earthquake) =