



Machine Learning and Computational Intelligence Lecture 7

Sanjeeb Prasad Panday, PhD

Associate Professor

Dept. of Electronics and Computer Engineering

Director (ICTC)

IOE, TU

Probability

- Joint probability
 - $P(A \& B), P(A, B), P(A \cap B)$
- Conditional probability
 - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
 - $P(A \cap B) = P(A|B) * P(B)$
- Independent variables
 - $P(A|B) = P(A)$
 - $P(A \cap B) = P(A) * P(B)$
- Bayes Rule
 - $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

Bayesian Networks: A Tutorial

Weng-Keen Wong

School of Electrical Engineering and Computer Science

Oregon State University

Modified by

Longin Jan Latecki

Temple University

latecki@temple.edu

Introduction



Suppose you are trying to determine if a patient has inhalational anthrax. You observe the following symptoms:

- The patient has a cough
- The patient has a fever
- The patient has difficulty breathing

Introduction



You would like to determine how likely the patient is infected with inhalational anthrax given that the patient has a cough, a fever, and difficulty breathing

We are not 100% certain that the patient has anthrax because of these symptoms. We are dealing with uncertainty!

Introduction



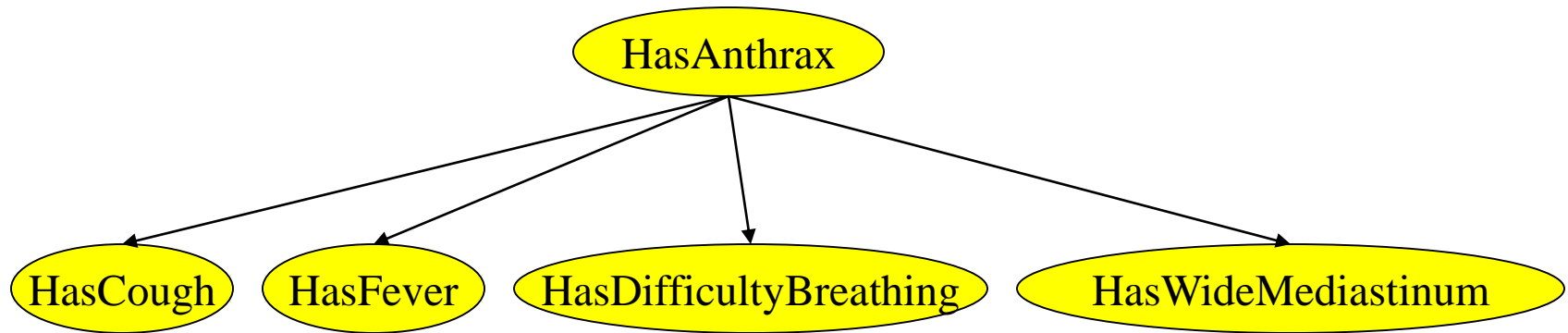
Now suppose you order an x-ray and observe that the patient has a wide mediastinum.

Your belief that that the patient is infected with inhalational anthrax is now much higher.

Introduction

- In the previous slides, what you observed affected your belief that the patient is infected with anthrax
- This is called reasoning with uncertainty
- Wouldn't it be nice if we had some methodology for reasoning with uncertainty? Well in fact, we do...

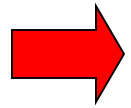
Bayesian Networks



- In the opinion of many AI researchers, Bayesian networks are the most significant contribution in AI in the last 10 years
- They are used in many applications eg. spam filtering, speech recognition, robotics, diagnostic systems and even syndromic surveillance

Outline

1. Introduction



2. Probability Primer

3. Bayesian networks

Probability Primer: Random Variables

- A **random variable** is the basic element of probability
- Refers to an event and there is some degree of uncertainty as to the outcome of the event
- For example, the random variable A could be the event of getting a head on a coin flip



Boolean Random Variables

- We will start with the simplest type of random variables – Boolean ones
- Take the values *true* or *false*
- Think of the event as occurring or not occurring
- Examples (Let A be a Boolean random variable):
 - A = Getting a head on a coin flip
 - A = It will rain today

The Joint Probability Distribution

- Joint probabilities can be between any number of variables
eg. $P(A = \text{true}, B = \text{true}, C = \text{true})$
- For each combination of variables, we need to say how probable that combination is
- The probabilities of these combinations need to sum to 1

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Sums to 1

The Joint Probability Distribution

- Once you have the joint probability distribution, you can calculate any probability involving A , B , and C
- Note: May need to use marginalization and Bayes rule, (both of which are not discussed in these slides)

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Examples of things you can compute:

- $P(A=true) = \text{sum of } P(A,B,C) \text{ in rows with } A=true$
- $P(A=true, B = true / C=true) =$
 $P(A = true, B = true, C = true) / P(C = true)$

The Problem with the Joint Distribution

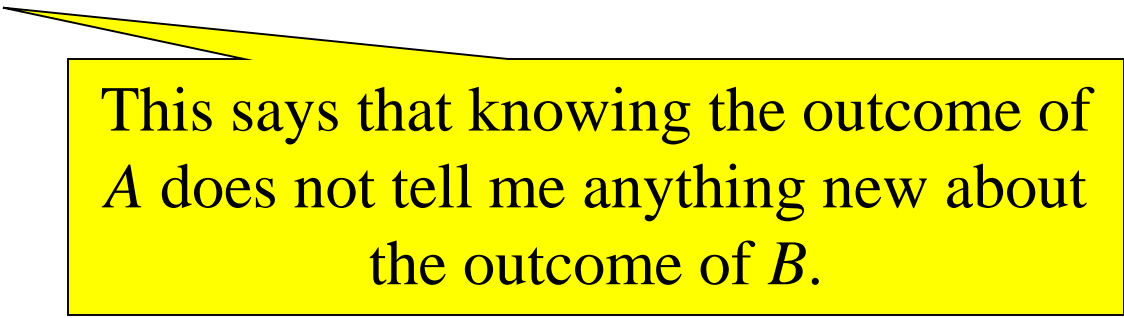
- Lots of entries in the table to fill up!
- For k Boolean random variables, you need a table of size 2^k
- How do we use fewer numbers? Need the concept of independence

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Independence

Variables A and B are independent if any of the following hold:

- $P(A, B) = P(A) P(B)$
- $P(A / B) = P(A)$
- $P(B / A) = P(B)$



This says that knowing the outcome of A does not tell me anything new about the outcome of B .

Independence

How is independence useful?

- Suppose you have n coin flips and you want to calculate the joint distribution $P(C_1, \dots, C_n)$
- If the coin flips are not independent, you need 2^n values in the table
- If the coin flips are independent, then

$$P(C_1, \dots, C_n) = \prod_{i=1}^n P(C_i)$$

Each $P(C_i)$ table has 2 entries and there are n of them for a total of $2n$ values

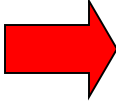
Conditional Independence

Variables A and B are conditionally independent given C if any of the following hold:

- $P(A, B / C) = P(A / C) P(B / C)$
- $P(A / B, C) = P(A / C)$
- $P(B / A, C) = P(B / C)$

Knowing C tells me everything about B . I don't gain anything by knowing A (either because A doesn't influence B or because knowing C provides all the information knowing A would give)

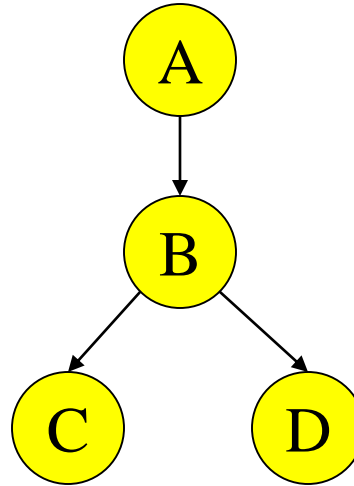
Outline

1. Introduction
2. Probability Primer
-  3. Bayesian networks

A Bayesian Network

A Bayesian network is made up of:

