Artificial Intelligence

Instructor: Kietikul Jearanaitanakij

Department of Computer Engineering

King Mongkut’s Institute of Technology Ladkrabang

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

- 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

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

Example: P(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).

Example of prior probability:

- P(Weather) = < 0.7, 0.2, 0.08, 0.02 >
- P(Weather) defines a probability distribution for the random variable Weather.

Weather	P(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has obtained some evidence concerning the previously unknown propositions making up the domain, prior probabilities are no longer applicable.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

  $$P(Cavity \mid Toothache) = 0.8$$

- $P(A|B)$ 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

- **Conditional probability** can be calculated by

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

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

$$P(A \wedge B) = P(A|B).P(B)$$

- Note that $P(A \wedge B) \equiv P(A, B)$.

The joint probability distribution

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

	Toothache	**¬Toothache**
Cavity	0.04	0.06
¬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.

$$P(Cavity) = 0.10$$

$$P(Cavity \vee Toothache) = \quad ?$$

$$P(Cavity \mid Toothache) = \quad ?$$

Bayes’ rule and its use

$P(A \wedge B) = P(A|B).P(B)$ ....(1)

$P(A \wedge B) = P(B|A).P(A)$ ....(2)

(1) = (2) , $P(B|A) = \frac{P(A|B).P(B)}{P(A)}$; Bayes’ rule.

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

Proof:

Example of Bayes' rule

In medical diagnosis, the doctor **knows P(symptoms | disease)** and **want to derive a diagnosis, P(disease | symptoms)**. **For example**, a doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 70% of the time.

$$P(s \mid m) \;=\; 0.7$$

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

$$P(m) \;=\; 1/50000$$

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

$$P(s) \;=\; 0.01$$

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

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

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.

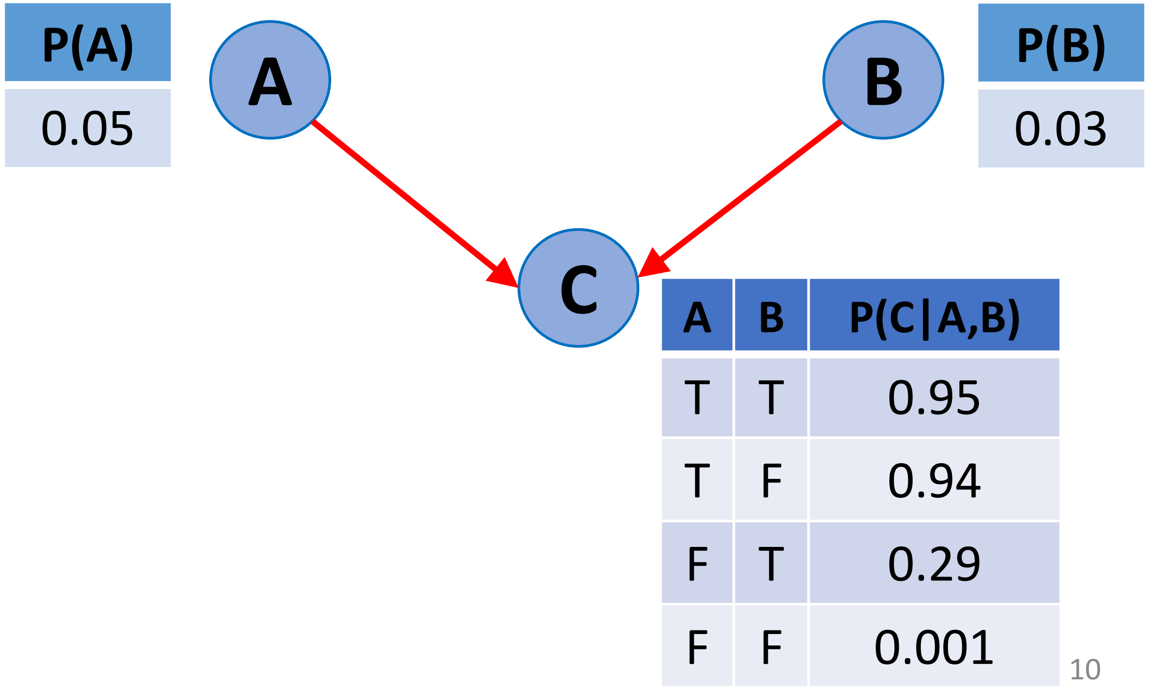
Probabilistic Reasoning Systems

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.

$$P(A\land B\land C\land D) = \underbrace{P(A|B\land C\land D).P(B\land C\land D)}_{\text{Not easy to calculate}}$$

Bayesian network

- 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.



