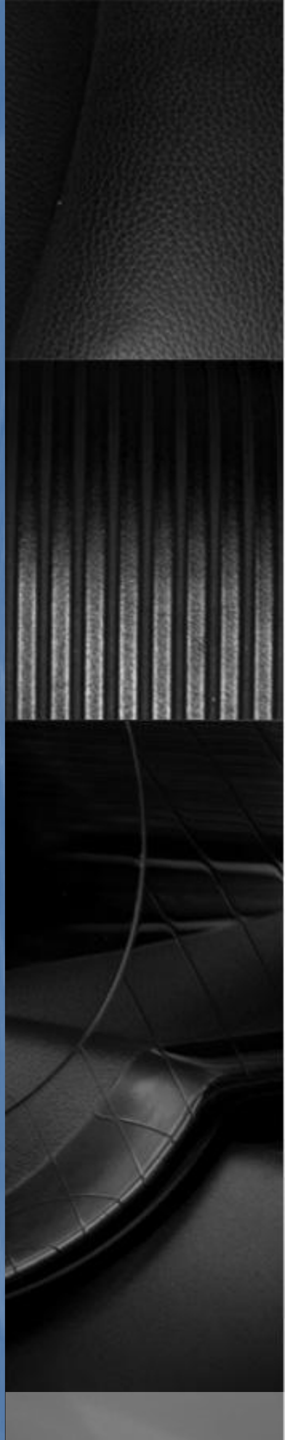
COMP4431 Artificial Intelligence

Bayes Theorem and Bayesian Network

Raymond Pang
Department of Computing
The Hong Kong Polytechnic University

based on Prof. Anita Wasilewska



Overview

- Bayesian Theorem and Conditional Probability
- Naive Bayes Classifier
- Bayesian Belief Network

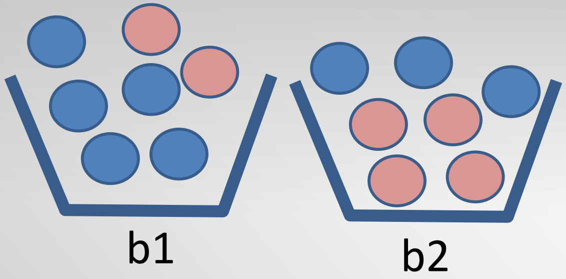
Joint Probability

- The joint probability is the probability of two (or more) simultaneous events, often described in terms of events A and B from two dependent random variables, e.g. X and Y.
- The joint probability is often summarized as just the outcomes, e.g. A and B.
  - – Joint Probability: Probability of two (or more) simultaneous events, e.g. P(A and B) or P(A, B).

Conditional Probability

- The conditional probability is the probability of one event given the occurrence of another event, often described in terms of events A and B from two dependent random variables e.g. X and Y.
- Conditional Probability: Probability of one (or more) event given the occurrence of another event, e.g. P(A given B) or P(A | B).

CONDITIONAL PROBABILITY



Example:

- There are 2 baskets. Basket 1 has 2 red balls and 5 blue balls. Basket 2 has 4 red ball and 3 blue ball.
- Find probability of picking a red ball from basket 1?
- The question above wants P(red ball | basket 1).
- The answer intuitively wants the probability of red ball from only the sample space of basket 1.
- So the answer is 2/7

Then, how about solving P(basket2 | red ball) ???

BAYESIAN THEOREM

$$P(A|B)=\frac{P(A)P(B|A)}{P(B)}=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\overline{A})P(B|\overline{A})}$$

- **A special case of Bayesian Theorem:**

P(A∩B) = P(B) x P(A|B)

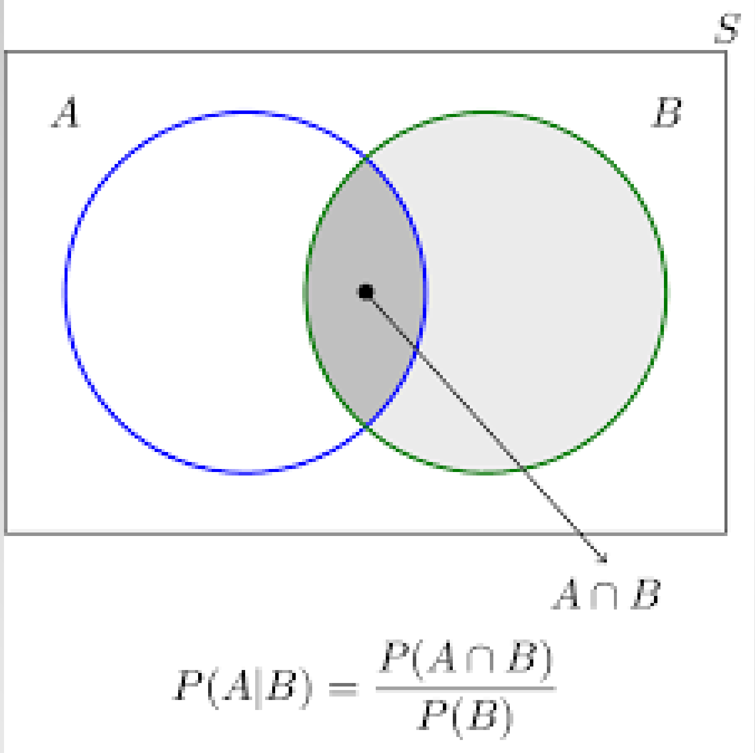
P(B∩A) = P(A) x P(B|A)

Since P(A∩B) = P(B∩A),

P(B) x P(A|B) = P(A) x P(B|A)

⇒P(A|B)

= [P(A) x P(B|A)] / P(B)



BAYESIAN THEOREM

- Firstly, in general, the result P(A|B) is referred to as the posterior probability and P(A) is referred to as the prior probability.
  - P(A|B): Posterior probability.
  - P(A): Prior probability.
- Sometimes P(B|A) is referred to as the likelihood and P(B) is referred to as the evidence.
  - P(B|A): Likelihood.
  - P(B): Evidence.
- This allows Bayes Theorem to be restated as:
  - Posterior = Likelihood * Prior / Evidence

BAYESIAN THEOREM

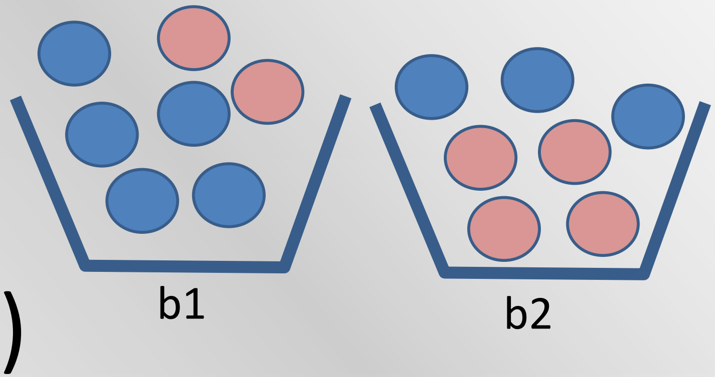
Solution to P(basket2 | red ball) ?

P(basket 2| red ball)

= [P(b2) x P(r | b2)] / P(r)

= (1/2) x (4/7)] / (6/14)

= 0.66



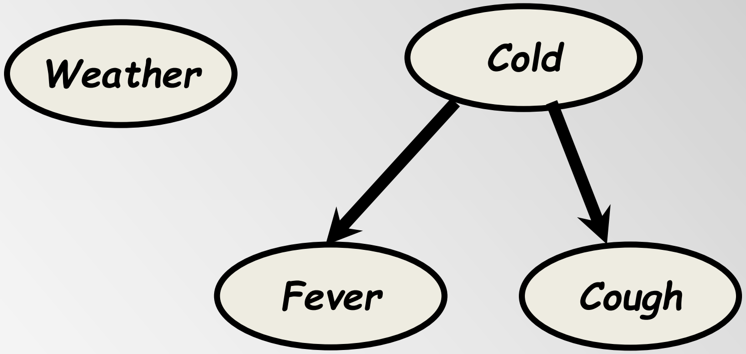
Bayesian Belief Networks

- Bayesian Belief Networks (BBNs) / Bayesian Networks can reason with networks of propositions and associated probabilities
- Useful for many AI problems
    - Diagnosis
    - Expert systems
    - Planning
    - Learning

NAÏVE BAYES CLASSIFIER

Naive Bayes Classifier

- A Naive Bayes classifier is based on a specific type of Bayesian network that makes strong independence assumptions.
  - ❑ In this network, the class variable is the parent node, and all feature variables are child nodes that are conditionally independent given the class.