1. A Directed Acyclic Graph



2. A set of tables for each node in the graph

A	P(A)
false	0.6
true	0.4

A	B	P(B A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

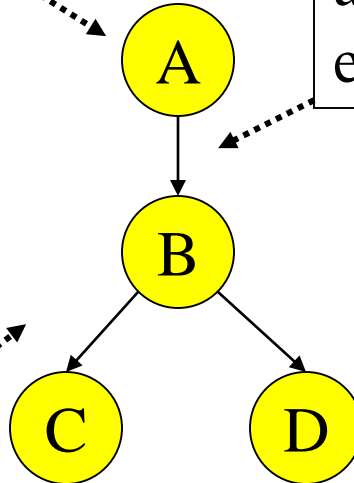
B	D	P(D B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

A Directed Acyclic Graph

Each node in the graph is a random variable

A node X is a parent of another node Y if there is an arrow from node X to node Y
eg. A is a parent of B



Informally, an arrow from node X to node Y means X has a direct influence on Y

A Set of Tables for Each Node

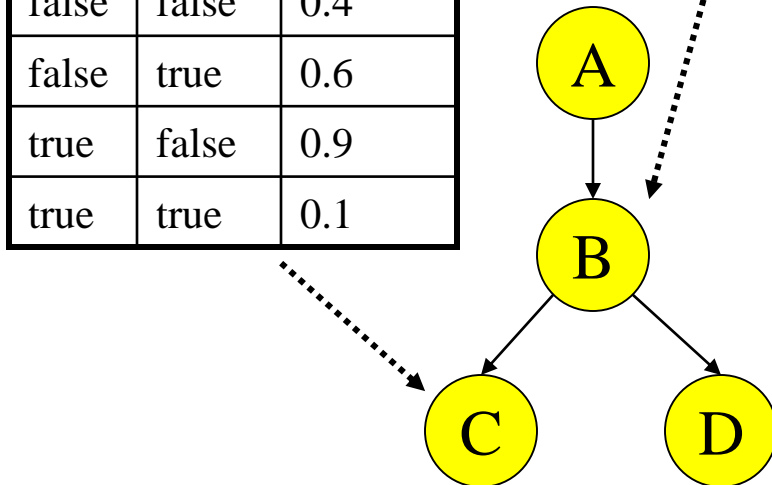
A	P(A)
false	0.6
true	0.4

A	B	P(B A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Each node X_i has a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$ that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability tables (CPTs)



B	D	P(D B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

A Set of Tables for Each Node

Conditional Probability
Distribution for C given B

B	C	P(C B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

For a given combination of values of the parents (B in this example), the entries for $P(C=\text{true} \mid B)$ and $P(C=\text{false} \mid B)$ must add up to 1
eg. $P(C=\text{true} \mid B=\text{false}) + P(C=\text{false} \mid B=\text{false}) = 1$

If you have a Boolean variable with k Boolean parents, this table has 2^{k+1} probabilities (but only 2^k need to be stored)

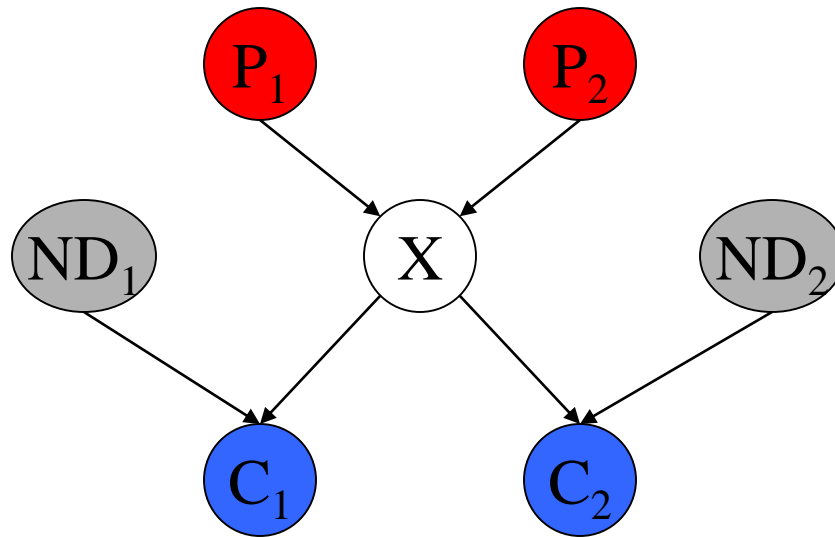
Bayesian Networks

Two important properties:

1. Encodes the conditional independence relationships between the variables in the graph structure
2. Is a compact representation of the joint probability distribution over the variables

Conditional Independence

The Markov condition: given its parents (P_1, P_2), a node (X) is conditionally independent of its non-descendants (ND_1, ND_2)



The Joint Probability Distribution

Due to the Markov condition, we can compute the joint probability distribution over all the variables X_1, \dots, X_n in the Bayesian net using the formula:

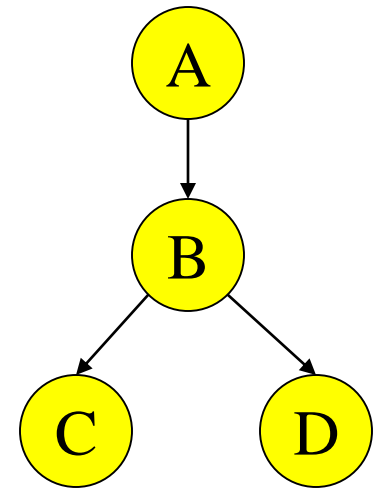
$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i \mid \text{Parents}(X_i))$$

Where $\text{Parents}(X_i)$ means the values of the Parents of the node X_i with respect to the graph

Using a Bayesian Network Example

Using the network in the example, suppose you want to calculate:

$$\begin{aligned} &P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true}) \\ &= P(A = \text{true}) * P(B = \text{true} \mid A = \text{true}) * \\ &\quad P(C = \text{true} \mid B = \text{true}) P(D = \text{true} \mid B = \text{true}) \\ &= (0.4)*(0.3)*(0.1)*(0.95) \end{aligned}$$



Using a Bayesian Network Example

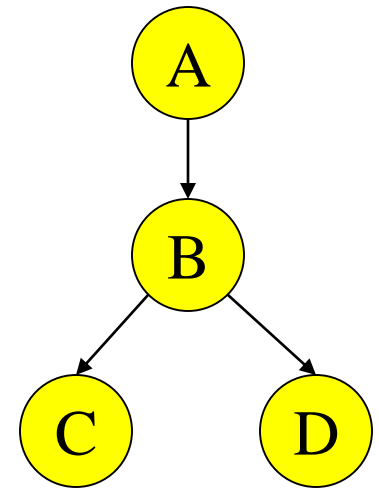
Using the network in the example, suppose you want to calculate:

$$\begin{aligned} &P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true}) \\ &= P(A = \text{true}) * P(B = \text{true} \mid A = \text{true}) * \\ &\quad P(C = \text{true} \mid B = \text{true}) P(D = \text{true} \mid B = \text{true}) \\ &= (0.4) * (0.3) * (0.1) * (0.95) \end{aligned}$$

This is from the
graph structure



These numbers are from the
conditional probability tables



Joint Probability Factorization

For any joint distribution of random variables the following factorization is always true:

$$P(A, B, C, D) = P(A)P(B | A)P(C | A, B)P(D | A, B, C)$$

We derive it by repeatedly applying the Bayes' Rule

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

$$\begin{aligned} P(A, B, C, D) &= P(B, C, D | A)P(A) \\ &= P(C, D | B, A)P(B | A)P(A) \\ &= P(D | C, B, A)P(C | B, A)P(B | A)P(A) \\ &= P(A)P(B | A)P(C | A, B)P(D | A, B, C) \end{aligned}$$

