

# Artificial Intelligence

Instructor: Kietikul Jearanaitanakij

Department of Computer Engineering

King Mongkut's Institute of Technology Ladkrabang

soon update gantt chart //

# Lecture 7

## Quantifying Uncertainty

- Prior probability
- Conditional probability
- Joint probability distribution
- Bayes' rule and its use
- Probabilistic reasoning (Bayesian network)
- Useful techniques for manipulating probability
- Naive Bayes Classifier

# Uncertainty

→ មហាវិស័យដែល agent មិន ប្រាកដ (uncertain)

→ មហាវិស័យដែល agent មិន ប្រាកដ (uncertain) ដែលមាន លក្ខណៈ ប្រាកដ ឬ មិនប្រាកដ

- In some problem domains, the agent's knowledge can at best provide only a “degree of belief” in the relevant sentences, i.e., assign a numerical degree of belief between 0 and 1 to sentences.

**Prior probability** → ការប្រាកដដំបូងដែលមិនមានព័ត៌មានផ្សេងៗទៀត →  $P(A)$  ដែល

- Let  $P(A)$  be the unconditional or prior probability that the proposition  $A$  is true.

**Example:**  $P(\text{Cavity}) = 0.1$  ; for any patient.

- $P(A)$  can only be used when there is no other information.
- If some new information  $B$  is known, we have to reason with the **conditional probability** of  $A$  given  $B$ ,  $P(A|B)$ , instead of  $P(A)$ .

→ ការប្រាកដដែលមានព័ត៌មានផ្សេងៗទៀត

## Example of prior probability:

- $P(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$
- $P(\text{Weather})$  defines a probability distribution for the random variable Weather.

Weather	$P(\text{Weather})$
Hot	0.7
Rainy	0.2
Cold	0.08
Warm	0.02

## Conditional probability

- Once the agent has obtained some evidence concerning the previously unknown propositions making up the domain, prior probabilities are no longer applicable.
- Instead, we use conditional or posterior probabilities,  $P(A | B)$ .

### Example

$$P(\text{Cavity} | \text{Toothache}) = 0.8$$

- $P(A | B)$  can only be used when all we know is B.
- If we also know C, we must compute  $P(A | B \wedge C)$  instead of  $P(A | B)$ .

- We can represent the conditional probability in table form as follow.

$P(A B)$		A	
		True	False
B	Hot	0.5	0.2
	Cold	0.1	0.2

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

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

- Conditional probability** can be calculated by

$$P(A|B) = \frac{P(A \cap B)}{P(B)} ; P(B) > 0.$$

- By moving a denominator to another side, we derive a **product rule** of two variables.

$$P(A \cap B) = P(A|B) \cdot P(B)$$

- Note that  $P(A \cap B) \equiv P(A, B)$ . *ସମସ୍ତ ସମ୍ଭାବନା*

## The joint probability distribution

The joint probability distribution  $P(x_1, x_2, \dots, x_n)$  assigns probabilities to all possible atomic event.

	Toothache	$\neg$ Toothache
Cavity	0.04	0.06
$\neg$ Cavity	0.01	0.89

Sum to 1.0

Atomic event

- Warning: when you see a probability table, it could be the conditional probability or the joint probability distribution. Read its description before using it.
- Atomic events are mutually exclusive.

$$P(\text{Cavity}) = 0.10$$

$$P(\text{Cavity} \vee \text{Toothache}) = 0.04 + 0.06 + 0.01 = 0.11$$

$$P(\text{Cavity} \mid \text{Toothache}) = \frac{P(C \wedge T)}{P(T)} = \frac{0.04}{(0.04 + 0.01)} = 0.8$$

## Bayes' rule and its use

$$P(A \cap B) = P(A|B) \cdot P(B) \quad \dots\dots\dots(1)$$

$$P(A \cap B) = P(B|A) \cdot P(A) \quad \dots\dots\dots(2)$$

$$(1) = (2), \quad P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \quad ; \text{ Bayes' rule.}$$

$$\text{Similarly, } P(Y|X, E) = \frac{P(X|Y, E) \cdot P(Y|E)}{P(X|E)}$$

Proof:

ଏକାଞ୍ଚିତା ନିୟମ

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

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

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

ଏକାଞ୍ଚିତା ନିୟମ



**Example of Bayes' rule** → (ဥပမာ) (ဝိသေသန) သိကမ်း ဟု ခေါ်သည်။ <sup>ရှာဖွေရန်</sup>

In medical diagnosis, the doctor **knows**  $P(\text{symptoms} \mid \text{disease})$  and **want to derive a diagnosis,  $P(\text{disease} \mid \text{symptoms})$** . <sup>သိကမ်း</sup> **For example**, a doctor knows that the disease meningitis <sup>မိနင်ဂိုက်တစ်</sup> causes the patient to have a stiff neck, say, 70% of the time.