P(A)
0.05

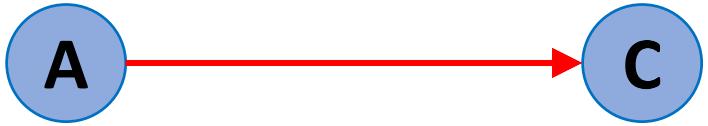
P(B)
0.03

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

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.



A has a direct influence on C.

- Each node Xi has a conditional probability distribution P(Xi |Parents(Xi)) that quantifies the effect of the parents on the node.

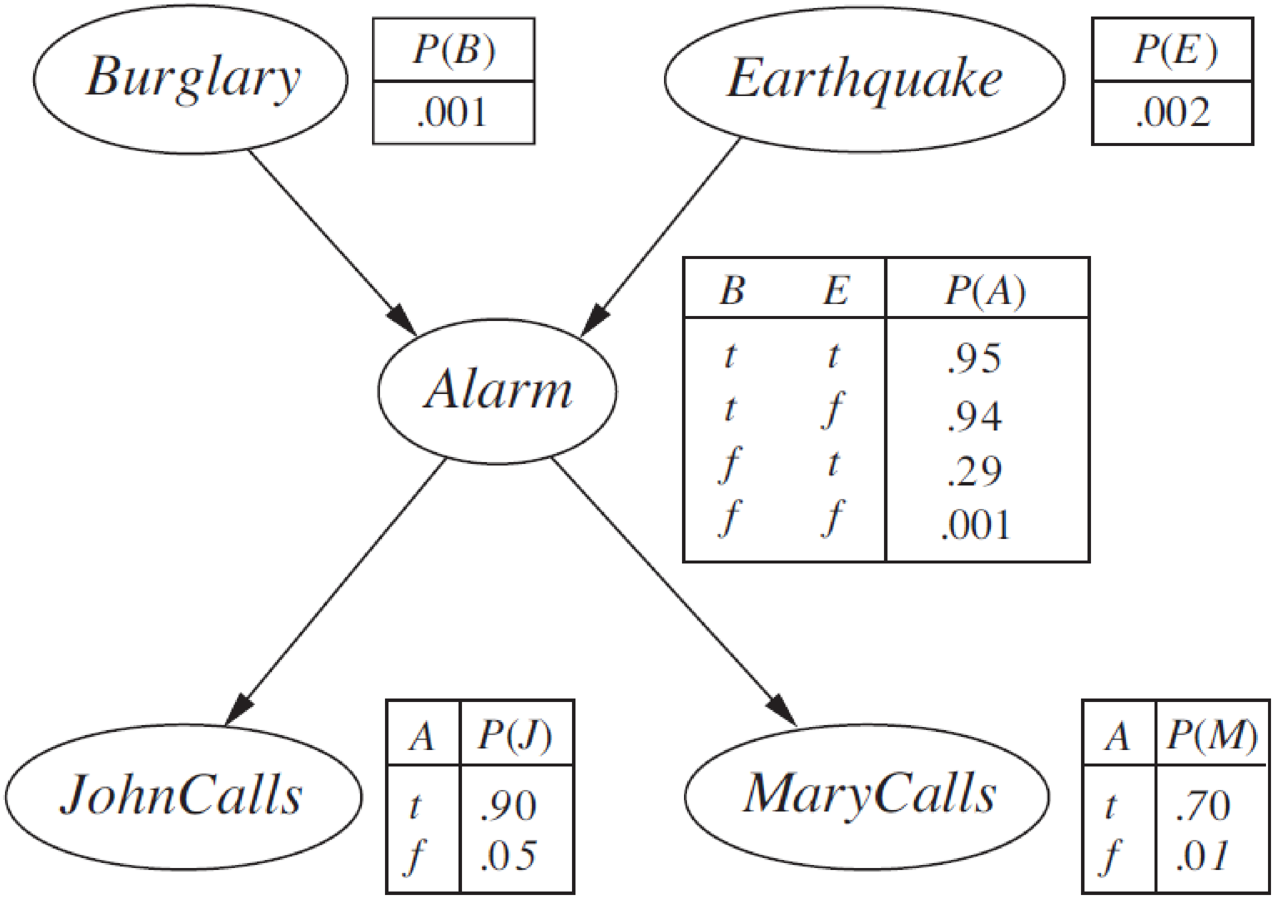
Example : Burglary alarm

John’s house

Mary's house

- 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.