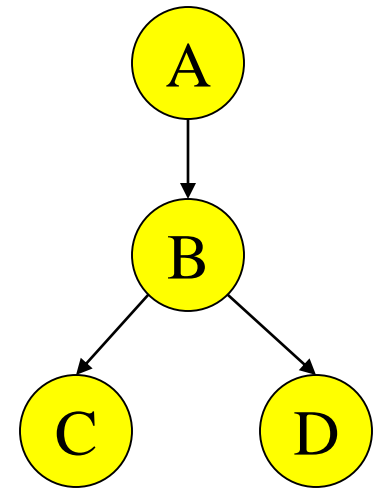
Joint Probability Factorization

Our example graph carries additional independence information, which simplifies the joint distribution:

$$\begin{aligned} P(A, B, C, D) &= P(A)P(B \mid A)P(C \mid A, B)P(D \mid A, B, C) \\ &= P(A)P(B \mid A)P(C \mid B)P(D \mid B) \end{aligned}$$

This is why, we only need the tables for $P(A)$, $P(B|A)$, $P(C|B)$, and $P(D|B)$ and why we computed

$$\begin{aligned} &P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true}) \\ &= P(A = \text{true}) * P(B = \text{true} \mid A = \text{true}) * \\ &\quad P(C = \text{true} \mid B = \text{true}) P(D = \text{true} \mid B = \text{true}) \\ &= (0.4)*(0.3)*(0.1)*(0.95) \end{aligned}$$



Inference

- Using a Bayesian network to compute probabilities is called inference
- In general, inference involves queries of the form:

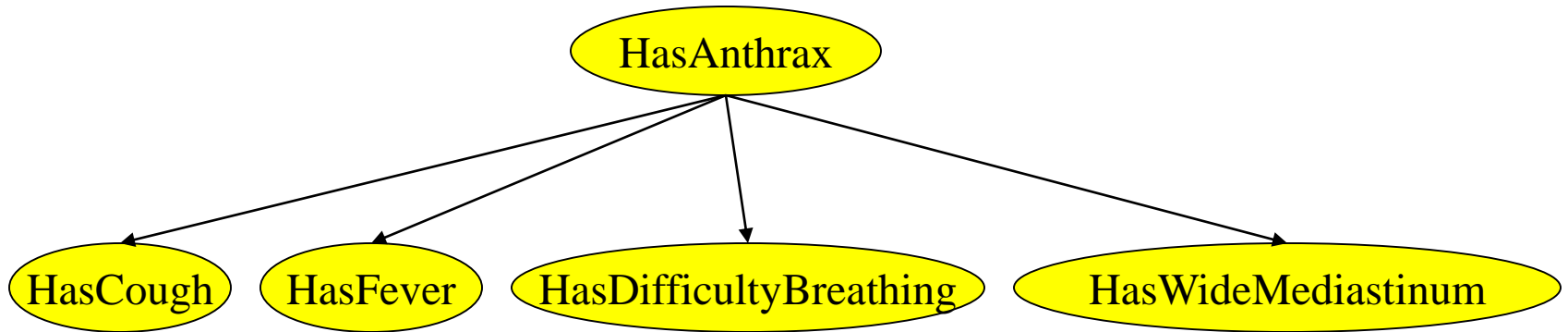
$$P(X \mid E)$$



E = The evidence variable(s)

X = The query variable(s)

Inference



- An example of a query would be:
 $P(HasAnthrax = true \mid HasFever = true, HasCough = true)$
- Note: Even though *HasDifficultyBreathing* and *HasWideMediastinum* are in the Bayesian network, they are not given values in the query (ie. they do not appear either as query variables or evidence variables)
- They are treated as unobserved variables and summed out.

Inference Example

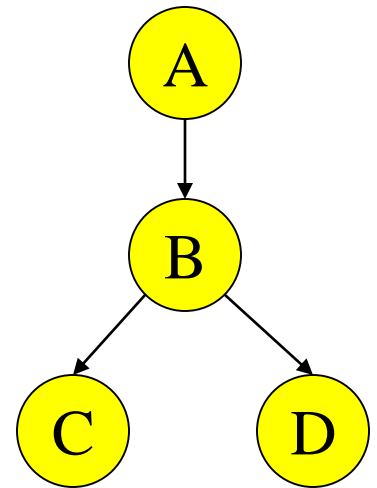
Supposed we know that $A=\text{true}$.

What is more probable $C=\text{true}$ or $D=\text{true}$?

For this we need to compute

$P(C=t \mid A=t)$ and $P(D=t \mid A=t)$.

Let us compute the first one.



$$P(C = t \mid A = t) = \frac{P(A = t, C = t)}{P(A = t)} = \frac{\sum_{b,d} P(A = t, B = b, C = t, D = d)}{P(A = t)}$$

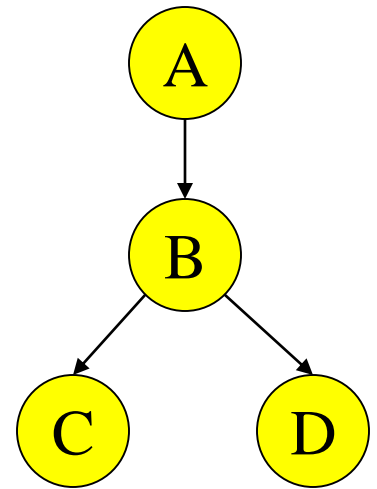
A	P(A)
false	0.6
true	0.4

A	B	P(B A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

What is $P(A=\text{true})$?



$$\begin{aligned}
 P(A=t) &= \sum_{b,c,d} P(A=t, B=b, C=c, D=d) \\
 &= \sum_{b,c,d} P(A=t)P(B=b \mid A=t)P(C=c \mid B=b)P(D=d \mid B=b) \\
 &= P(A=t) \sum_{b,c,d} P(B=b \mid A=t)P(C=c \mid B=b)P(D=d \mid B=b) \\
 &= P(A=t) \sum_b P(B=b \mid A=t) \sum_{c,d} P(C=c \mid B=b)P(D=d \mid B=b) \\
 &= P(A=t) \sum_b P(B=b \mid A=t) \sum_c P(C=c \mid B=b) \sum_d P(D=d \mid B=b) \\
 &= P(A=t) \sum_b P(B=b \mid A=t) \sum_c P(C=c \mid B=b) * 1 \\
 &= 0.4(P(B=t \mid A=t) \sum_c P(C=c \mid B=t) + P(B=f \mid A=t) \sum_c P(C=c \mid B=f)) = \dots
 \end{aligned}$$

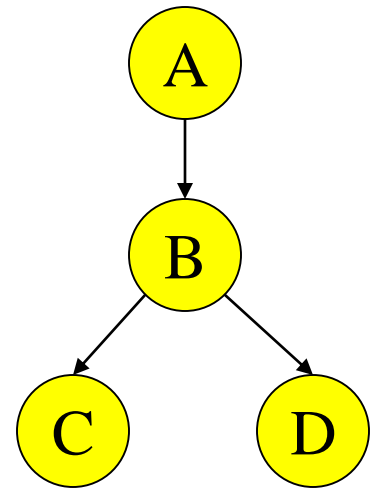
A	P(A)
false	0.6
true	0.4

A	B	P(B A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

What is $P(C=\text{true}, A=\text{true})$?