$$P(s \mid m) = 0.7 \quad \text{ခေါ်ဝေါ်ပါသည်။ (မိနင်ဂိုက်တစ်) သောအခါ S = 0.7}$$

The doctor also knows some unconditional facts: the prior probability that a patient has meningitis is 1/50,000,

$$P(m) = 1/50000 \quad \text{မူလက သိသော အချက်}$$

and the prior probability that any patient has a stiff neck is 1%.

$$P(s) = 0.01 \quad \text{မူလက သိသော အချက်} \quad = 0.0014$$

We expect a patient with a stiff neck to have meningitis with probability

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014 \quad \text{← ရှာဖွေရန်}$$

Notice that even though a stiff neck is quite strongly indicated by meningitis (with probability 0.7), the probability of meningitis in the patient remains small. This is because the prior probability of stiff necks is much higher than that of meningitis.

↓  
ဒီအရာကို မလေးချင်ဘူး

# Probabilistic Reasoning Systems

We saw that the full joint probability distribution can answer any question about the domain, but can become intractably large as the number of variables grows.

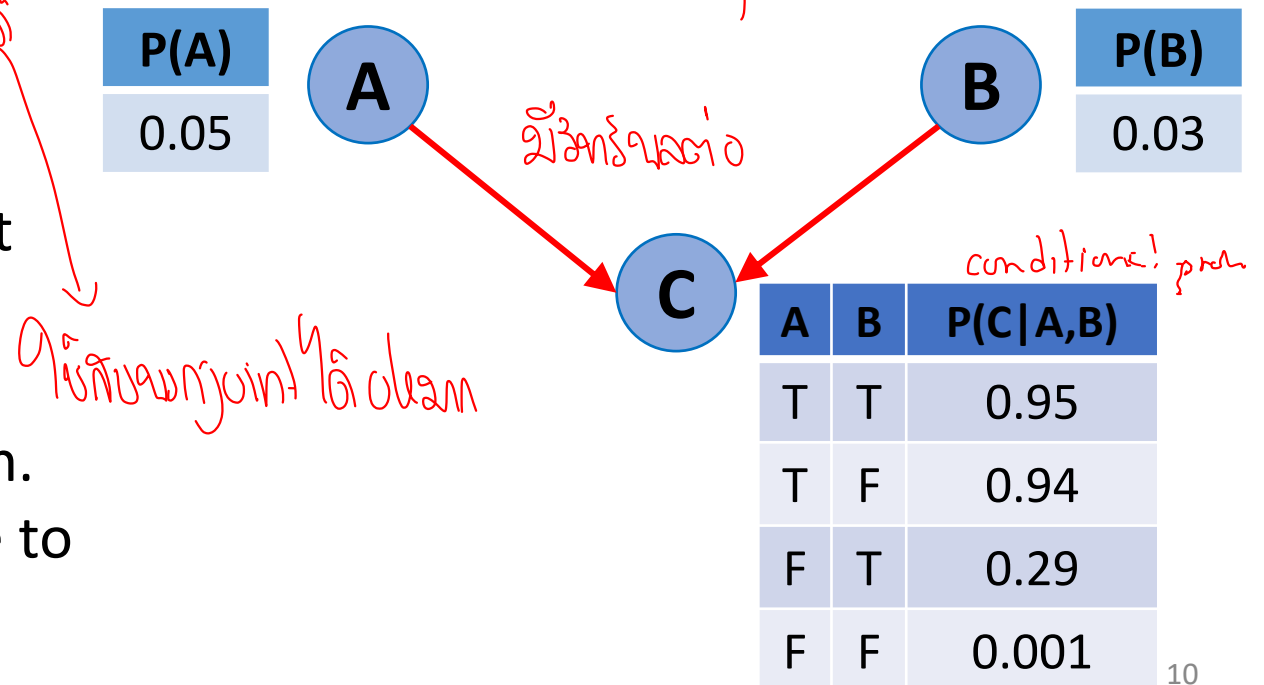
$$P(A \wedge B \wedge C \wedge D) = P(A | B \wedge C \wedge D) \cdot P(B \wedge C \wedge D)$$

Not easy to calculate  
မလေးဘူး conditional prob  
ဒါဟာ conditional prob ဟုလည်းကောင်း