Weather is independent of other variables

Fever and Cough are independent given Cold

Naive Bayes Classifier

- The key assumption in a Naive Bayes classifier is that all features X are conditionally independent of each other given the class label C. This can be expressed as:

$$P(X_1, X_2, \ldots, X_n \mid C) = \prod_{i=1}^{n} P(X_i \mid C)$$

- This assumption simplifies the computation of the posterior probability $P(C|X_1,X_2,\ldots,X_n)$

Naive Bayes Classifier

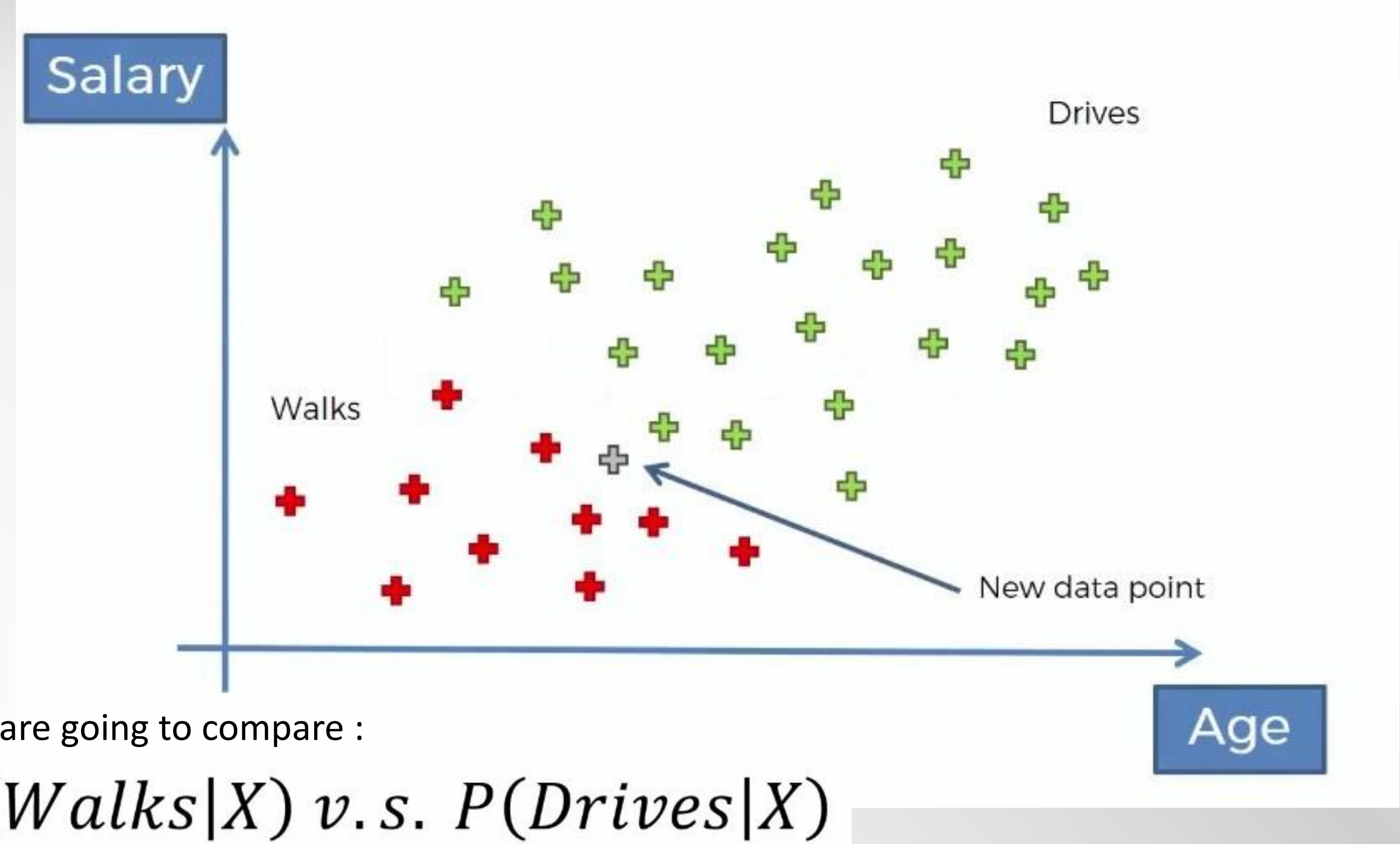
- The goal of the Naive Bayes classifier is to predict the class label *C* for a given set of features X.
- Using Bayes' theorem, the posterior probability is computed as:

$$P(C \mid X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n \mid C) \cdot P(C)}{P(X_1, X_2, \ldots, X_n)}$$

- Now, we want to find the class *C* that maximizes the posterior probability

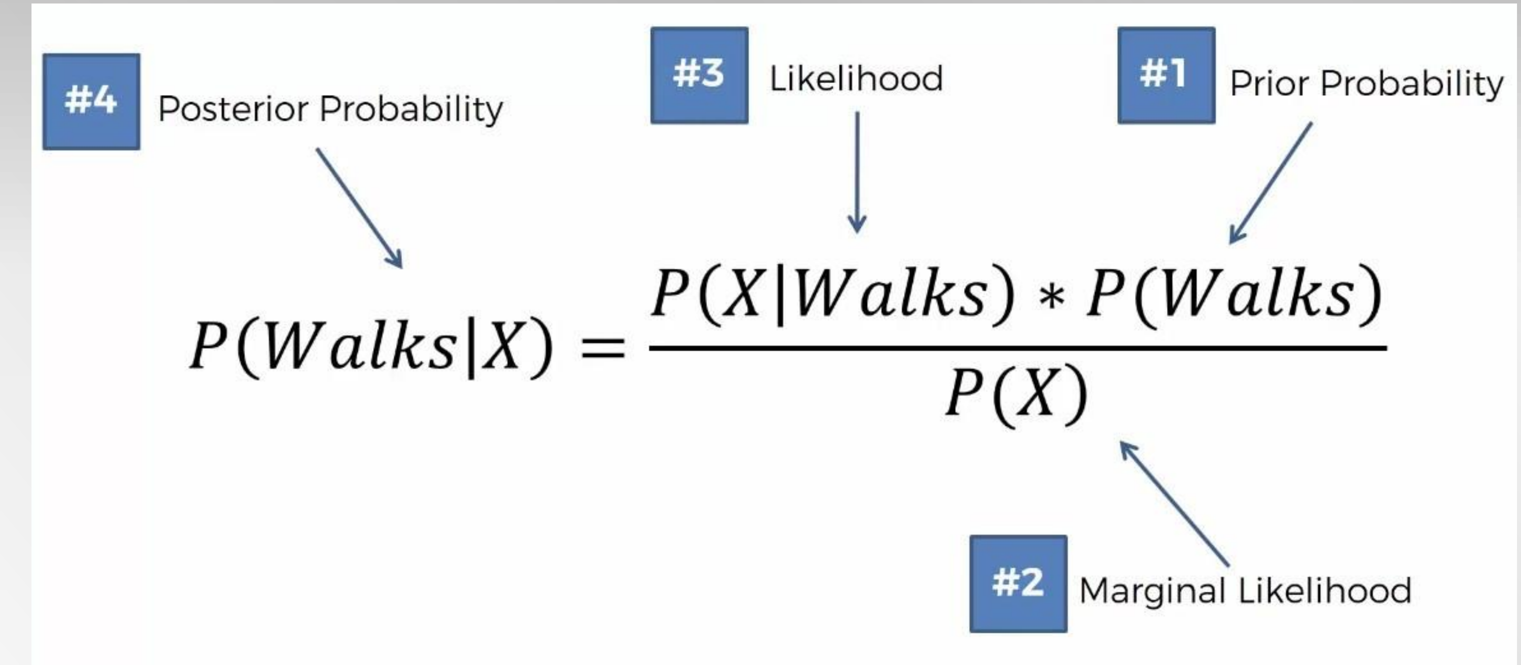
$$\hat{C} = \arg\max_{C} \left(P(C) \prod_{i=1}^{n} P(X_i \mid C) \right)$$

Illustrative Example



We are going to compare :

$$P(Walks|X)\ v.s.\ P(Drives|X)$$



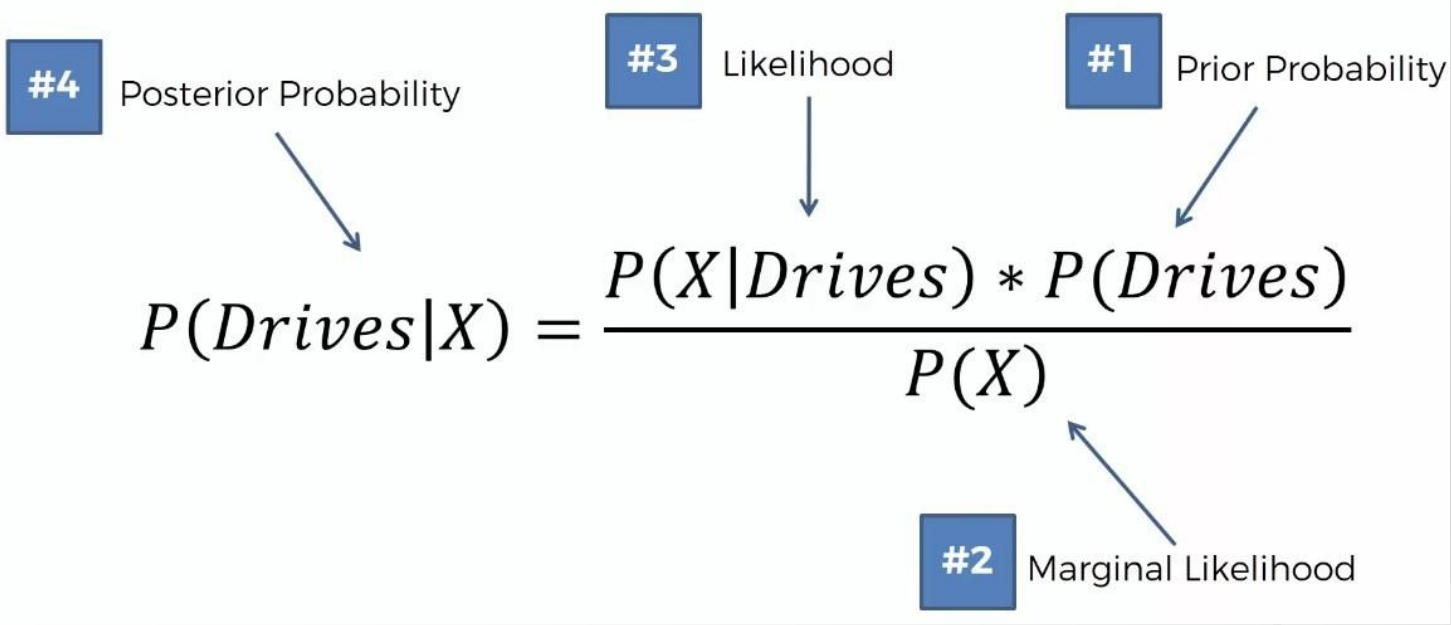
#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