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



P(B)
.001

P(E)
.002

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J)
t	.90
f	.05

A	P(M)
t	.70
f	.01

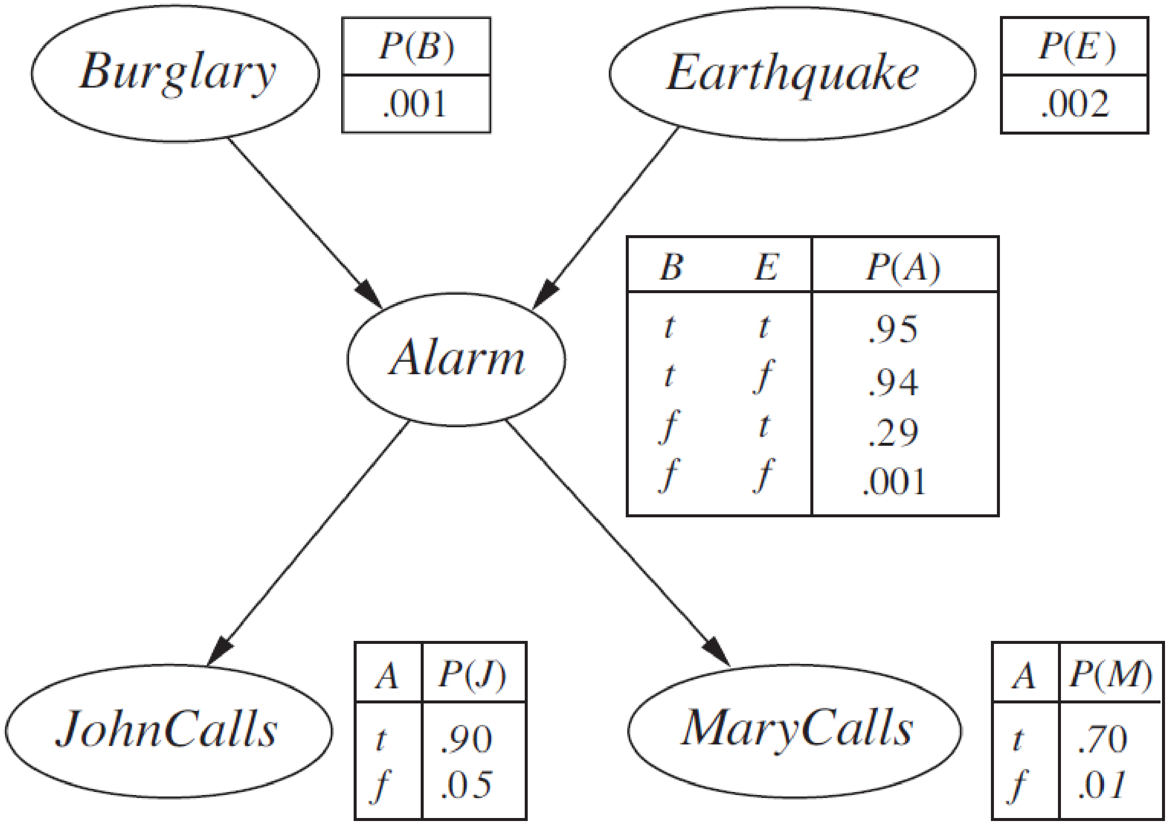
Did you see something unreasonable ?