## Bayesian network

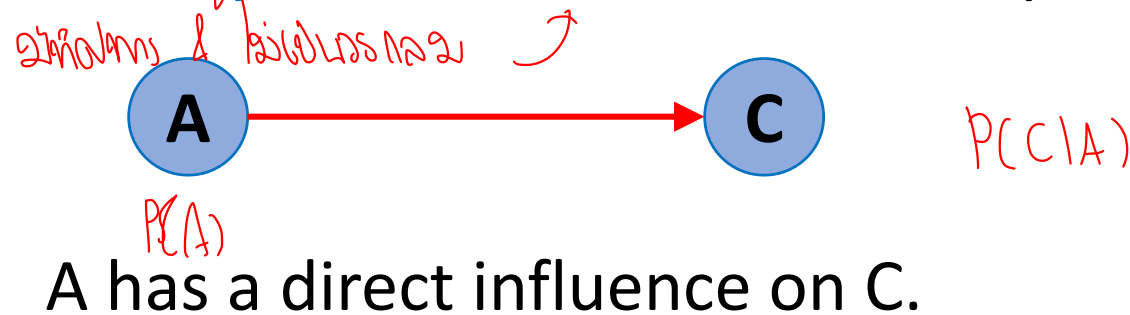
ဘာကို bayesian network ဟုလည်းကောင်း

- A data structure that represents the dependencies among variables and to give a concise specification of the joint probability distribution.
- A Bayesian network can be far more compact than the full joint distribution.
- This property is what makes it feasible to handle domains with many variables.



## The full specification of a Bayesian network is as follows:

- Each node corresponds to a random variable.
- A set of directed links or arrows connects pairs of nodes. If there is an arrow from node A to node C, A is said to be a parent of C. The graph has **no directed cycles**; hence a directed acyclic graph, or DAG.