$$\begin{aligned}
 P(A=t, C=t) &= \sum_{b,d} P(A=t, B=b, C=t, D=d) \\
 &= \sum_{b,d} P(A=t)P(B=b \mid A=t)P(C=t \mid B=b)P(D=d \mid B=b) \\
 &= P(A=t) \sum_b P(B=b \mid A=t)P(C=t \mid B=b) \sum_d P(D=d \mid B=b) \\
 &= 0.4(P(B=t \mid A=t)P(C=t \mid B=t) \sum_d P(D=d \mid B=t) \\
 &\quad + P(B=f \mid A=t)P(C=t \mid B=f) \sum_d P(D=d \mid B=f)) \\
 &= 0.4(0.3 * 0.1 * 1 + 0.7 * 0.6 * 1) = 0.4(0.03 + 0.42) = 0.4 * 0.45 = 0.18
 \end{aligned}$$

A	P(A)
false	0.6
true	0.4

A	B	P(B A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

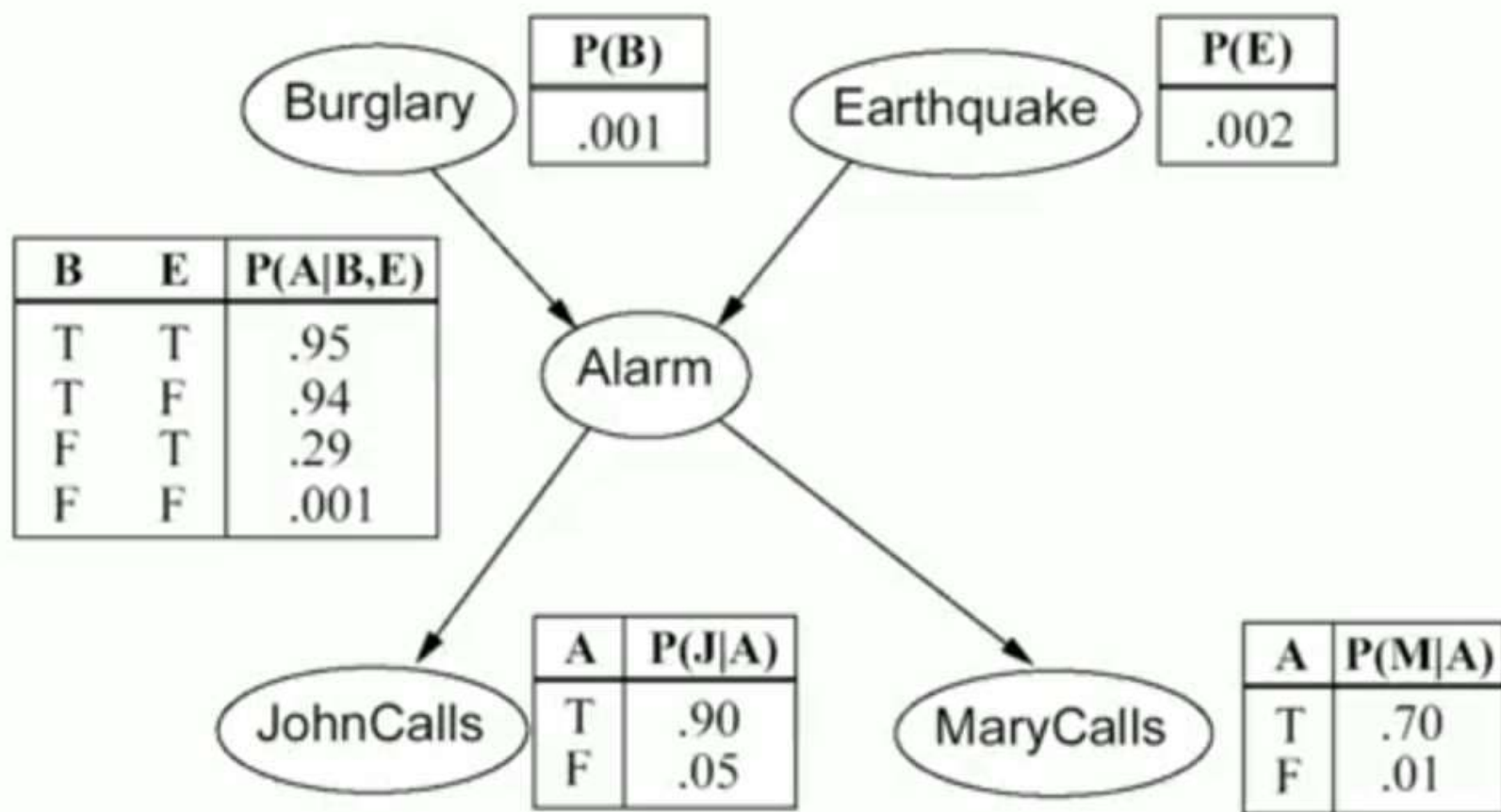
B	D	P(D B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

BAYESIAN BELIEF NETWORKS – EXAMPLE – 1

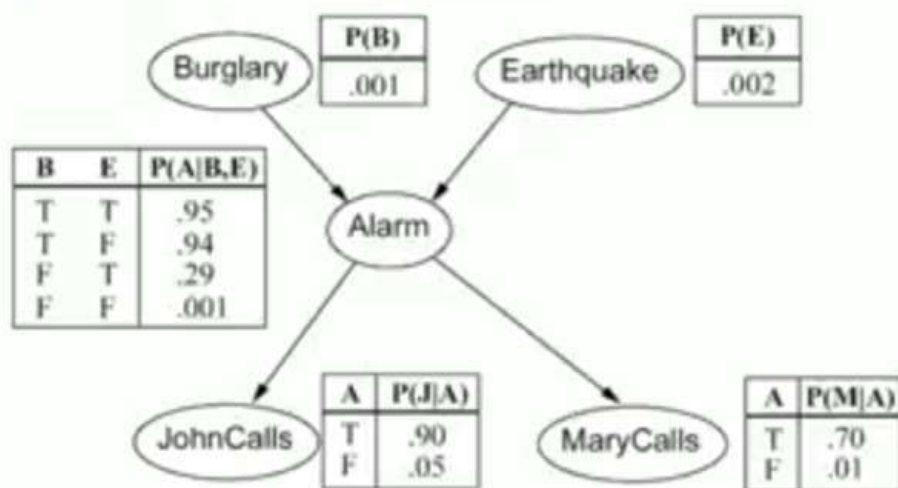
- You have a new burglar alarm installed at home.
- It is fairly reliable at detecting burglary, but also sometimes responds to minor earthquakes.
- You have two neighbors, John and Merry , who promised to call you at work when they hear the alarm.
- John always calls when he hears the alarm, but sometimes confuses telephone ringing with the alarm and calls too.
- Merry likes loud music and sometimes misses the alarm.
- Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

BAYESIAN BELIEF NETWORKS – EXAMPLE – 1



BAYESIAN BELIEF NETWORKS – EXAMPLE – 1

1. What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Merry call?



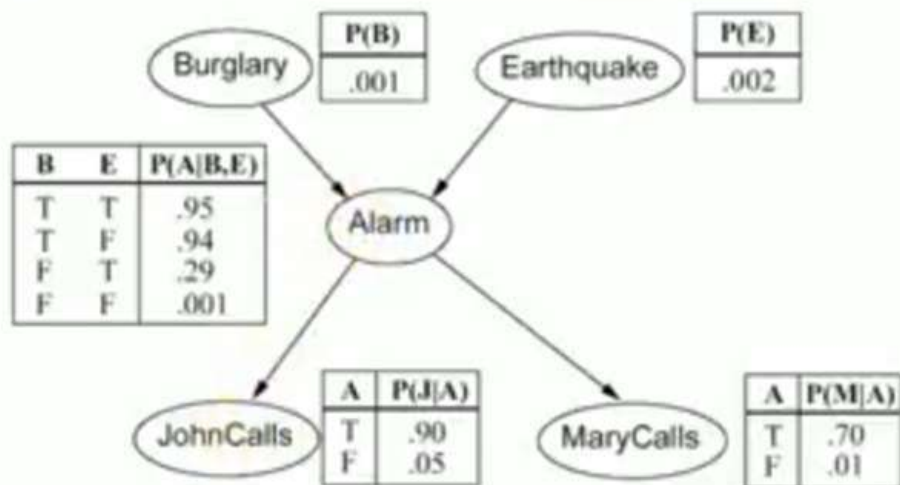
Solution:

$$\begin{aligned} P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) &= P(j \mid a) P(m \mid a) P(a \mid \neg b, \neg e) P(\neg b) P(\neg e) \\ &= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 \\ &= 0.00062 \end{aligned}$$