$$P(Walks|X) = \frac{P(X|Walks) * P(Walks)}{P(X)}$$

#2 Marginal Likelihood



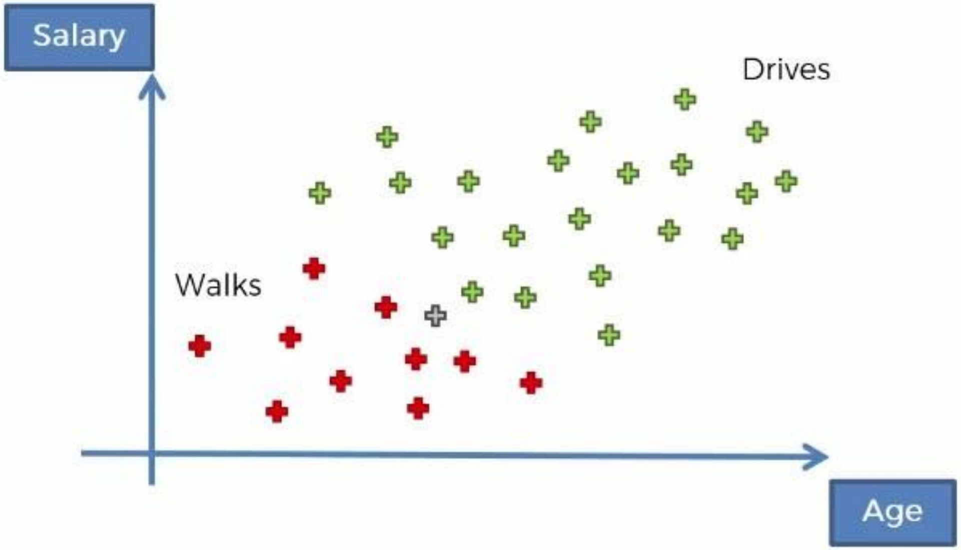
#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

$$P(Drives|X) = \frac{P(X|Drives) * P(Drives)}{P(X)}$$

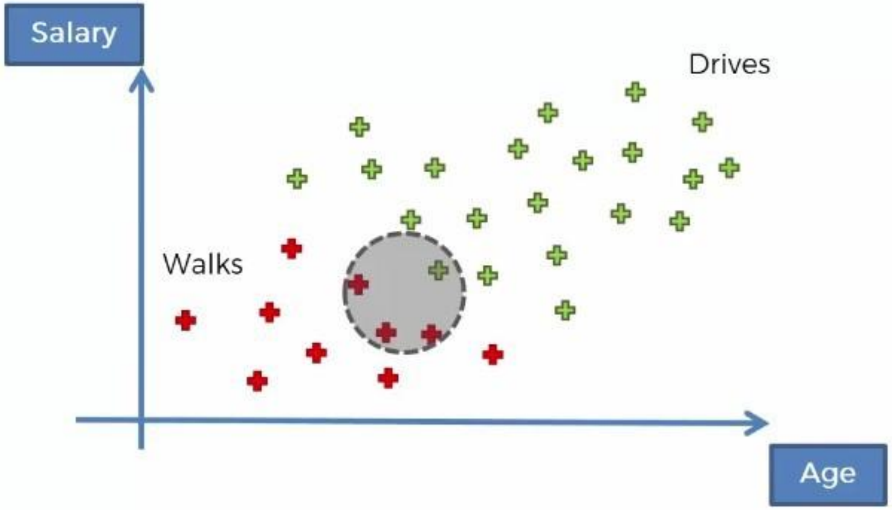
#2 Marginal Likelihood



#1. P(Walks)

$$P(Walks) = \frac{Number\ of\ Walkers}{Total\ Observations}$$

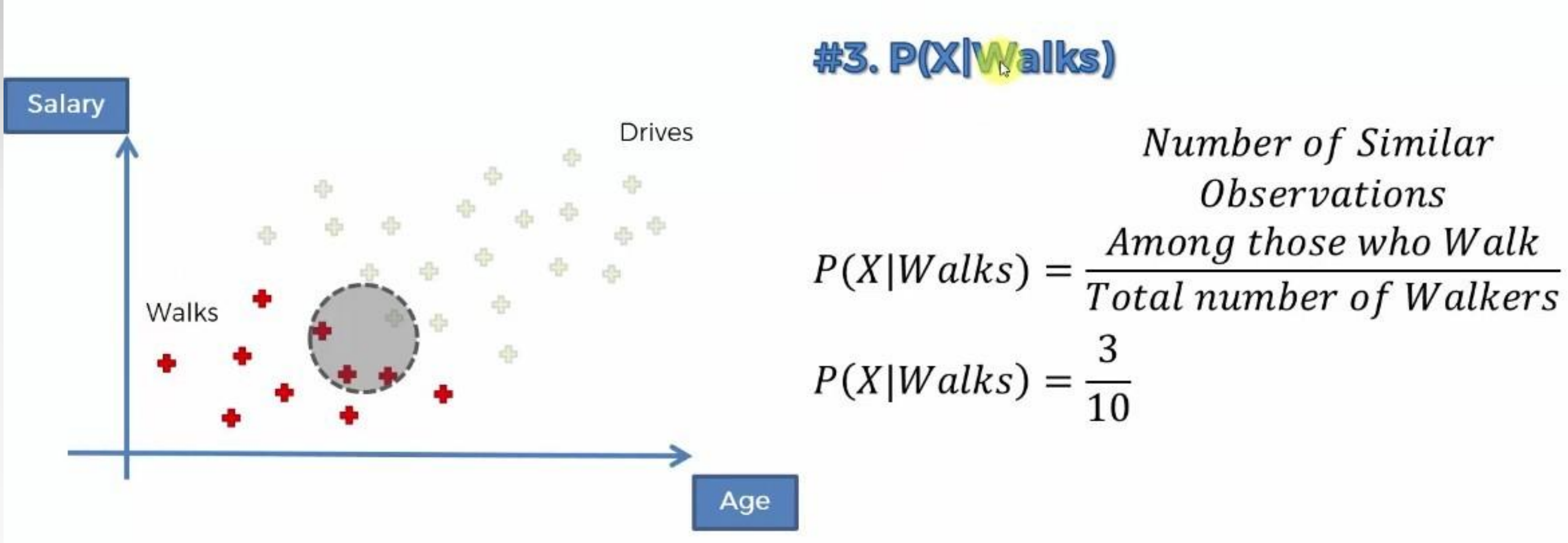
$$P(Walks) = \frac{10}{30}$$



#2. P(X)