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

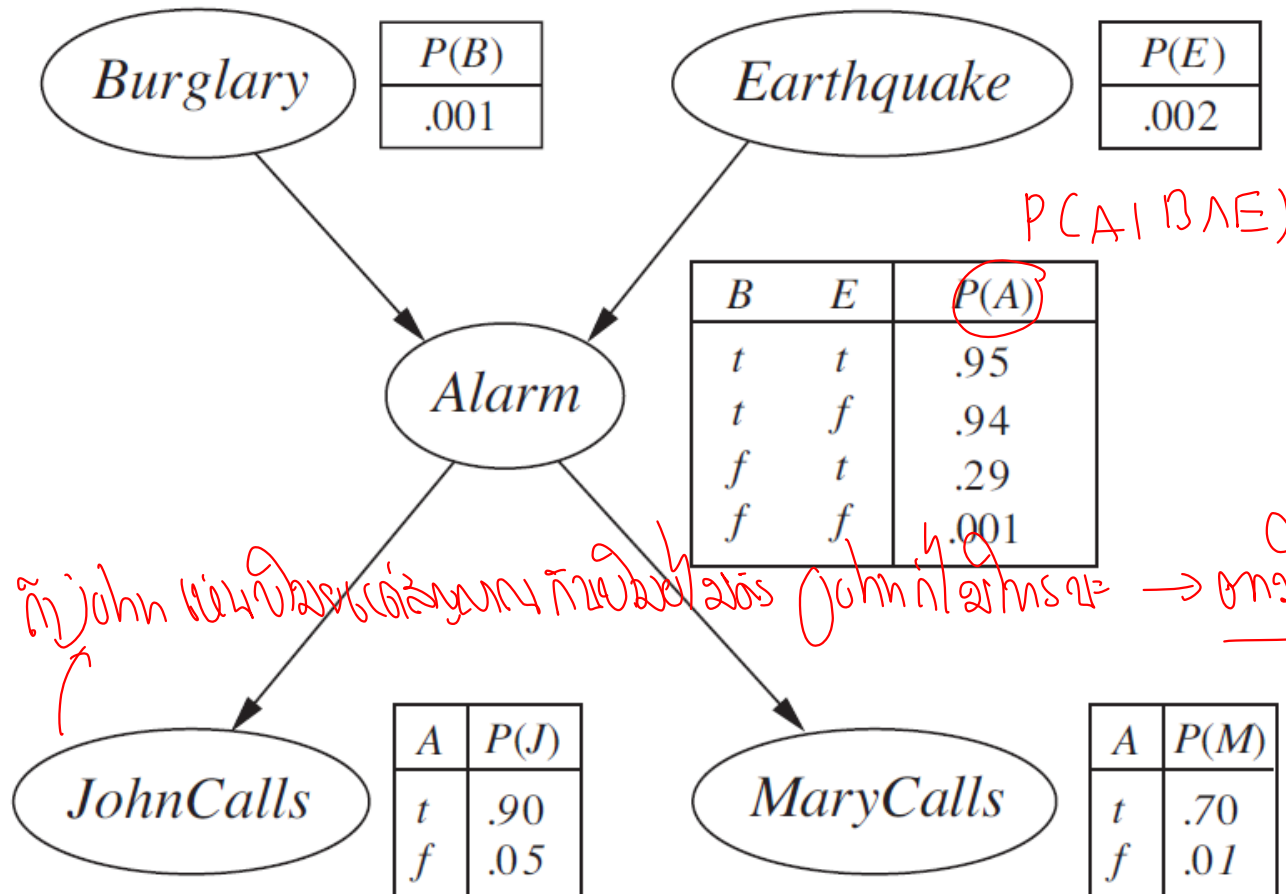
## Example : Burglary alarm



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

การคำนวณ Bayesian network ได้ดังนี้

Drawing a directed acyclic graph, or DAG, to represent the Bayesian network.



တိကျသော အချက်အလက်များကို အောက်တွင် ဖော်ပြပါမည်။

$$P(A|B \cap E)$$

## Did you see something unreasonable ?



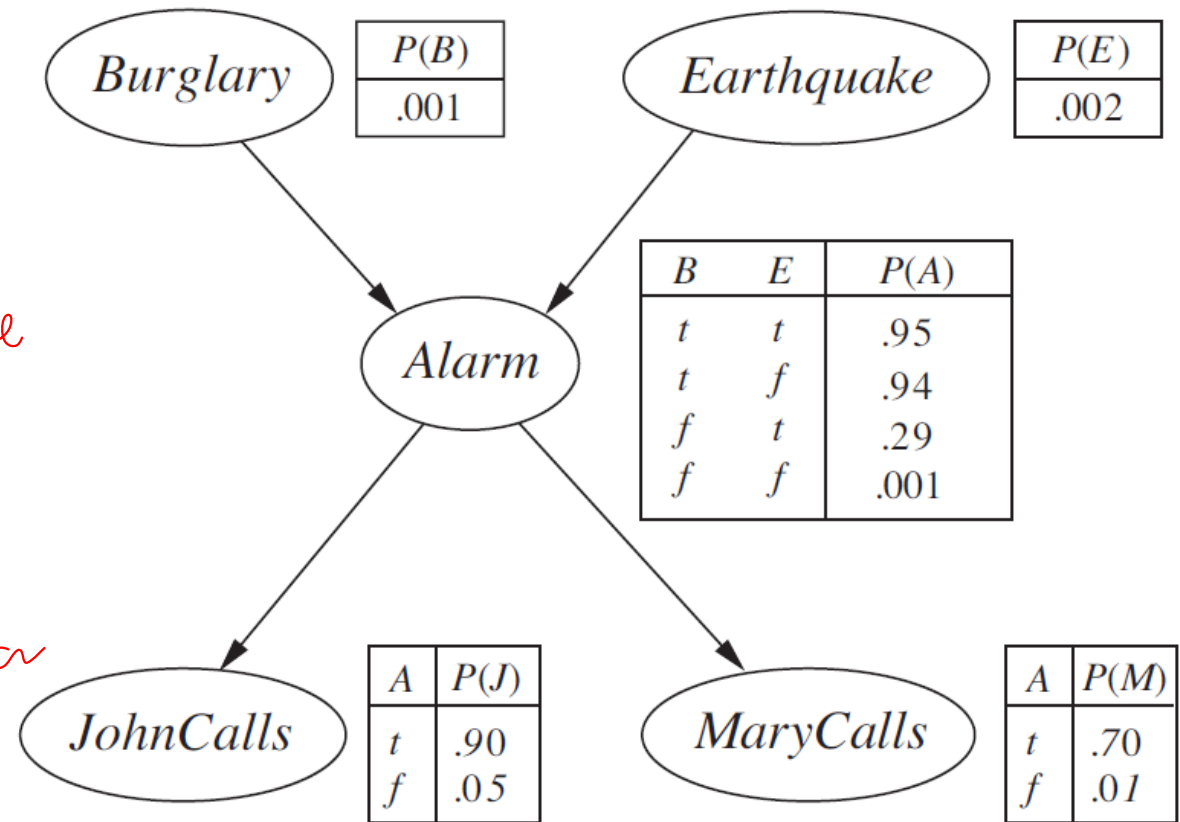
1. John kann das Gesetz nicht verstehen → Handlung

We can calculate the probability that **the alarm has sounded**, but **neither a burglary nor an earthquake has occurred**, and **both John and Mary call**.