BAYESIAN BELIEF NETWORKS – EXAMPLE – 1

2. What is the probability that John call?

Solution:



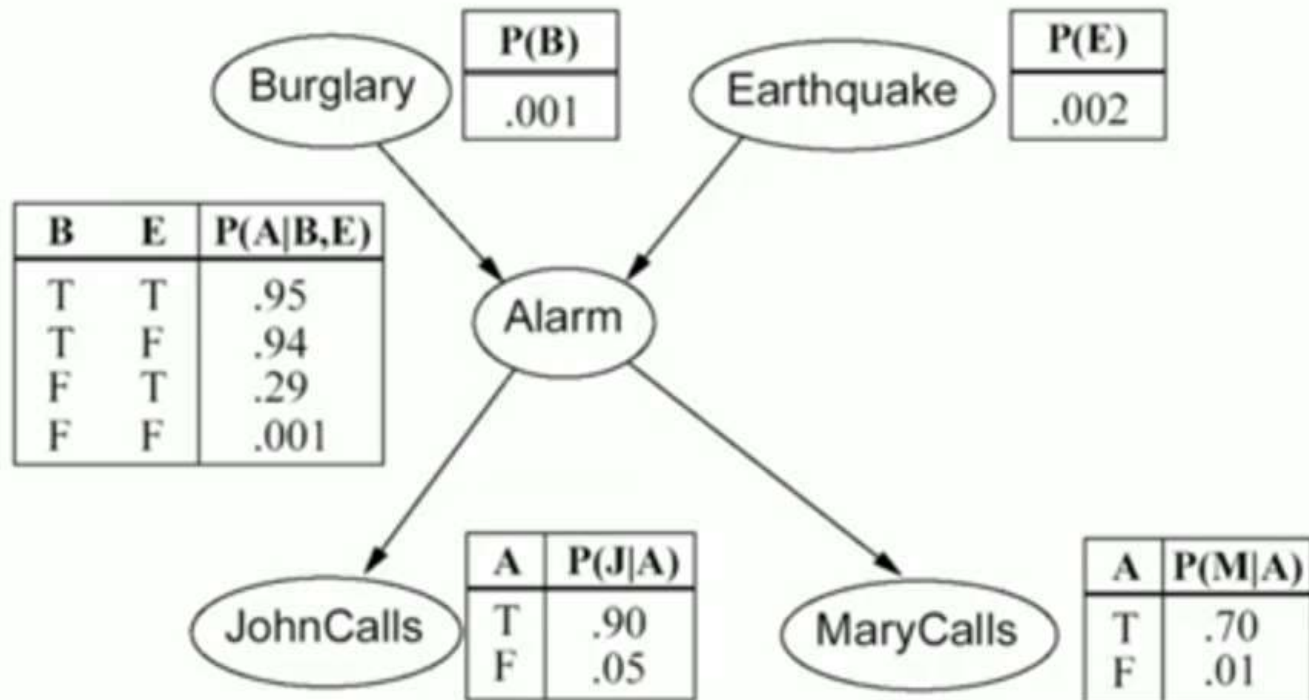
$$P(j) = P(j | a) P(a) + P(j | \neg a) P(\neg a)$$

$$= P(j | a) \{P(a | b, e) * P(b, e) + P(a | \neg b, e) * P(\neg b, e) + P(a | b, \neg e) * P(b, \neg e) + P(a | \neg b, \neg e) * P(\neg b, \neg e)\}$$

$$+ P(j | \neg a) \{P(\neg a | b, e) * P(b, e) + P(\neg a | \neg b, e) * P(\neg b, e) + P(\neg a | b, \neg e) * P(b, \neg e) + P(\neg a | \neg b, \neg e) * P(\neg b, \neg e)\}$$

$$= 0.90 * 0.00252 + 0.05 * 0.9974 = 0.0521$$

BAYESIAN BELIEF NETWORKS – EXAMPLE – 2



3. What is the probability that there is a burglary given that John and Merry calls?

BAYESIAN BELIEF NETWORKS – EXAMPLE – 2

- Suppose, we are given for the evidence variables E_1, \dots, E_m , their values e_1, \dots, e_m , and we want to predict whether the query variable X has the value x or not.
- For this we compute and compare the following:

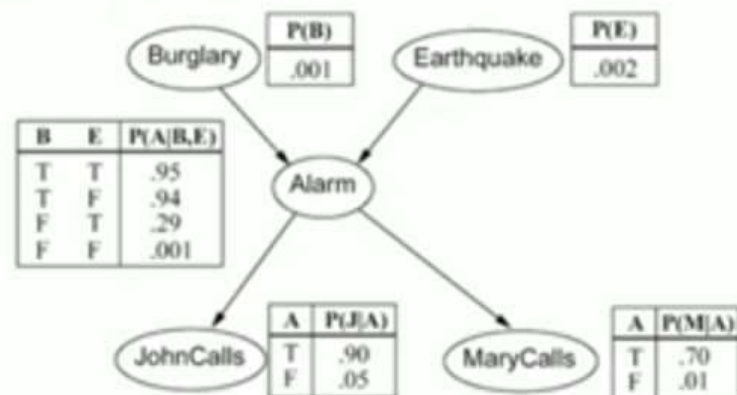
$$P(x | e_1, \dots, e_m) = \frac{P(x, e_1, \dots, e_m)}{P(e_1, \dots, e_m)} = \alpha P(x, e_1, \dots, e_m)$$

$$P(\neg x | e_1, \dots, e_m) = \frac{P(\neg x, e_1, \dots, e_m)}{P(e_1, \dots, e_m)} = \alpha P(\neg x, e_1, \dots, e_m)$$

$$\alpha = \frac{1}{(P(x, e_1, \dots, e_m) + P(\neg x, e_1, \dots, e_m))}$$

BAYESIAN BELIEF NETWORKS – EXAMPLE – 2

3. What is the probability that there is a burglary given that John and Merry calls?

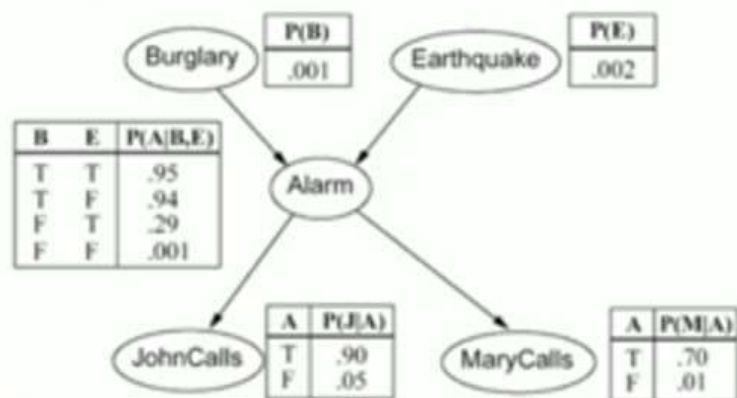


$$\begin{aligned}
 P(b | j, m) &= \alpha P(b) \sum_a P(j|a) P(m|a) \sum_e P(a|b, e) P(e) \\
 &= \alpha P(b) \sum_a P(j|a) P(m|a) \{ P(a|b, e) P(e) + P(a|b, \neg e) P(\neg e) \} \\
 &= \alpha P(b) [P(j|a) P(m|a) \{ P(a|b, e) P(e) + P(a|b, \neg e) P(\neg e) \} \\
 &\quad + P(j|\neg a) P(m|\neg a) \{ P(\neg a|b, e) P(e) + P(\neg a|b, \neg e) P(\neg e) \}] \\
 &= \alpha * .001 * (.9 * .7 * (.95 * .002 + .94 * .998) + .05 * .01 * (.05 * .002 + .71 * .998)) \\
 &= \alpha * .00059
 \end{aligned}$$