$$P(X) = \frac{Number\ of\ Similar\ Observations}{Total\ Observations}$$

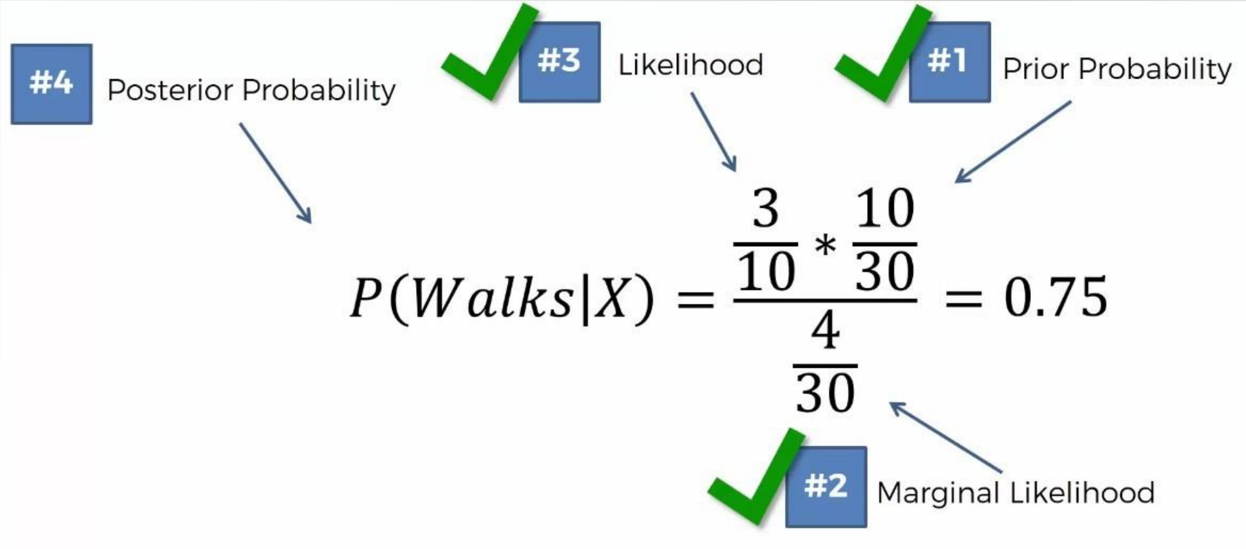
$$P(X) = \frac{4}{30}$$



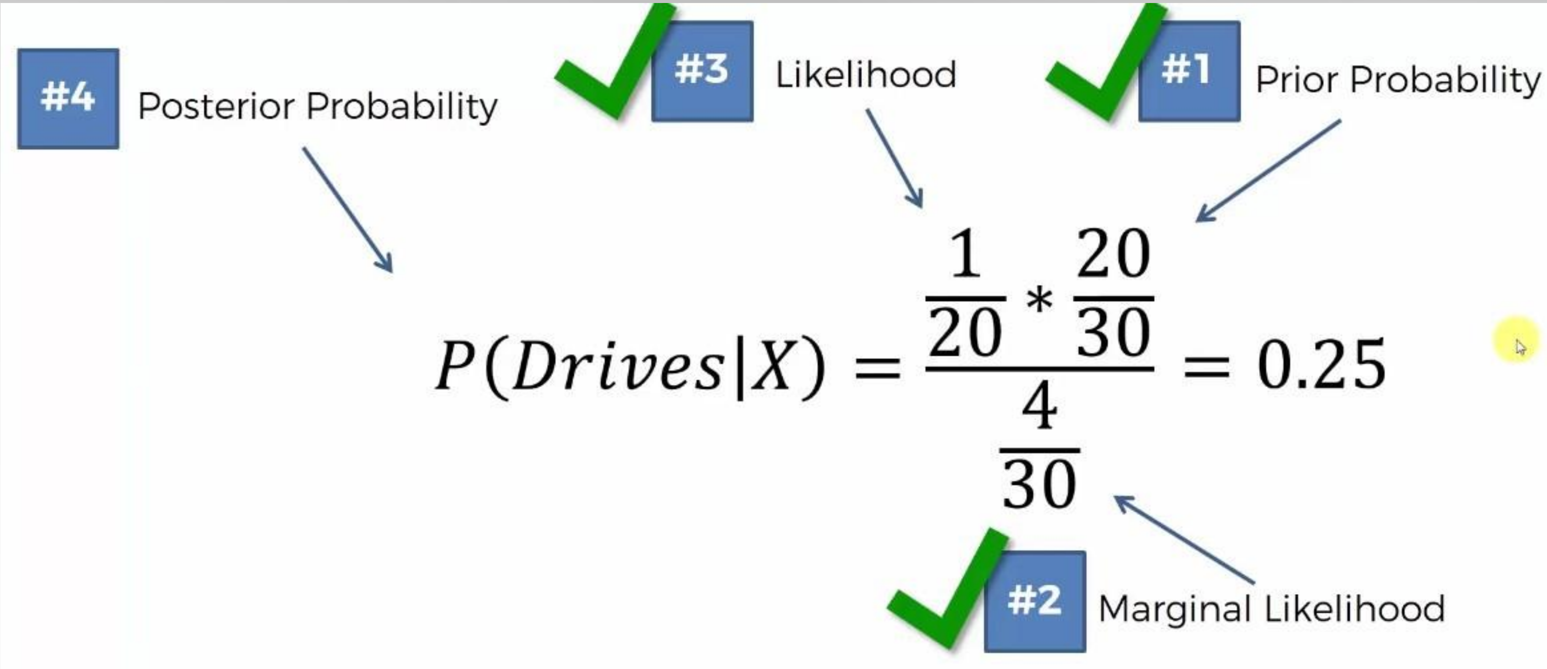
#3. P(X|Walks)

$$P(X|Walks) = \frac{\begin{array}{c}Number\ of\ Similar\\ Observations\\ Among\ those\ who\ Walk\end{array}}{Total\ number\ of\ Walkers}$$

$$P(X|Walks) = \frac{3}{10}$$



$$P(Walks|X) = \frac{\frac{3}{10} * \frac{10}{30}}{\frac{4}{30}} = 0.75$$



#4 Posterior Probability

✓ #3 Likelihood

✓ #1 Prior Probability

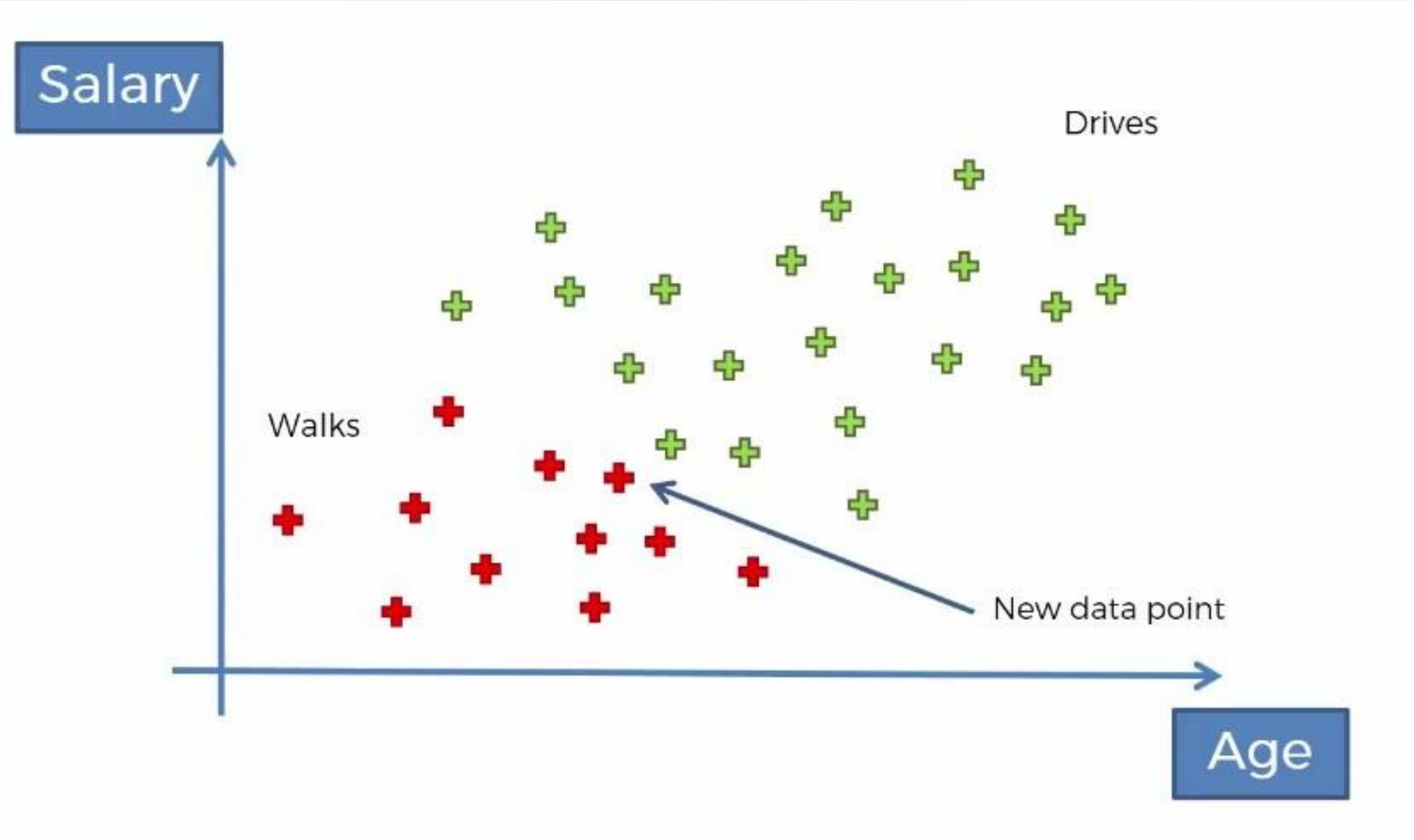
$$P(Drives|X) = \frac{\frac{1}{20} * \frac{20}{30}}{\frac{4}{30}} = 0.25$$