$P(j, m, a, \neg b, \neg e) = (?)$  *တကယ်တမ်းမသိရဘူး -> sense မရှိဘူး*

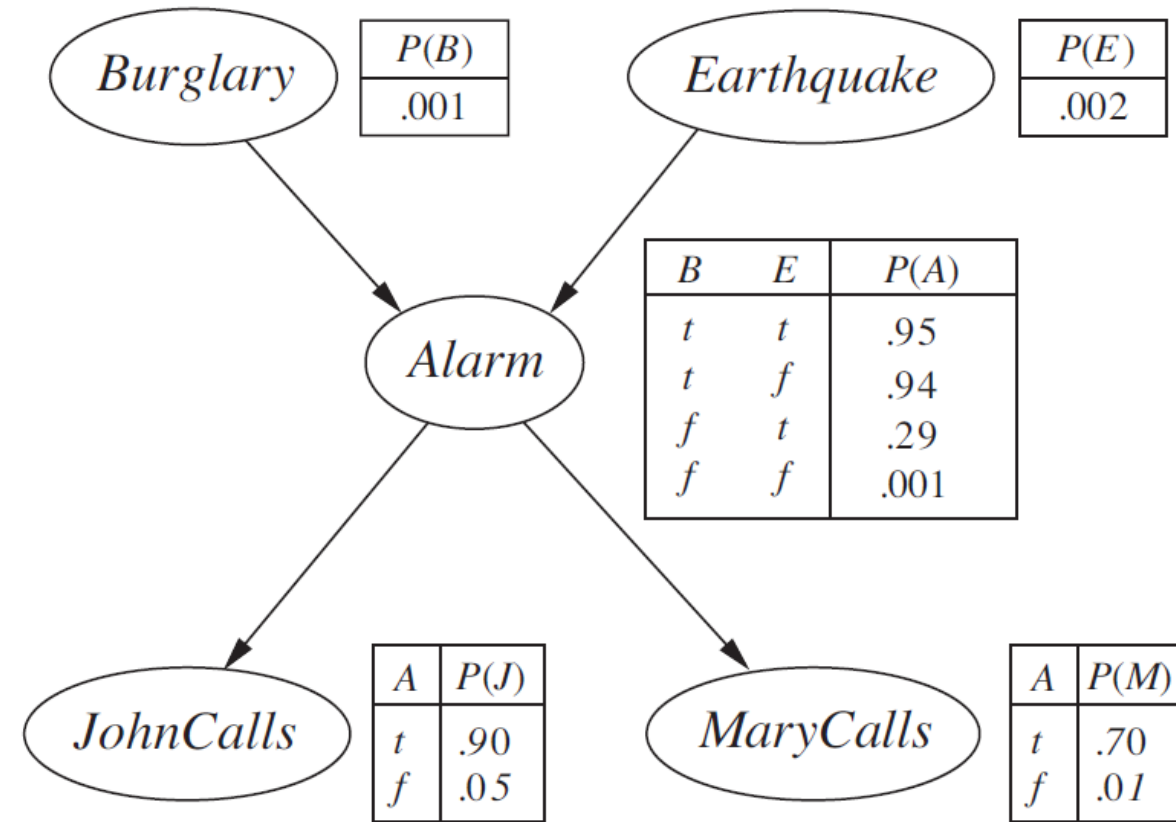
The joint probability distribution can be represented by

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parents}(X_i)) .$$
 *production*  $\Sigma \rightarrow$  summation *အသံအသံ*



We multiply entries from the joint distribution (using single-letter names for the variables):

b, e, a, j, and m stand for Burglary, Earthquake, Alarm, JohnCalls, and MaryCalls, respectively.



$$\begin{aligned}
 P(j, m, a, \neg b, \neg e) &= \underbrace{P(j | a)}_{\text{red underline}} \underbrace{P(m | a)}_{\text{red underline}} \underbrace{P(a | \neg b \wedge \neg e)}_{\text{red underline}} \underbrace{P(\neg b)}_{\text{red underline}} \underbrace{P(\neg e)}_{\text{red underline}} \\
 &= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = \underline{0.000628} .
 \end{aligned}$$

น้อยนิด



# Useful techniques for manipulating probability

**Case study:** A domain consisting of 3 Boolean variables **Toothache**, **Cavity**, and **Catch** (whether the dentist's steel probe catches something wrong in my tooth).

$P(\text{Toothache}, \text{Cavity}, \text{Catch})$  is shown in 2x2x2 table as follows:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

**Figure 13.3** A full joint distribution for the *Toothache*, *Cavity*, *Catch* world.

- Notice that the probabilities in the joint distribution sum to 1.
- $P(\text{cavity} \vee \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$
- We can extract the distribution over some subset of variables or a single variable by  
 $P(\text{cavity}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.20$  ; **Marginal probability** of cavity

(or **summing out**)



- **Marginalization:**  $P(Y) = \sum_{z \in Z} P(Y, z)$  only  
ခံယူ

Here we sum over all the possible combinations of values of the set of variables  $Z$ .

$$P(Cavity) = \sum_{z \in \{Catch, Toothache\}} P(Cavity, z)$$

- **Conditioning:**

Instead of using joint probabilities, we can apply the conditional probability to calculate the marginalization,  $P(Y)$ .

$$P(Y) = \sum_z \underbrace{P(Y|z)} \underbrace{P(z)}_{\text{ခံယူလိမ့်မည်}}$$

- In most cases, we are more interested in computing conditional probabilities than full joint distribution since the size of probability table is smaller.