BAYESIAN BELIEF NETWORKS – EXAMPLE – 2

3. What is the probability that there is a burglary given that John and Merry calls?

$$\begin{aligned}
 P(\neg b \mid j, m) &= \alpha P(\neg b) \sum_a P(j|a)P(m|a) \sum_e P(a|\neg b, e)P(e) \\
 &= \alpha P(\neg b) \sum_a P(j|a)P(m|a) \{ P(a|\neg b, e)P(e) + P(a|\neg b, \neg e)P(\neg e) \} \\
 &= \alpha P(\neg b) [P(j|a)P(m|a) \{ P(a|\neg b, e)P(e) + P(a|\neg b, \neg e)P(\neg e) \} \\
 &\quad + P(j|\neg a)P(m|\neg a) \{ P(\neg a|\neg b, e)P(e) + P(\neg a|\neg b, \neg e)P(\neg e) \}] \\
 &= \alpha * .999 * (.9 * .7 * (.29 * .002 + .001 * .998) + .05 * .01 * (.71 * .002 + .999 * .998)) \\
 &= \alpha * .0015
 \end{aligned}$$



BAYESIAN BELIEF NETWORKS – EXAMPLE – 2

3. What is the probability that there is a burglary given that John and Merry calls?

$$\alpha = \frac{1}{(P(b, j, m) + P(\neg b, j, m))}$$

$$\begin{aligned}\alpha &= \frac{1}{(.00059 + .0015)} \\ &= 478.5\end{aligned}$$

$$\begin{aligned}P(b \mid j, m) &= \alpha * P(b, j, m) \\ &= 478.5 * .00059 \\ &= 0.28\end{aligned}$$

$$\begin{aligned}P(\neg b \mid j, m) &= \alpha * P(\neg b, j, m) \\ &= 478.5 * .0015 \\ &= 0.72\end{aligned}$$

Naïve Bayes Classifier

QUIZZ: Probability Basics

- **Quiz:** We have two six-sided dice. When they are rolled, it could end up with the following occurrence: (*A*) dice 1 lands on side "3", (*B*) dice 2 lands on side "1", and (*C*) Two dice sum to eight. Answer the following questions:

1) $P(A) = ?$

2) $P(B) = ?$

3) $P(C) = ?$

4) $P(A | B) = ?$

5) $P(C | A) = ?$

6) $P(A, B) = ?$

7) $P(A, C) = ?$

8) Is $P(A, C)$ equals $P(A) * P(C)$?



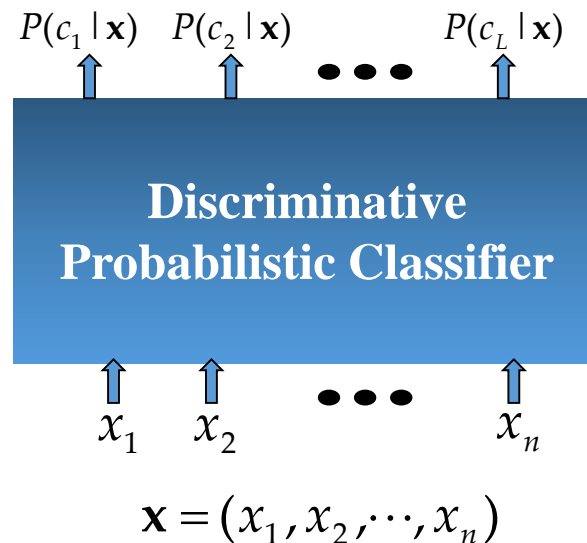
Probabilistic Classification

Probabilistic Classification

- Establishing a probabilistic model for classification
 - Discriminative model**

$$P(C | \mathbf{X}) \quad C = c_1, \dots, c_L, \mathbf{X} = (X_1, \dots, X_n)$$

What is a
discriminative
Probabilistic
Classifier?

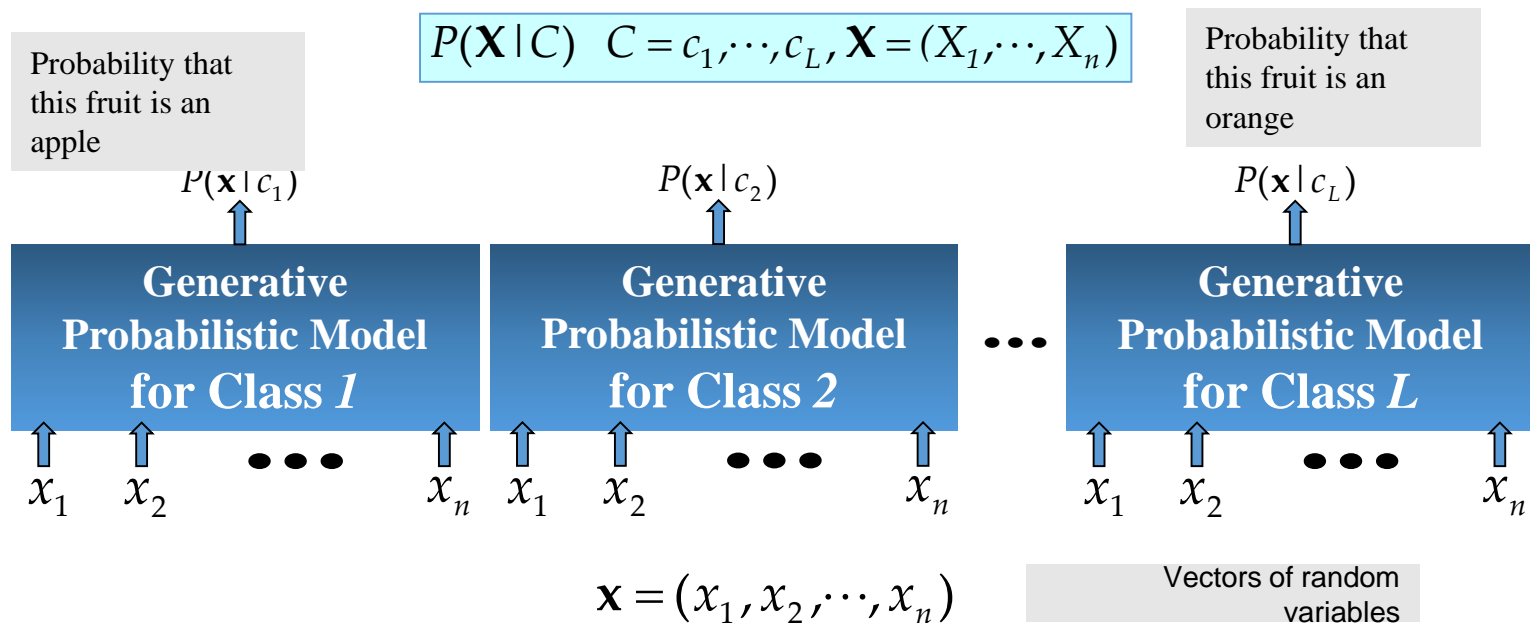


Example

- C_1 – benign mole
- C_2 – cancer

Probabilistic Classification

- Establishing a probabilistic model for classification (cont.)
 - Generative model**



Background: Methods to create classifiers

- There are three methods to establish a classifier

a) Model a classification rule directly

Examples: k-NN, decision trees, perceptron, SVM

b) Model the probability of class memberships given input data

Example: Perceptron with the cross-entropy cost

c) Make a probabilistic model of data within each class

Examples: ***Naive Bayes***, model based classifiers

- **a)** and **b)** are examples of **discriminative** classification
- **c)** is an example of **generative** classification
- **b)** and **c)** are both examples of **probabilistic** classification

GOOD NEWS: You can create your own hardware/software classifiers!