✓ #2 Marginal Likelihood

$$P(Walks|X)\ v.s.\ P(Drives|X)$$

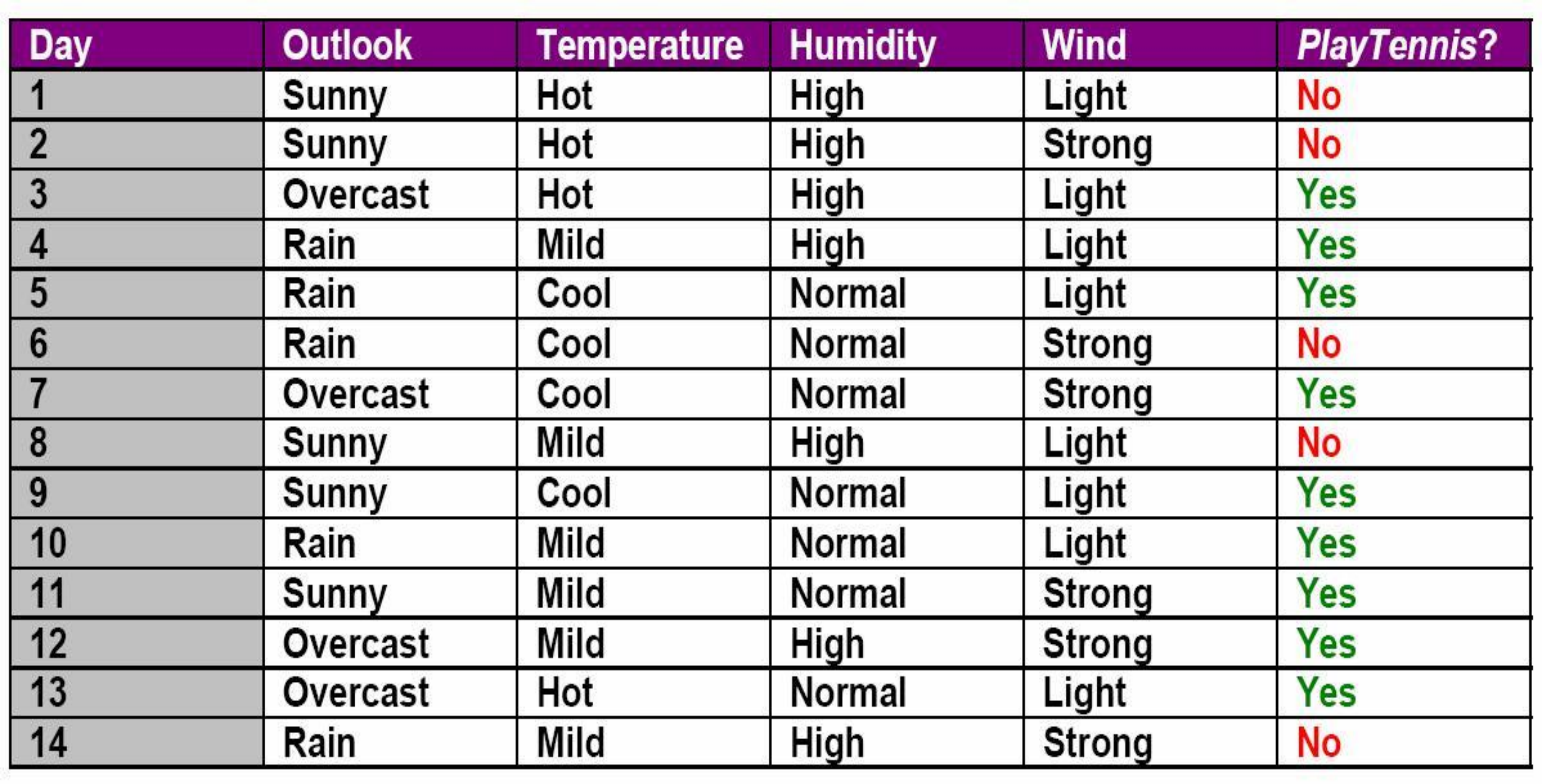
$$0.75\ v.s.\ 0.25$$

Final Classification



Naive Bayes Classifier

Tennis-Example



Day	Outlook	Temperature	Humidity	Wind	*PlayTennis?*
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

Naive Bayes Classifier

- Problem:

Use training data from above to classify the following instances:

a) <Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>

b) <Outlook=overcast, Temperature=cool, Humidity=high, Wind=strong>

Naive Bayes Classifier

<Outlook=sunny, Temperature=cool, Humidity=high,
Wind=strong>

Answer to (a):

P(PlayTennis=yes) = 9/14 = 0.64

P(PlayTennis=n) = 5/14 = 0.36

P(Outlook=sunny|PlayTennis=yes) = 2/9 = 0.22

P(Outlook=sunny|PlayTennis=no) = 3/5 = 0.60

P(Temperature=cool|PlayTennis=yes) = 3/9 = 0.33

P(Temperature=cool|PlayTennis=no) = 1/5 = .20

P(Humidity=high|PlayTennis=yes) = 3/9 = 0.33

P(Humidity=high|PlayTennis=no) = 4/5 = 0.80

P(Wind=strong|PlayTennis=yes) = 3/9 = 0.33

P(Wind=strong|PlayTennis=no) = 3/5 = 0.60

Naive Bayes Classifier

P(yes)xP(sunny|yes)xP(cool|yes)xP(high|yes)xP(strong|yes) = 0.0053

P(no)xP(sunny|no)xP(cool|no)xP(high|no)x P(strong|no) = 0.0206

So the class for this instance is ‘no’. We can normalize the probility by:

[0.0206]/[0.0206+0.0053] = 0.795

Naive Bayes Classifier

<Outlook=overcast, Temperature=cool, Humidity=high, Wind=strong>

Answer to (b):

P(PlayTennis=yes) = 9/14 = 0.64

P(PlayTennis=no) = 5/14 = 0.36

P(Outlook=overcast|PlayTennis=yes) = 4/9 = 0.44

P(Outlook=overcast|PlayTennis=no) = 0/5 = 0

P(Temperature=cool|PlayTennis=yes) = 3/9 = 0.33

P(Temperature=cool|PlayTennis=no) = 1/5 = .20

P(Humidity=high|PlayTennis=yes) = 3/9 = 0.33

P(Humidity=high|PlayTennis=no) = 4/5 = 0.80

P(Wind=strong|PlayTennis=yes) = 3/9 = 0.33

P(Wind=strong|PlayTennis=no) = 3/5 = 0.60

Naive Bayes Classifier

Estimating Probabilities:

In the previous example, P(overcast|no) = 0 which causes the formula-

P(no)xP(overcast|no)xP(cool|no)xP(high|no) xP(strong|nno) = 0.0

This causes problems in comparing because the other probabilities are not considered. We can avoid this difficulty by using m-estimate.

Naive Bayes Classifier : **M-Estimate**

We adjust the probability and avoid zero probabilities

Formula:

[c + k] / [n + m] where c/n is the original probability used before,

k=1 and m= the number of possible values for "Outlook" (e.g., sunny, overcast, rainy)..

Using this method our new values of probability is given below-

Naive Bayes Classifier

New answer to (b):

P(PlayTennis=yes) = 10/16 = 0.63

P(PlayTennis=no) = 6/16 = 0.37

P(Outlook=overcast|PlayTennis=yes) = 4+1/9+3 = 5/12 = 0.42

P(Outlook=overcast|PlayTennis=no) = 0+1/5+3 = 1/8 = .13

P(Temperature=cool|PlayTennis=yes) = 4/12 = 0.33

P(Temperature=cool|PlayTennis=no) = 2/8 = .25

P(Humidity=high|PlayTennis=yes) = 4/11 = 0.36

P(Humidity=high|PlayTennis=no) = 5/7 = 0.71

P(Wind=strong|PlayTennis=yes) = 4/11 = 0.36

P(Wind=strong|PlayTennis=no) = 4/7 = 0.57

Naive Bayes Classifier

P(yes) x P(overcast|yes) x P(cool|yes) x P(high|yes) x P(strong|yes) = 0.011

P(no) x P(overcast|no) x P(cool|no) x P(high|no) xP(strong|nno) = 0.00486

So the class of this instance is ‘yes’

Naive Bayes Classifier

- The conditional probability values of all the attributes with respect to the class are pre-computed and stored on disk.
- This prevents the classifier from computing the conditional probabilities every time it runs.
- This stored data can be reused to reduce the latency of the classifier.

Bayesian Belief Network

Bayesian Belief Network

- In Naïve Bayes Classifier we make the assumption of class conditional independence, that is given the class label of a sample, the value of the attributes are conditionally independent of one another.
- However, there can be dependences between value of attributes. To avoid this we use Bayesian Belief Network which provide joint conditional probability distribution.

Bayesian Networks

- A Bayesian network is a form of probabilistic graphical model.
- Specifically, a Bayesian network is a directed acyclic graph of nodes representing variables and arcs representing dependence relations among the variables.
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link = ‘direct influences’)
  - a conditional distribution (CPT) for each node given its parents:

    $P(X_i|Parents(X_i))$

Example

I’m at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn’t call. Sometime it’s set off by a minor earthquake. Is there a burglar?

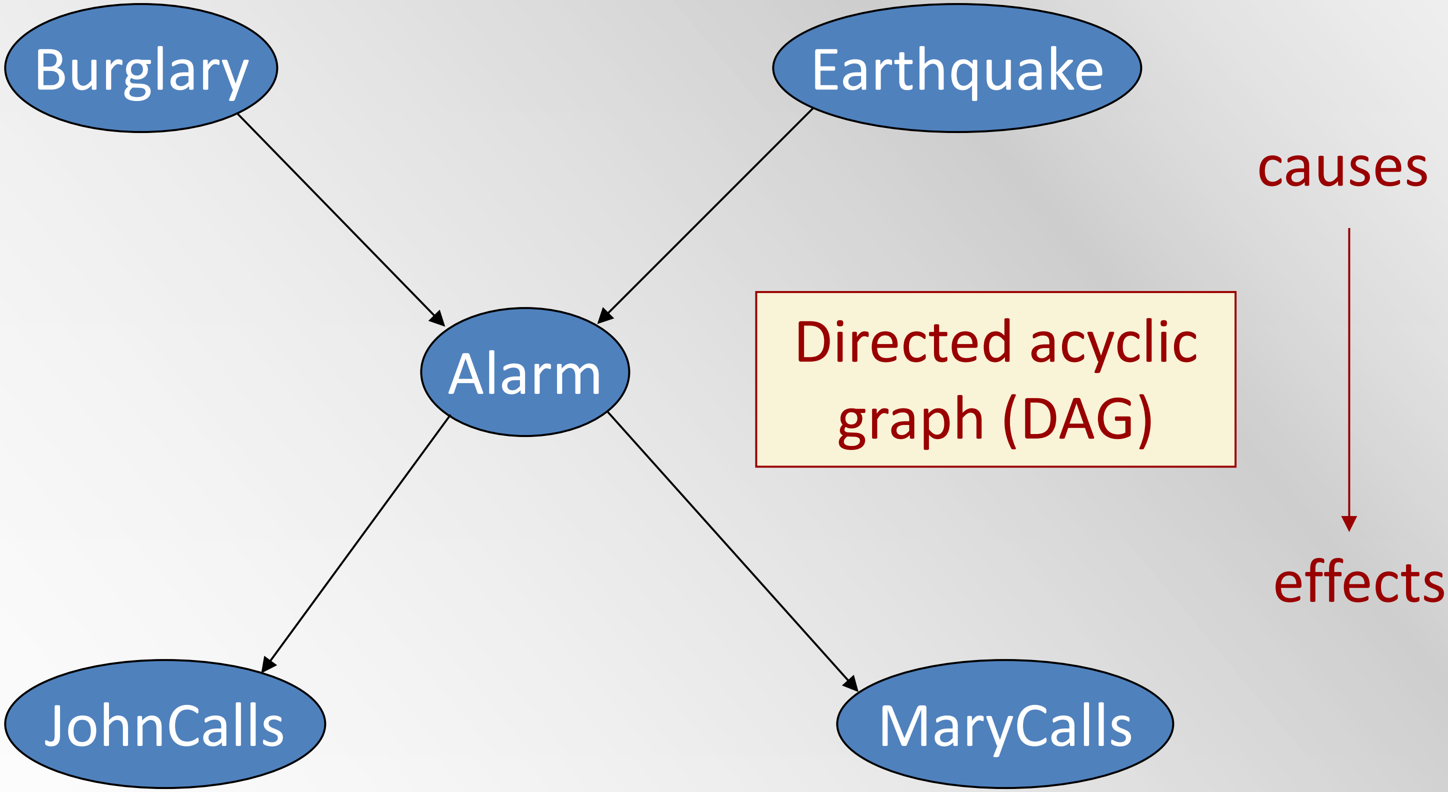
Variables: Burglar, Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects “causal” knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