→ កំណត់អោយបានជា 1/P(toothache)

- **Normalization constant:**

$$\begin{aligned} P(cavity | toothache) &= \frac{P(cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6 . \end{aligned}$$

sum = 1

If the prior probability is not available, we can use a normalization constant to calculate the conditional probability.

→ កំណត់អោយបានជា 1/P(toothache)  
បានដឹង (sum = 1) ហើយយើងអាចគណនាបាន

$$\begin{aligned} P(\neg cavity | toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 . \end{aligned}$$

- Notice that in these two calculations the term **1/P(toothache)** remains constant, no matter which value of Cavity we calculate.
- In fact, it can be viewed as a **normalization constant**, denoted by  $\alpha$ , for the distribution  $P(\text{Cavity} | \text{toothache})$ , ensuring that it adds up to 1.

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

**Figure 13.3** A full joint distribution for the *Toothache*, *Cavity*, *Catch* world.

$$P(\text{Cavity} \mid \text{toothache}) = \alpha P(\text{Cavity}, \text{toothache})$$

$$= \alpha [P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg \text{catch})]$$

$$= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] = \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$$

Variable: *Cavity*  
 value of variable: *toothache*  
 normalization factor:  $\alpha$   
 marginalization:  $P(Y|Z)$   
 Divide each term by  $(0.12+0.08)$ .  $\frac{x_i}{\text{sum}(x)}$  msscale up

$$\frac{t_1}{h} + \frac{t_2}{h} = 1$$

$$\frac{1}{h}(t_1 + t_2) = 1$$

- We can calculate  $P(\text{Cavity} \mid \text{toothache})$  even if we don't know the value of  $P(\text{toothache})$ !
- Normalization turns out to be a useful shortcut in many probability calculations, both to make the computation easier and to allow us to proceed when some probability assessment (such as  $P(\text{toothache})$ ) is not available.

- **Bayes' rule : (Revisited)** *ግንባታ*

Bayes' rule can be viewed as the relationship between **cause** and **effect**.

$$\overbrace{P(\text{cause} | \text{effect})}^{\text{Diagnostic}} = \frac{\overbrace{P(\text{effect} | \text{cause})P(\text{cause})}^{\text{Causal}}}{P(\text{effect})}$$

- From previous example, a doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 70% of the time.
- Letting **s** be the proposition that the patient has a **stiff neck** and **m** be the proposition that the patient has **meningitis**.

$$\left. \begin{aligned} P(s | m) &= 0.7 \\ P(m) &= 1/50000 \\ P(s) &= 0.01 \end{aligned} \right\} \text{Given by a doctor.}$$

$$P(m | s) = \frac{P(s | m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014 .$$

## • Using independence of variables to simplify Bayes' rule

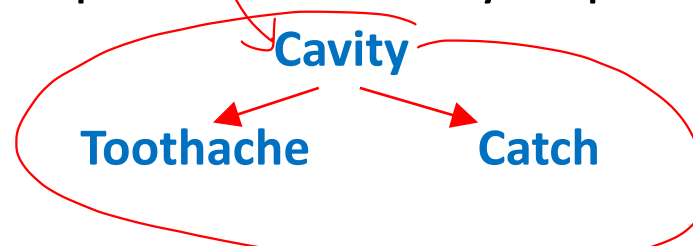
→ simplifying bayes' rule

- What happens when we have two or more pieces of evidence? For example, what can a dentist conclude if her steel probe catches in the aching tooth of a patient?

$$\frac{\mathbf{P}(\text{Cavity} \mid \text{toothache} \wedge \text{catch})}{= \alpha \mathbf{P}(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) \mathbf{P}(\text{Cavity})}$$

$$\alpha = \frac{1}{\mathbf{P}(\text{toothache} \wedge \text{catch})}$$

- We need to know the conditional probabilities of the conjunction (toothache  $\wedge$  catch) for each value of Cavity. That might be feasible for just two evidence variables, but it does **not scale up**.  
→ 1.512.000.000
- We can assert the notion of independence between variables to simplify the expression.  
→ 1.512.000.000
- In fact, **Toothache** and **Catch** are independent. Each is directly caused by **Cavity**, but neither has a direct effect on the other (toothache depends on the state of the nerves in the tooth, whereas the probe's accuracy depends on the dentist's skill).  
→ 1.512.000.000