Probability Basics

- We defined prior, conditional and joint probability for random variables
 - Prior probability: $P(X)$
 - Conditional probability: $P(X_1 | X_2), P(X_2 | X_1)$
 - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
 - Relationship: $P(X_1, X_2) = P(X_2 | X_1)P(X_1) = P(X_1 | X_2)P(X_2)$
 - Independence: $P(X_2 | X_1) = P(X_2), P(X_1 | X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$
- **Bayesian Rule**

$$P(C | \mathbf{X}) = \frac{P(\mathbf{X} | C)P(C)}{P(\mathbf{X})}$$



$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Method: Probabilistic Classification with MAP

- MAP classification rule
 - **MAP: Maximum A Posterior**
 - Assign x to c^* if

We use this rule in many applications

$$P(C = c^* | \mathbf{X} = \mathbf{x}) > P(C = c | \mathbf{X} = \mathbf{x}) \quad c \neq c^*, c = c_1, \dots, c_L$$

- **Method of** Generative classification with the MAP rule
 1. Apply Bayesian rule to convert them into posterior probabilities

$$\begin{aligned} P(C = c_i | \mathbf{X} = \mathbf{x}) &= \frac{P(\mathbf{X} = \mathbf{x} | C = c_i)P(C = c_i)}{P(\mathbf{X} = \mathbf{x})} \\ &\propto P(\mathbf{X} = \mathbf{x} | C = c_i)P(C = c_i) \\ &\quad \text{for } i = 1, 2, \dots, L \end{aligned}$$

2. Then apply the MAP rule

Naïve Bayes

Naïve Bayes

For a class, the previous generative model can be decomposed by **n** generative models of a single input.

- Bayes classification

$$P(C | \mathbf{X}) \propto P(\mathbf{X} | C)P(C) = P(X_1, \dots, X_n | C)P(C)$$

Difficulty: learning the joint probability $P(X_1, \dots, X_n | C)$

- Naïve Bayes classification

- Assumption that **all input attributes are conditionally independent!**

$$\begin{aligned} P(X_1, X_2, \dots, X_n | C) &= P(X_1 | X_2, \dots, X_n, C)P(X_2, \dots, X_n | C) \\ &= P(X_1 | C)P(X_2, \dots, X_n | C) \\ &= P(X_1 | C)P(X_2 | C) \cdots P(X_n | C) \end{aligned}$$

Product of individual probabilities

- MAP classification rule: for $\mathbf{x} = (x_1, x_2, \dots, x_n)$

$$[P(x_1 | c^*) \cdots P(x_n | c^*)]P(c^*) > [P(x_1 | c) \cdots P(x_n | c)]P(c), \quad c \neq c^*, c = c_1, \dots, c_L$$

Naïve Bayes Algorithm

- Naïve Bayes Algorithm (for discrete input attributes) has two phases

- **1. Learning Phase:** Given a training set S ,

Learning is easy, just create probability tables.

For each target value of c_i ($c_i = c_1, \dots, c_L$)

$\hat{P}(C = c_i) \leftarrow$ estimate $P(C = c_i)$ with examples in S ;

For every attribute value x_{jk} of each attribute X_j ($j = 1, \dots, n; k = 1, \dots, N_j$)

$\hat{P}(X_j = x_{jk} | C = c_i) \leftarrow$ estimate $P(X_j = x_{jk} | C = c_i)$ with examples in S ;

Output: conditional probability tables; for $X_j, N_j \times L$ elements

- **2. Test Phase:** Given an unknown instance $\mathbf{X}' = (a'_1, \dots, a'_n)$,

Look up tables to assign the label c^* to \mathbf{X}' if

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_n | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_n | c)] \hat{P}(c), \quad c \neq c^*, c = c_1, \dots, c_L$$

Classification is easy, just multiply probabilities

Tennis Example

- Example: Play Tennis

PlayTennis: training examples

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

PlayTennis: training examples

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

The learning phase for tennis example

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Play}=\text{No}) = 5/14$$

We have four variables, we calculate for each we calculate the conditional probability table

Outlook	Play=Yes	Play=No
<i>Sunny</i>	2/9	3/5
<i>Overcast</i>	4/9	0/5
<i>Rain</i>	3/9	2/5

Temperature	Play=Yes	Play=No
<i>Hot</i>	2/9	2/5
<i>Mild</i>	4/9	2/5
<i>Cool</i>	3/9	1/5

Humidity	Play=Yes	Play=No
<i>High</i>	3/9	4/5
<i>Normal</i>	6/9	1/5

Wind	Play=Yes	Play=No
<i>Strong</i>	3/9	3/5
<i>Weak</i>	6/9	2/5

Formulation of a Classification Problem

- **Given** the data as found in last slide:
- **Find** for a new point in space (vector of values) to which group it belongs (classify)

The *test phase* for the tennis example

- Test Phase

- Given a new instance of variable values,

$x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

- Given calculated Look up tables

$$P(\text{Outlook}=\text{Sunny} \mid \text{Play}=\text{Yes}) = 2/9$$

$$P(\text{Outlook}=\text{Sunny} \mid \text{Play}=\text{No}) = 3/5$$

$$P(\text{Temperature}=\text{Cool} \mid \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Temperature}=\text{Cool} \mid \text{Play}=\text{No}) = 1/5$$

$$P(\text{Humidity}=\text{High} \mid \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Humidity}=\text{High} \mid \text{Play}=\text{No}) = 4/5$$

$$P(\text{Wind}=\text{Strong} \mid \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Wind}=\text{Strong} \mid \text{Play}=\text{No}) = 3/5$$

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Play}=\text{No}) = 5/14$$

- Use the MAP rule to calculate Yes or No

$$P(\text{Yes} \mid x'): [P(\text{Sunny} \mid \text{Yes})P(\text{Cool} \mid \text{Yes})P(\text{High} \mid \text{Yes})P(\text{Strong} \mid \text{Yes})]P(\text{Play}=\text{Yes}) = 0.0053$$

$$P(\text{No} \mid x'): [P(\text{Sunny} \mid \text{No})P(\text{Cool} \mid \text{No})P(\text{High} \mid \text{No})P(\text{Strong} \mid \text{No})]P(\text{Play}=\text{No}) = 0.0206$$

Given the fact $P(\text{Yes} \mid x') < P(\text{No} \mid x')$, we label x' to be “No”.

Issues Relevant to Naïve Bayes

Issues Relevant to Naïve Bayes

1. Violation of Independence Assumption
2. Zero conditional probability Problem

Issues Relevant to Naïve Bayes

First Issue

1. Violation of Independence Assumption

Events are correlated

- For many real world tasks, $P(X_1, \dots, X_n | C) \neq P(X_1 | C) \dots P(X_n | C)$
- Nevertheless, naïve Bayes works surprisingly well anyway!