A Simple Belief Network

- Nodes are random variables
- Edge from x to y has meaning of “x has direct influence on y

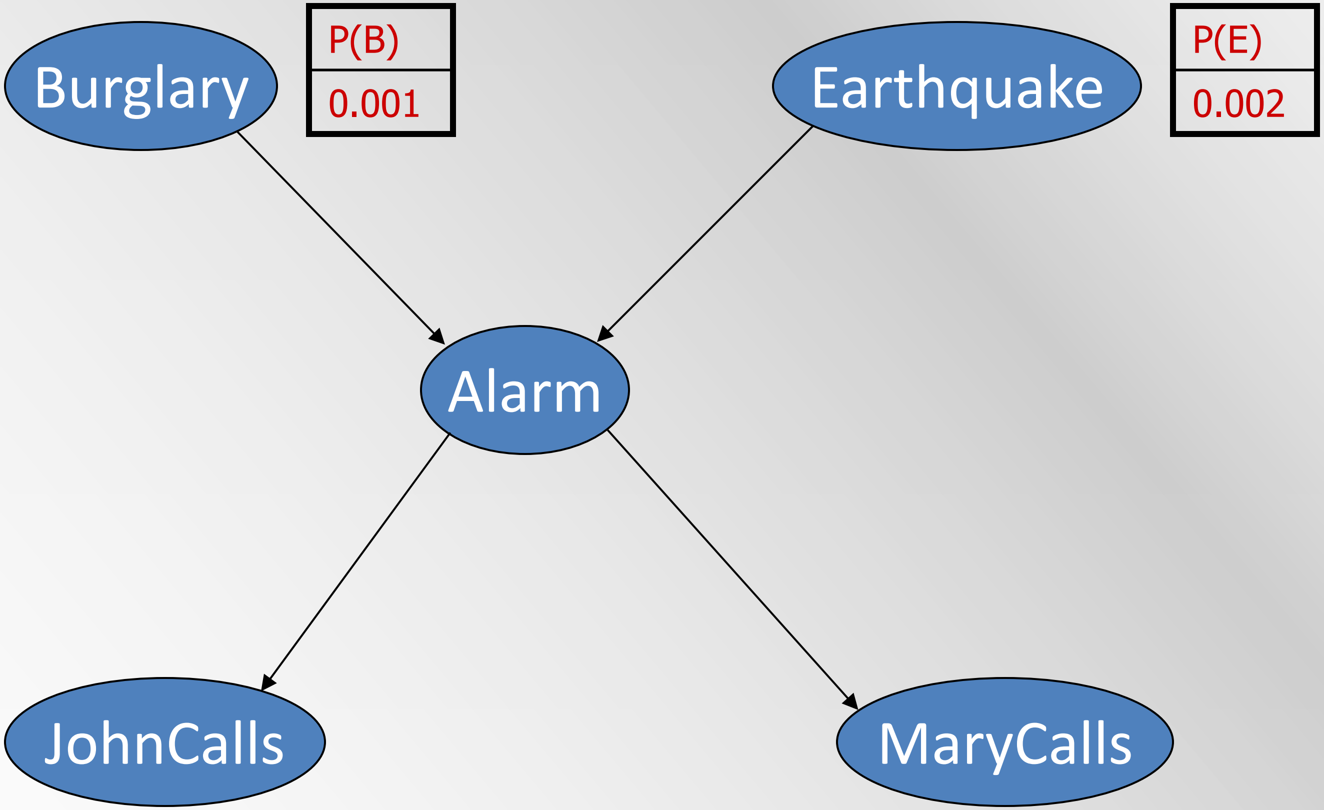


Directed acyclic graph (DAG)

Construction of Belief Network

- Identify Variables
  - □ Determine the key variables involved in the problem domain.
- Determine Relationships
  - □ Establish the causal or dependency relationships between these variables. This often involves domain knowledge and understanding of the problem context.
- Draw the Graph
  - □ Create a directed acyclic graph (DAG) where nodes represent variables and edges represent direct dependencies.
- Define Conditional Probability Tables (CPTs)
  - □ For each node, specify the probabilities of its possible values given its parent nodes.