toothache and catch are independent

- Here we have the conditional independence of toothache and catch given Cavity:

$$\mathbf{P}(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) = \mathbf{P}(\text{toothache} \mid \text{Cavity}) \mathbf{P}(\text{catch} \mid \text{Cavity})$$

non dependent

joint was variable

- Plugging this conditional independence into the equation in previous page, we have:

$$\begin{aligned} \mathbf{P}(\text{Cavity} \mid \text{toothache} \wedge \text{catch}) \\ = \alpha \mathbf{P}(\text{toothache} \mid \text{Cavity}) \mathbf{P}(\text{catch} \mid \text{Cavity}) \mathbf{P}(\text{Cavity}) . \end{aligned}$$

- This is good because the joint probability disappears.
- The conditional independence also decomposes the full joint distribution into pieces.

$$\begin{aligned} \mathbf{P}(\text{Toothache}, \text{Catch}, \text{Cavity}) &\leftarrow \text{bias} \\ &= \mathbf{P}(\text{Toothache}, \text{Catch} \mid \text{Cavity}) \mathbf{P}(\text{Cavity}) \quad (\text{product rule}) \\ &= \mathbf{P}(\text{Toothache} \mid \text{Cavity}) \mathbf{P}(\text{Catch} \mid \text{Cavity}) \mathbf{P}(\text{Cavity}) \end{aligned}$$

- The general form for the full joint distribution can be written as

$$\mathbf{P}(Cause, Effect_1, \dots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i | Cause) .$$

- Such a probability distribution is called a **naive Bayes model**—“naive” because it is often used (as a simplifying assumption) in cases where the “effect” variables are not actually conditionally independent given the cause variable.
- The naive Bayes model is sometimes called a Bayesian classifier.

# Naive Bayes Classifier

→ เราได้ข้อสังเกตเป็น evidence<sub>n</sub> | ข้อสังเกต, ตัวสังเกต  
↓ ข้อความที่เราสนใจเป็น  $y$ ;  $x$  | ข้อความที่เราสนใจ, ข้อความสังเกต

- Naive Bayes is a kind of classifier which uses the Bayes theorem.
- It predicts membership probabilities for each class such as the probability that given data point belongs to a particular class.
- The class with the highest probability is considered as the most likely class, i.e., Maximum A Posteriori (MAP).

- The MAP for a hypothesis is:

$$MAP(H) = \max(P(H|E))$$

$$= \max\left(\frac{P(E|H)P(H)}{P(E)}\right)$$

$$= \max(P(E|H) \cdot P(H))$$

↓ ถ้าเราสนใจหา hypothesis → ให้ max

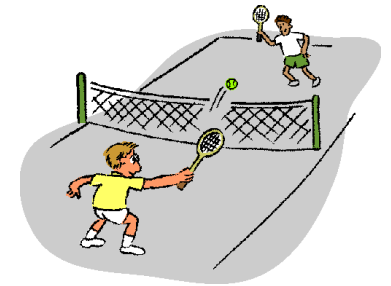
Marginalization constant  
(removing it won't affect.)

- Naive Bayes classifier assumes that all the features are unrelated to each other.



## Simple example of Naive Bayes Classifier (single feature)

- Suppose we have a training data set of weather and corresponding target variable 'Play' (suggesting possibilities of playing tennis).
- Now, we need to classify whether players will play or not based on weather condition.



14 instances

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Frequency table

Convert a raw data into the frequency table.

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Calculate the likelihood table.

- Will players play tennis if weather is sunny?

*2 hypo*

$$\begin{aligned} P(\text{Yes} \mid \text{Sunny}) &= P(\text{Sunny} \mid \text{Yes}) * P(\text{Yes}) / P(\text{Sunny}) \\ &= (3/9) * 0.64 / 0.36 = 0.60 \end{aligned}$$

$$\begin{aligned} P(\text{No} \mid \text{Sunny}) &= P(\text{Sunny} \mid \text{No}) * P(\text{No}) / P(\text{Sunny}) \\ &= (2/5) * 0.36 / 0.36 = 0.40 \end{aligned}$$

MAP(H) = max(0.60, 0.40) = 0.60. Therefore, H = Yes.

## Another example of Naive Bayes Classifier (multiple features)



- We have 3 classes (500 instances each) associated with Animal Types: Parrot, Dog, Fish.
- There are 4 features associate with each class: Swim, Wings, GreenColor, DangerousTeeth.
- All the features are categorical variables with either of the 2 values: T(True) or F(False).