Issues Relevant to Naïve Bayes

Second Issue

1. Zero conditional probability Problem

- Such problem exists when no example contains the attribute value $P(X_1, \dots, X_n | C) \neq P(X_1 | C) \dots P(X_n | C)$

$$\hat{P}(x_1 | c_i) \dots \hat{P}(a_{jk} | c_i) \dots \hat{P}(x_n | c_i) = 0$$

- In this circumstance, $X_j = a_{jk}$, $\hat{P}(X_j = a_{jk} | C = c_i) = 0$ during test
- For a remedy, conditional probabilities are estimated with

$$\hat{P}(X_j = a_{jk} | C = c_i) = \frac{n_c + mp}{n + m}$$

n_c : number of training examples for which $X_j = a_{jk}$ and $C = c_i$

n : number of training examples for which $C = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of X_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)

1 The Classifier

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values a_1, a_2, \dots, a_n . This results in:

$$V_{nb} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod P(a_i | v_j) \quad (1)$$

We generally estimate $P(a_i | v_j)$ using m-estimates:

$$P(a_i | v_j) = \frac{n_c + mp}{n + m} \quad (2)$$

where:

- $n =$ the number of training examples for which $v = v_j$
- $n_c =$ number of examples for which $v = v_j$ and $a = a_i$
- $p =$ a priori estimate for $P(a_i | v_j)$
- $m =$ the equivalent sample size

2 Car theft Example

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

2.1 data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

2.2 Training example

We want to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set. Looking back at equation (2) we can see how to compute this. We need to calculate the probabilities

$$P(\text{Red}|\text{Yes}), P(\text{SUV}|\text{Yes}), P(\text{Domestic}|\text{Yes}) ,$$

$$P(\text{Red}|\text{No}) , P(\text{SUV}|\text{No}), \text{ and } P(\text{Domestic}|\text{No})$$

and multiply them by $P(\text{Yes})$ and $P(\text{No})$ respectively . We can estimate these values using equation (3).

Yes:	No:
Red:	Red:
n = 5	n = 5
n_c= 3	n_c = 2
p = .5	p = .5
m = 3	m = 3
SUV:	SUV:
n = 5	n = 5
n_c = 1	n_c = 3
p = .5	p = .5
m = 3	m = 3
Domestic:	Domestic:
n = 5	n = 5
n_c = 2	n_c = 3
p = .5	p = .5
m = 3	m =3

Looking at $P(Red|Yes)$, we have 5 cases where $v_j = Yes$, and in 3 of those cases $a_i = Red$. So for $P(Red|Yes)$, $n = 5$ and $n_c = 3$. Note that all attribute are binary (two possible values). We are assuming no other information so, $p = 1 / (\text{number-of-attribute-values}) = 0.5$ for all of our attributes. Our m value is arbitrary, (We will use $m = 3$) but consistent for all attributes. Now we simply apply equation (3) using the precomputed values of n , n_c , p , and m .

$$P(Red|Yes) = \frac{3 + 3 * .5}{5 + 3} = .56$$

$$P(SUV|Yes) = \frac{1 + 3 * .5}{5 + 3} = .31$$

$$P(Domestic|Yes) = \frac{2 + 3 * .5}{5 + 3} = .43$$

$$P(Red|No) = \frac{2 + 3 * .5}{5 + 3} = .43$$

$$P(SUV|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

$$P(Domestic|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

We have $P(Yes) = .5$ and $P(No) = .5$, so we can apply equation (2). For $v = Yes$, we have

$$\begin{aligned} &P(Yes) * P(Red \mid Yes) * P(SUV \mid Yes) * P(Domestic \mid Yes) \\ &= .5 * .56 * .31 * .43 = .037 \end{aligned}$$

and for $v = No$, we have

$$\begin{aligned} &P(No) * P(Red \mid No) * P(SUV \mid No) * P(Domestic \mid No) \\ &= .5 * .43 * .56 * .56 = .069 \end{aligned}$$

Since $0.069 > 0.037$, our example gets classified as 'NO'

Naïve Bayesian Classifier: Training Dataset



Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

New Data:

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayesian Classifier: An Example



Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X|C_i)P(C_i)$, for $i=1,2$

First step: Compute $P(C)$ The prior probability of each class can be computed based on the training tuples:

$$P(\text{buys_computer=yes})=9/14=0.643$$

$$P(\text{buys_computer=no})=5/14=0.357$$

Naïve Bayesian Classifier: An Example



Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X|C_i)P(C_i)$, for $i=1,2$

Second step: compute $P(X|C_i)$

$$\begin{aligned} P(X|\text{buys_computer=yes}) &= P(\text{age=youth} | \text{buys_computer=yes}) \times \\ &\quad P(\text{income=medium} | \text{buys_computer=yes}) \times \\ &\quad P(\text{student=yes} | \text{buys_computer=yes}) \times \\ &\quad P(\text{credit_rating=fair} | \text{buys_computer=yes}) \\ &= 0.044 \end{aligned}$$

$$P(\text{age=youth} | \text{buys_computer=yes}) = 0.222$$

$$P(\text{income=medium} | \text{buys_computer=yes}) = 0.444$$

$$P(\text{student=yes} | \text{buys_computer=yes}) = 6/9 = 0.667$$

$$P(\text{credit_rating=fair} | \text{buys_computer=yes}) = 6/9 = 0.667$$

Naïve Bayesian Classifier: An Example



Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X|C_i)P(C_i)$, for $i=1,2$

Second step: compute $P(X|C_i)$

$$\begin{aligned} P(X|\text{buys_computer=no}) &= P(\text{age=youth}|\text{buys_computer=no}) \times \\ &\quad P(\text{income=medium}|\text{buys_computer=no}) \times \\ &\quad P(\text{student=yes}|\text{buys_computer=no}) \times \\ &\quad P(\text{credit_rating=fair}|\text{buys_computer=no}) \\ &= 0.019 \end{aligned}$$

$$P(\text{age=youth}|\text{buys_computer=no}) = 3/5 = 0.666$$

$$P(\text{income=medium}|\text{buys_computer=no}) = 2/5 = 0.400$$

$$P(\text{student=yes}|\text{buys_computer=no}) = 1/5 = 0.200$$

$$P(\text{credit_rating=fair}|\text{buys_computer=no}) = 2/5 = 0.400$$

Naïve Bayesian Classifier: An Example



Given X (age=youth, income=medium, student=yes, credit=fair)

Maximize $P(X|C_i)P(C_i)$, for $i=1,2$

We have computed in the first and second steps:

$$P(\text{buys_computer}=\text{yes})=9/14=0.643$$

$$P(\text{buys_computer}=\text{no})=5/14=0.357$$

$$P(X|\text{buys_computer}=\text{yes})= 0.044$$

$$P(X|\text{buys_computer}=\text{no})= 0.019$$

Third step: compute $P(X|C_i)P(C_i)$ for each class

$$P(X|\text{buys_computer}=\text{yes})P(\text{buys_computer}=\text{yes})=0.044 \times 0.643=0.028$$

$$P(X|\text{buys_computer}=\text{no})P(\text{buys_computer}=\text{no})=0.019 \times 0.357=0.007$$

The naïve Bayesian Classifier predicts **X belongs to class (“buys_computer = yes”)**

Avoiding the 0-Probability Problem



- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original : } P(A_i | C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace : } P(A_i | C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate : } P(A_i | C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Naïve Bayes (Summary)



- Advantage
 - Robust to isolated noise points
 - Handle missing values by ignoring the instance during probability estimate calculations
 - Robust to irrelevant attributes
- Disadvantage
 - Assumption: class conditional independence, which may cause loss of accuracy
 - Independence assumption may not hold for some attribute. Practically, dependencies exist among variables
 - Use other techniques such as Bayesian Belief Networks (BBN)

Remember



- Bayes' rule can be turned into a classifier
- Maximum A Posteriori (MAP) hypothesis estimation incorporates prior knowledge; Max Likelihood (ML) doesn't
- Naive Bayes Classifier is a simple but effective Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attributes are independent given the class.
- Bayesian classification is a generative approach to classification



**Thank you for your
attention**