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.

$$P(j, m, a, \neg b, \neg e) = ?$$

The joint probability distribution can be represented by

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \,|\, parents(X_i)) \,.$$



$P(B)$
.001

$P(E)$
.002

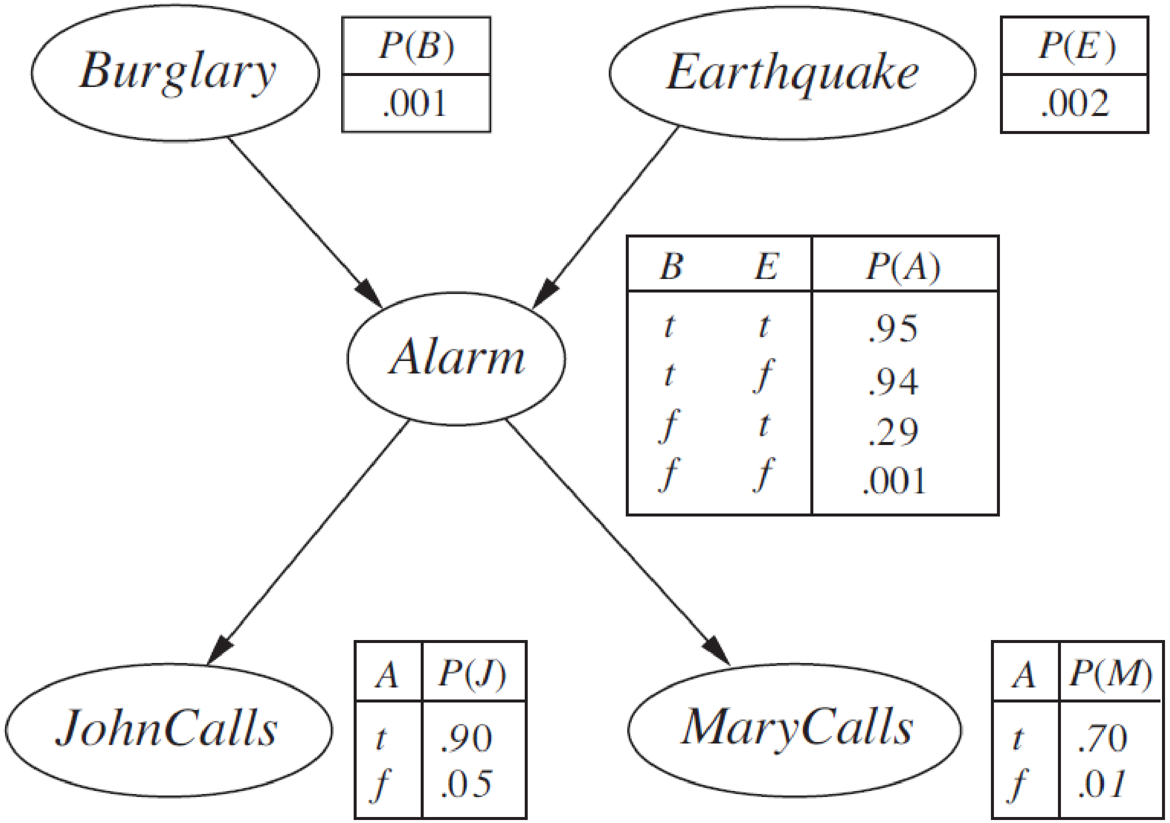
B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

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.



$P(B)$
.001

$P(E)$
.002

B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

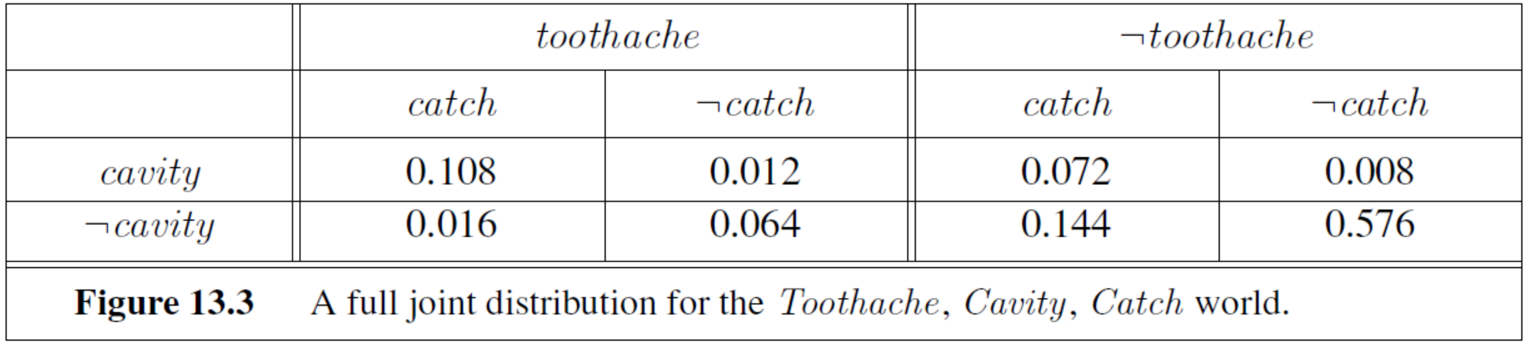
A	$P(M)$
t	.70
f	.01

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

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(Toothache, Cavity, Catch) is shown in 2x2x2 table as follows:



	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	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.
- $P(cavity \lor toothache) =$ 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

  $P(cavity) =$ 0.108+0.012+0.072+0.008 = 0.20 ; **Marginal probability** of cavity (or **summing out**)

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

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).

$$P(Y) = \sum_{z} P(Y|z) P(z)$$

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

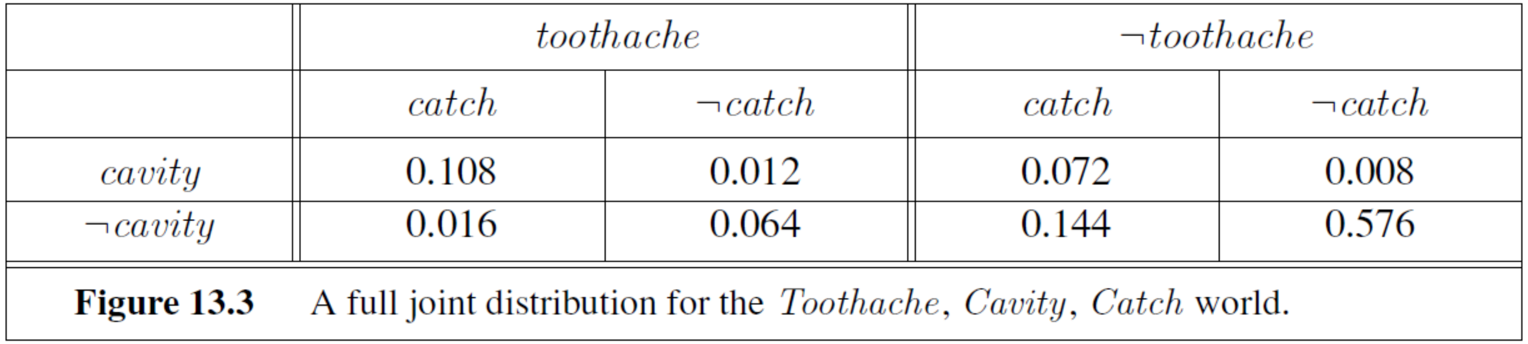
- **Normalization constant:**

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

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

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

- 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 ∝**,** for the distribution P(Cavity | toothache), ensuring that it adds up to 1.



	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

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

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

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

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

Divide each term by (0.12+0.08).

- We can calculate P(Cavity | toothache) even if we don’t know the value of P(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(toothache)) is not available.

- **Bayes’ rule : (Revisited)**

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

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

- 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.**

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

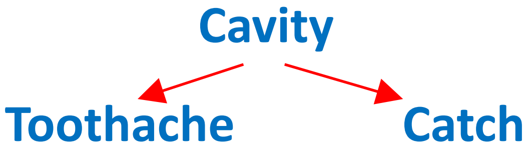
$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014\,.$$

Using independence of variables to simplify 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?

$$\mathbf{P}(Cavity \mid toothache \land catch)$$
$$= \alpha\, \mathbf{P}(toothache \land catch \mid Cavity)\, \mathbf{P}(Cavity)$$

- We need to know the conditional probabilities of the conjunction (toothache ∧ catch) for each value of Cavity. That might be feasible for just two evidence variables, but it does not scale up.
- We can assert the notion of independence between variables to simplify the expression.
- 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).



Cavity

Toothache

Catch

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

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

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

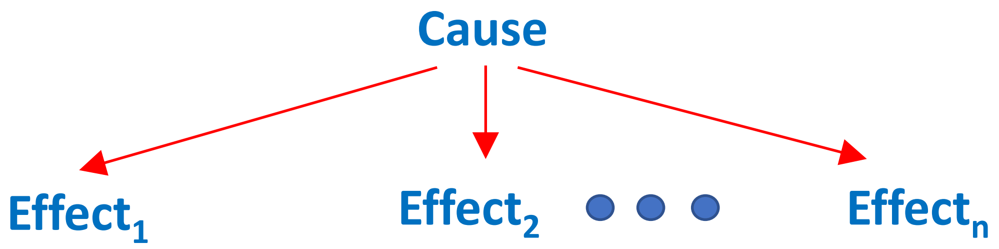
$$\begin{aligned}&\mathbf{P}(Cavity \mid toothache \wedge catch)\\ &\quad = \alpha\, \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity)\, \mathbf{P}(Cavity)\ .\end{aligned}$$

- This is good because the joint probability disappears.