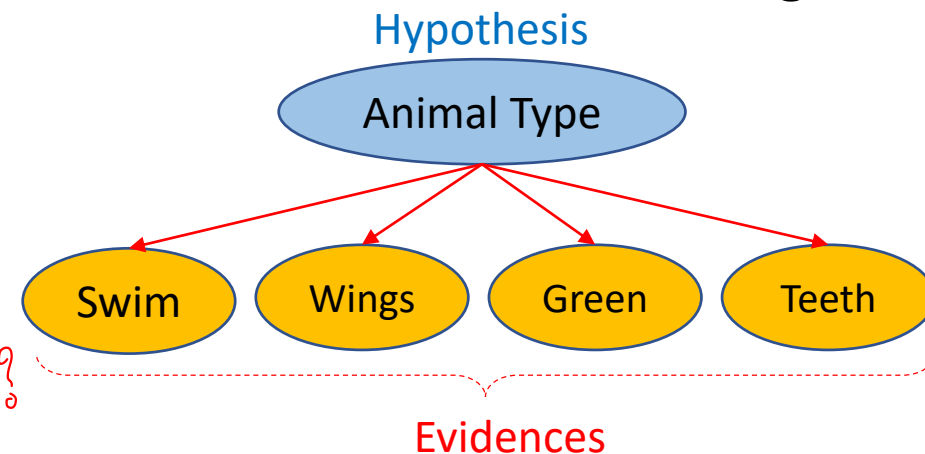
Swim	Wings	Green Color	Dangerous Teeth	Animal Type
50/500	500/500	400/500	0	Parrot
450/500	0	0	500/500	Dog
500/500	0	100/500	50/500	Fish

The above table shows a frequency table of our data. In our training data:

- **Parrots** have 50(10%) values for Swim, i.e., 10% parrot can swim according to our data, 500 out of 500(100%) parrots have wings, 400 out of 500(80%) parrots are Green and 0(0%) parrots have Dangerous Teeth.
- Classes with Animal type **Dogs** shows that 450 out of 500(90%) can swim, 0(0%) dogs have wings, 0(0%) dogs are of Green color and 500 out of 500(100%) dogs have Dangerous Teeth.
- Classes with Animal type **Fishes** shows that 500 out of 500(100%) can swim, 0(0%) fishes have wings, 100(20%) fishes are of Green color and 50 out of 500(10%) dogs have Dangerous Teeth.

- Given a record that has values in their feature set, predict whether the animal is a Dog, a Parrot or a Fish.

Swim	Wings	Green Color	Dangerous Teeth
True	False	True	False



- We will use the Naive Bayes approach

$$P(H | \text{Multiple Evidences}) = P(E1 | H) * P(E2 | H) * \dots * P(En | H) * P(H) / P(\text{Multiple Evidences})$$

- Let's consider a record. The Evidence here is Swim & Green.

~~$$P(\text{Dog} | \text{Swim, Green}) = P(\text{Swim} | \text{Dog}) * P(\text{Green} | \text{Dog}) * P(\text{Dog}) / P(\text{Swim, Green})$$~~

$$= 0.9 * 0 * 0.333 / P(\text{Swim, Green}) = 0 \frac{1}{3}$$

$$P(\text{Parrot} | \text{Swim, Green}) = P(\text{Swim} | \text{Parrot}) * P(\text{Green} | \text{Parrot}) * P(\text{Parrot}) / P(\text{Swim, Green})$$

$$= 0.1 * 0.80 * 0.333 / P(\text{Swim, Green}) = 0.0264 / P(\text{Swim, Green})$$

$$P(\text{Fish} | \text{Swim, Green}) = P(\text{Swim} | \text{Fish}) * P(\text{Green} | \text{Fish}) * P(\text{Fish}) / P(\text{Swim, Green})$$

$$= 1 * 0.2 * 0.333 / P(\text{Swim, Green})$$

$$= 0.0666 / P(\text{Swim, Green})$$

Therefore, the Naïve Bayes classifier predicts that this record is a fish.

## Advantages

เหมาะสำหรับ binary, event ของ binary, event non-zero, event non-zero

- Naive Bayes Algorithm is a fast, highly scalable algorithm.
- Naive Bayes can be use for Binary and Multiclass classification.
- It is a simple algorithm that depends on doing a bunch of counts.
- Great choice for text classification problems. It's a popular choice for spam email classification.

เหมาะสำหรับ spam mail ที่ติด spam 555+

## Disadvantages

ไม่เหมาะกับ continuous value, event ที่ขึ้นต่อกัน

- It considers all the features to be unrelated, so it cannot learn the relationship between features. E.g., Let's say Remo is going to a party. Remo likes to wear a white color shirt. He likes to wear a brown Jeans, But Remo doesn't like wearing a white shirt with Brown Jeans.