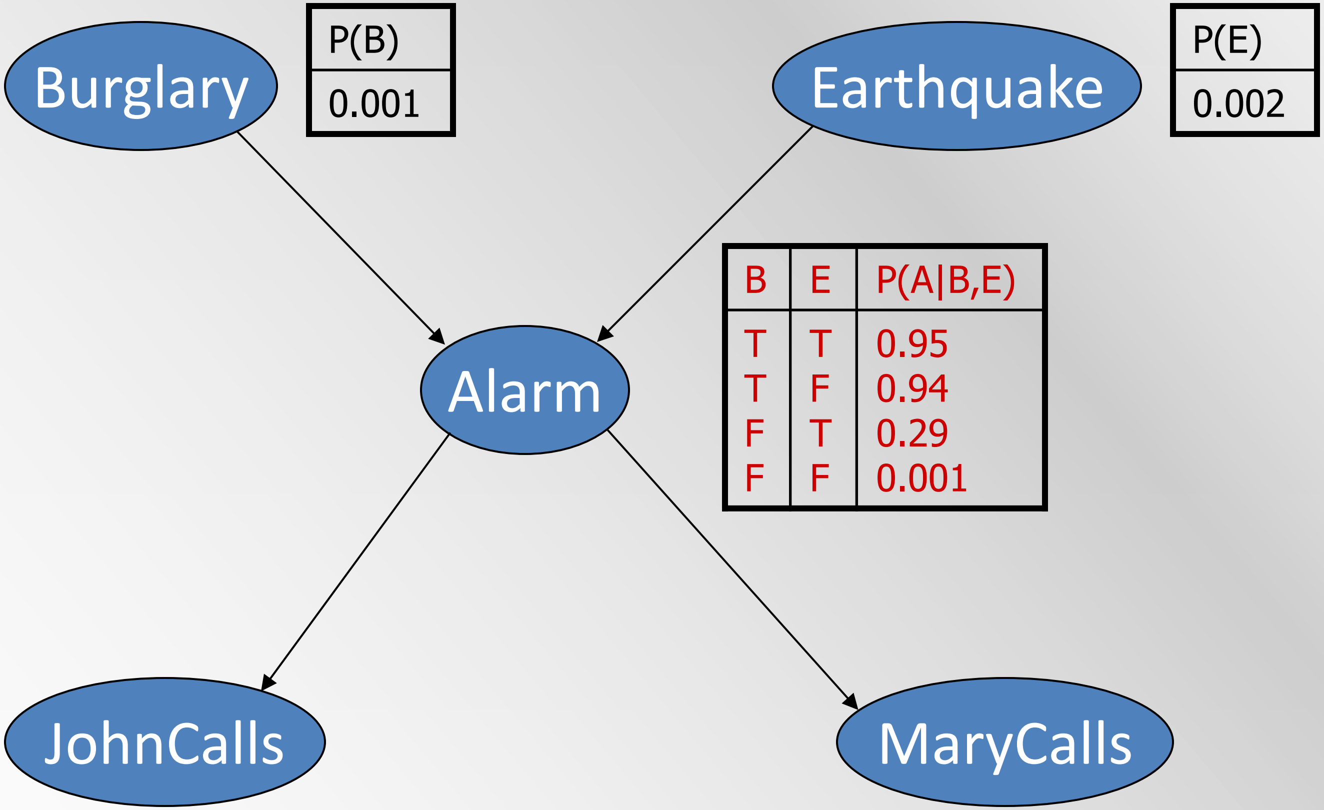
Assigning Probabilities to Roots



P(B)
0.001

P(E)
0.002

Conditional Probability Tables



Burglary

P(B)
0.001

Earthquake

P(E)
0.002

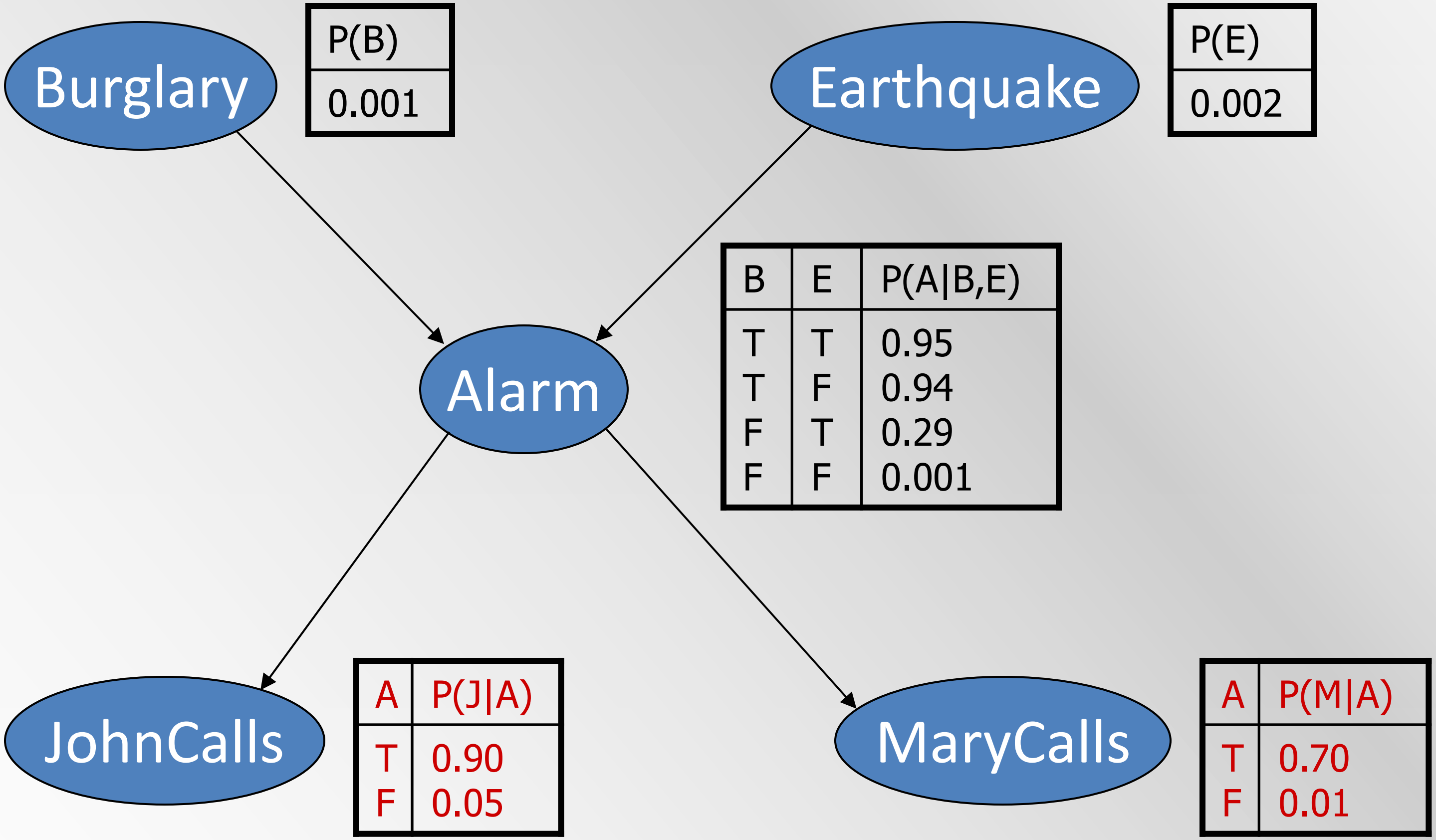
Alarm

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

JohnCalls

MaryCalls

Conditional Probability Tables



Burglary

Earthquake

Alarm

JohnCalls

MaryCalls

P(B)
0.001

P(E)
0.002

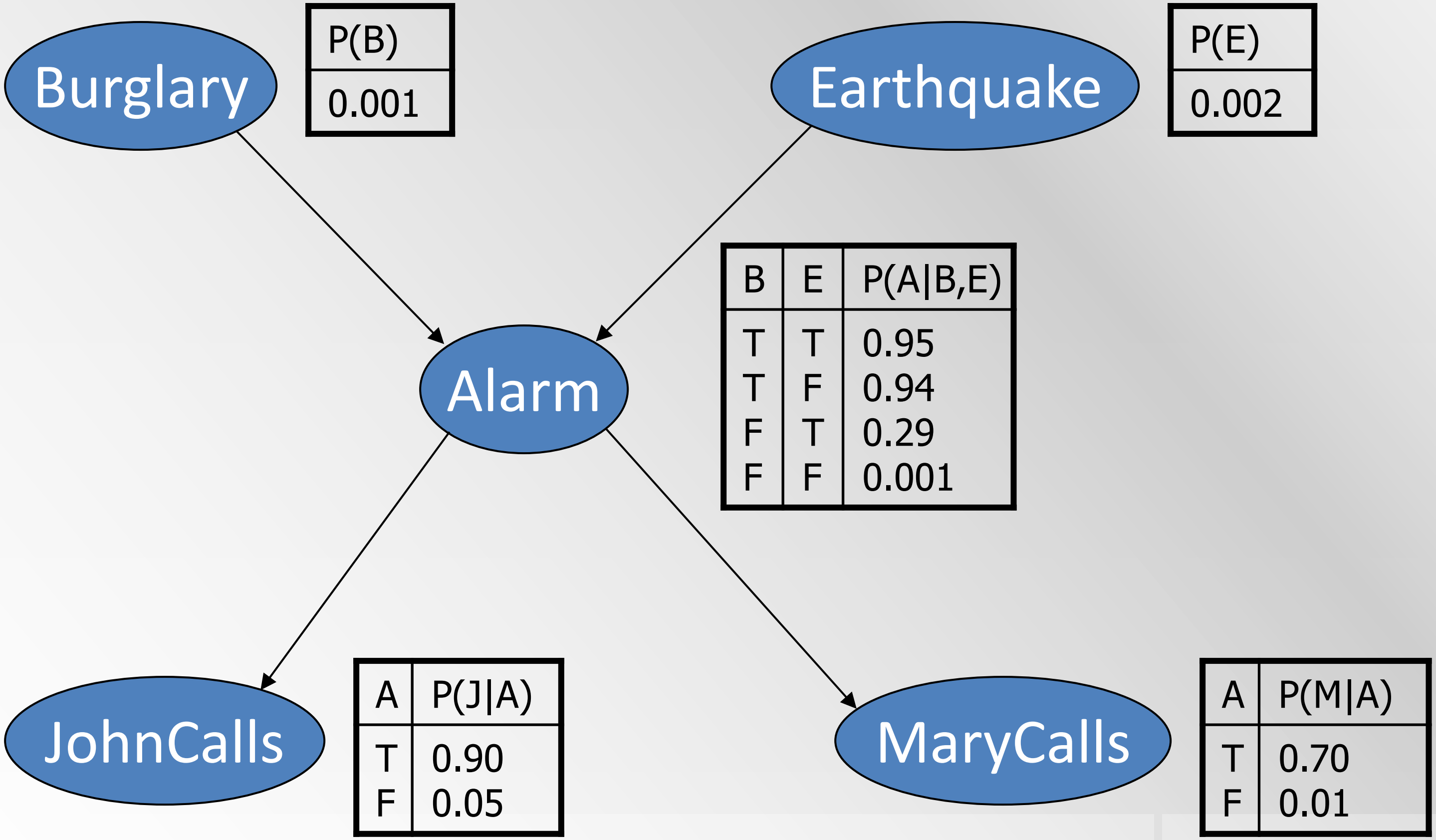
B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