- The conditional independence also decomposes the full joint distribution into pieces.

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

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



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

- 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

- 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:

$$\begin{aligned} MAP(H) &= \max\big(P(H|E)\big) \\ &= \max\left(\frac{P(E|H)P(H)}{P(E)}\right) \\ &= \max(P(E|H).P(H)) \end{aligned}$$

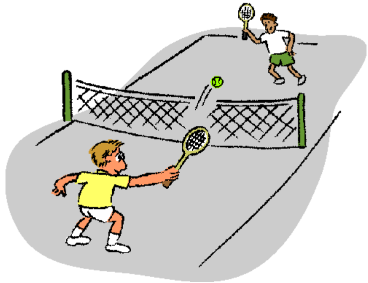
Marginalization constant
(removing it won't affect.)

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

Reference: http://dataaspirant.com

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

Convert a raw data into the frequency table.

Frequency Table

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

Calculate the likelihood 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		

- Will players play tennis if weather is sunny?

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)
= (3/9) * 0.64 / 0.36 = 0.60

P(No | Sunny) = P(Sunny | No) * P(No) / P (Sunny)
= (2/5) * 0.36 / 0.36 = 0.40

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).

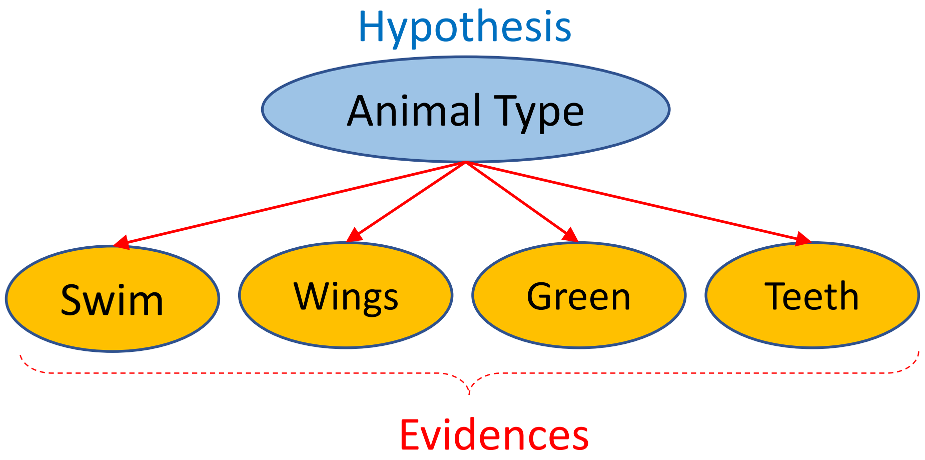
Swim	Wings	Green Color	Dangerous Teeth	Animal Type
50/500	500/500	400/500	0	Parrot
450/500	0	0	500/500	Dog
500/500	0	100/500	50/500	Fish

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|Multiple Evidences) = P(E1| H)* P(E2|H) ……*P(En|H) * P(H) / P(Multiple Evidences)

- Let’s consider a record. The Evidence here is Swim & Green.
- P(Dog | Swim, Green) = P(Swim|Dog) * P(Green|Dog) * P(Dog) / P(Swim, Green)

= 0.9 * 0 * 0.333 / P(Swim, Green) = **0**

P(Parrot| Swim, Green) = P(Swim|Parrot) * P(Green|Parrot) * P(Parrot) / P(Swim, Green)

= 0.1 * 0.80 * 0.333 / P(Swim, Green) = **0.0264**/ P(Swim, Green)

P(Fish | Swim, Green) = P(Swim|Fish) * P(Green|Fish) * P(Fish) / P(Swim, Green)

= 1 * 0.2 * 0.333 / P(Swim, Green)

= **0.0666**/ P(Swim, Green)

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

Advantages

- 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.

Disadvantages

- 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.