What the BN Means

$$P(x_1,x_2,\ldots,x_n) = \prod_{i=1,\ldots,n} P(x_i|\mathrm{Parents}(X_i))$$



P(B)
0.001

P(E)
0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

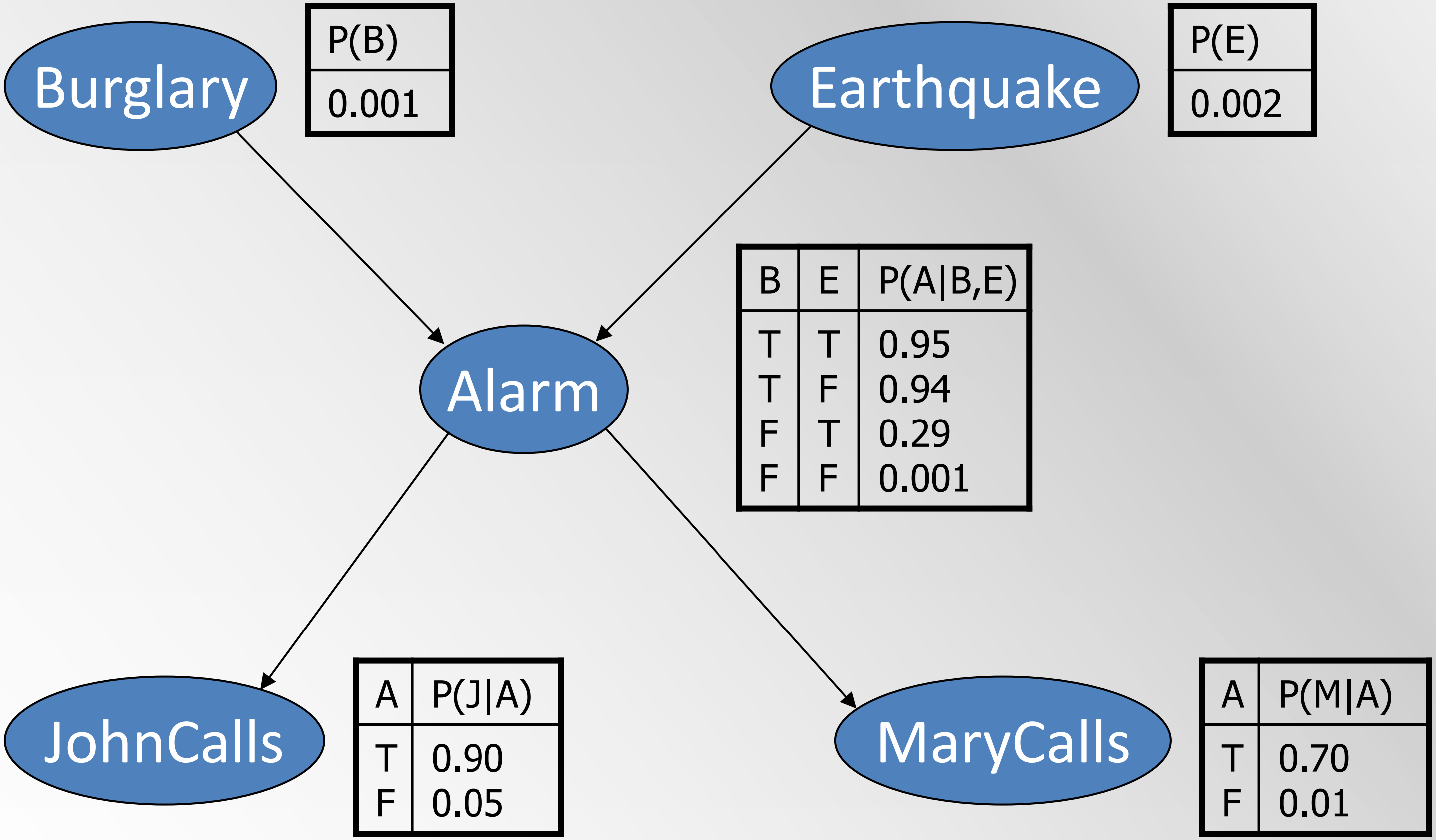
Calculation of Joint Probability

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$

$= P(J|A)P(M|A)P(A|\neg B,\neg E)P(\neg B)P(\neg E)$

$= 0.9 \text{ x } 0.7 \text{ x } 0.001 \text{ x } 0.999 \text{ x } 0.998$

$= 0.00062$



Burglary

Earthquake

Alarm

JohnCalls

MaryCalls

P(B)
0.001

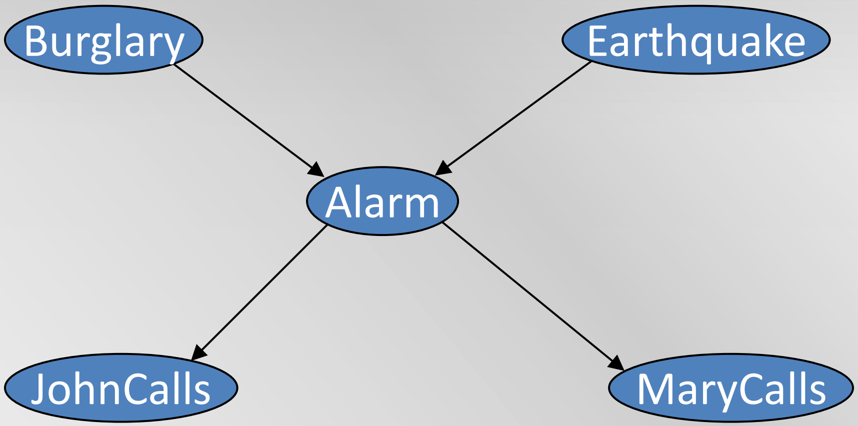
P(E)
0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

What The BN Encodes



- Each of the beliefs JohnCalls and MaryCalls is independent of Burglary and Earthquake given Alarm or ¬Alarm

For example, John does not observe any burglaries directly

- The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

For instance, given the alarm status, the reasons why John may not call are independent of the reasons why Mary may not call.

Structure of BN

E.g., JohnCalls is influenced by Burglary, but not directly. JohnCalls is directly influenced by Alarm

- The relation:

  $$P(x_1,x_2,\ldots,x_n) = \prod_{i=1,\ldots,n} P(x_i | \text{Parents}(X_i))$$

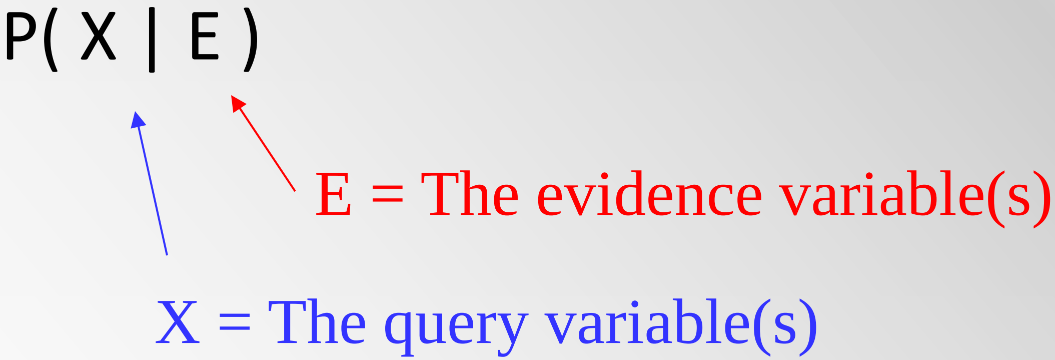
  means that each belief is independent of its predecessors in the BN given its parents

- Said otherwise, the parents of a belief X_i are all the beliefs that "directly influence" X_i

- Usually (but not always) the parents of X_i are its causes and X_i is the effect of these causes

Inference

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



$P(X \mid E)$

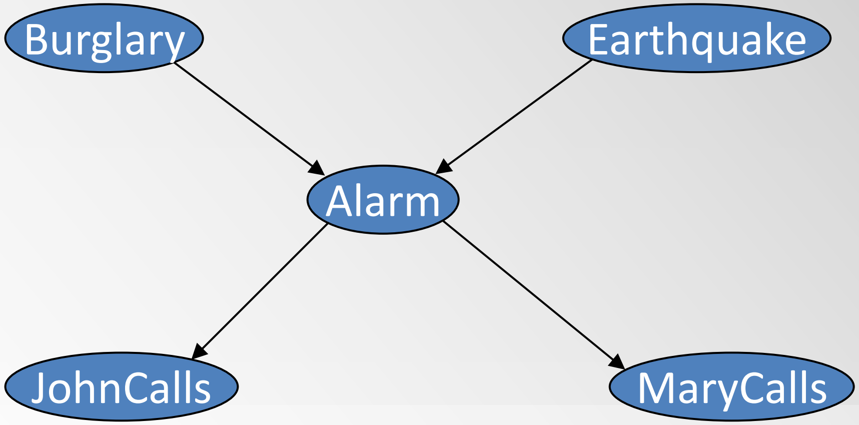
E = The evidence variable(s)

X = The query variable(s)

Inference: Example

- Target : P(B=true|J=true,M=true)
- Using Bayes' theorem:

$$P(B = \text{true}|J = \text{true}, M = \text{true}) = \frac{P(J = \text{true}, M = \text{true}|B = \text{true}) \cdot P(B = \text{true})}{P(J = \text{true}, M = \text{true})}$$



Inference: Example

P(B)
0.001

P(E)
0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

- Calculate P(J=true,M=true|B=true)
  - Using the law of total probability, we sum over all possible values of A and E:

$$P(J,M|B) = \sum_A \sum_E P(J,M,A,E|B)$$

  - According to the graph, we can simplify as :

$$= \sum_A \sum_E P(J = \text{true}|A) \cdot P(M = \text{true}|A) \cdot P(A|B = \text{true}, E) \cdot P(E)$$

Inference: Example

- Breaking it down for different A and E cases:
- For A=true, E = true

$$P(J = \text{true}|A = \text{true}) \cdot P(M = \text{true}|A = \text{true}) \cdot P(A = \text{true}|B = \text{true}, E = \text{true}) \cdot P(E = \text{true})$$
$$= 0.90 \cdot 0.70 \cdot 0.95 \cdot 0.02 = 0.01197$$

- For A=true, E = false

$$P(J = \text{true}|A = \text{true}) \cdot P(M = \text{true}|A = \text{true}) \cdot P(A = \text{true}|B = \text{true}, E = \text{false}) \cdot P(E = \text{false})$$
$$= 0.90 \cdot 0.70 \cdot 0.94 \cdot 0.98 = 0.057834$$

Inference: Example

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

- For A=false,E=true and A=false,*E*=false, they are just similar. We omit them here as the terms are negligible due to low probabilities of A=false.
- So, summing the above two terms :

$$P(J=\text{true}, M=\text{true}|B=\text{true}) \approx 0.01197 + 0.057834 = 0.069804$$

Inference: Example

- Then, we compute *P*(*J*=true,*M*=true)

$$P(J = \text{true}, M = \text{true}) = \sum_B \sum_A \sum_E P(J = \text{true}|A) \cdot P(M = \text{true}|A) \cdot P(A|B, E) \cdot P(B) \cdot P(E)$$

- This involves similar calculations as before, just further sum both *B*=true and *B*=false.
- Here, let's focus on the significant terms:
  - For B=true*B*=true, we already calculated 0.0698040.
  - For B=false*B*=false, similar calculations yield a smaller value due to lower probabilities of *A* being true.

Inference: Example

- Just assume P(*J*=true,*M*=true|*B*=false)≈0.0001.

$$P(J = \text{true}, M = \text{true}) \approx 0.01 \cdot 0.069804 + 0.99 \cdot 0.0001 = 0.00079804$$

Inference: Example

- Finally, we got all components, and calculate our result as :

$$P(B = \text{true} | J = \text{true}, M = \text{true}) = \frac{P(J = \text{true}, M = \text{true} | B = \text{true}) \cdot P(B = \text{true})}{P(J = \text{true}, M = \text{true})}$$

$$= \frac{0.069804 \cdot 0.01}{0.00079804} \approx 0.874$$

- So, the probability of a burglary given that both John and Mary have called is approximately 87.4%.

Advantages of Bayesian Approach

- Bayesian networks can readily handle incomplete data sets.
- Bayesian networks allow one to learn about causal relationships
- Bayesian networks readily facilitate use of prior knowledge.

Summary

- Bayesian Theorem and Conditional Probability
- Naive Bayes Classifier
- Bayesian Belief Network