Lecture 3
From Uncertainty management to Naïve Bayes

1. **Introduction, or what is uncertainty?**
2. **Basic probability theory**
3. **Bayesian reasoning**
4. **Naïve Bayes Classifiers**
5. **Certainty factors theory and evidential reasoning**

1. Introduction, or what is uncertainty?

- Information can be incomplete, inconsistent, uncertain, or all three. In other words, information is often unsuitable for solving a problem.
- **Uncertainty** is defined as the lack of the exact knowledge that would enable us to reach a perfectly reliable conclusion. Classical logic permits only exact reasoning. It assumes that perfect knowledge always exists and the ***law of the excluded middle*** can always be applied:

IF ***A*** **is true**	**IF** ***A*** **is false**
THEN ***A*** **is not false**	**THEN** ***A*** **is not true**

Sources of *uncertain* knowledge

- **Weak implications.** Domain experts and knowledge engineers have the painful task of establishing concrete correlations between IF (condition) and THEN (action) parts of the rules. Therefore, expert systems need to have the ability to handle vague associations, for example by accepting the degree of correlations as numerical certainty factors.

- **Imprecise language.** Our natural language is ambiguous and imprecise. We describe facts with such terms as *often* and *sometimes*, *frequently* and *hardly ever*. As a result, it can be difficult to express knowledge in the precise IF-THEN form of production rules. However, if the meaning of the facts is quantified, it can be used in expert systems. In 1944, Ray Simpson asked 355 high school and college students to place 20 terms like *often* on a scale between 1 and 100. In 1968, Milton Hakel repeated this experiment.

Quantification of ambiguous and imprecise terms on a time-frequency scale

Ray Simpson (1944)		*Milton Hakel* (1968)	
Term	*Mean value*	*Term*	*Mean value*
Always	99	Always	100
Very often	88	Very often	87
Usually	85	Usually	79
Often	78	Often	74
Generally	78	Rather often	74
Frequently	73	Frequently	72
Rather often	65	Generally	72
About as often as not	50	About as often as not	50
Now and then	20	Now and then	34
Sometimes	20	Sometimes	29
Occasionally	20	Occasionally	28
Once in a while	15	Once in a while	22
Not often	13	Not often	16
Usually not	10	Usually not	16
Seldom	10	Seldom	9
Hardly ever	7	Hardly ever	8
Very seldom	6	Very seldom	7
Rarely	5	Rarely	5
Almost never	3	Almost never	2
Never	0	Never	0

- **Unknown data.** When the data is incomplete or missing, the only solution is to accept the value "unknown" and proceed to an approximate reasoning with this value.
- **Combining the views of different experts.** Large expert systems usually combine the knowledge and expertise of a number of experts. Unfortunately, experts often have contradictory opinions and produce conflicting rules. To resolve the conflict, the knowledge engineer has to attach a weight to each expert and then calculate the composite conclusion. But no systematic method exists to obtain these weights.

2. Basic probability theory

- The concept of probability has a long history that goes back thousands of years when words like "probably", "likely", "maybe", "perhaps" and "possibly" were introduced into spoken languages. However, the mathematical theory of probability was formulated only in the 17th century.
- The **probability** of an event is the proportion of cases in which the event occurs. Probability can also be defined as a ***scientific measure of chance***.

- Probability can be expressed mathematically as a numerical index with a range between zero (an absolute impossibility) to unity (an absolute certainty).
- Most events have a probability index strictly between 0 and 1, which means that each event has *at least* two possible outcomes: favourable outcome or success, and unfavourable outcome or failure.

$$P(success) = \frac{the\ number\ of\ successes}{the\ number\ of\ possible\ outcomes}$$

$$P(failure) = \frac{the\ number\ of\ failures}{the\ number\ of\ possible\ outcomes}$$

- If s is the number of times success can occur, and f is the number of times failure can occur, then

$$\boxed{\begin{aligned} P(success) &= p = \frac{s}{s+f} \\ P(failure) &= q = \frac{f}{s+f} \end{aligned}}$$

and $$\boxed{p + q = 1}$$

- If we throw a coin, the probability of getting a head will be equal to the probability of getting a tail. In a single throw, $s = f = 1$, and therefore the probability of getting a head (or a tail) is 0.5.

Conditional probability

- Let A be an event in the world and B be another event. Suppose that events A and B are not mutually exclusive, but occur conditionally on the occurrence of the other. The probability that event A will occur if event B occurs is called the **conditional probability**. Conditional probability is denoted mathematically as $p(A|B)$ in which the vertical bar represents *GIVEN* and the complete probability expression is interpreted as *"Conditional probability of event A occurring given that event B has occurred"*.

$$\boxed{p(A|B) = \frac{\text{the number of times } A \text{ and } B \text{ can occur}}{\text{the number of times } B \text{ can occur}}}$$

- The number of times *A* and *B* can occur, or the probability that both *A* and *B* will occur, is called the **joint probability** of *A* and *B*. It is represented mathematically as $p(A \cap B)$. The number of ways *B* can occur is the probability of *B*, $p(B)$, and thus

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

- Similarly, the conditional probability of event *B* occurring given that event *A* has occurred equals

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

Hence, $$\boxed{p(B \cap A) = p(B|A) \times p(A)}$$

and $$\boxed{p(A \cap B) = p(A|B) \times p(B)}$$

Substituting the last equation into the equation

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

yields the **Bayesian rule**:

Bayesian rule

$$p\left(A\middle|B\right)=\frac{p\left(B\middle|A\right)\times p\left(A\right)}{p\left(B\right)}$$

where:

$p(A|B)$ is the conditional probability that event A occurs given that event B has occurred;

$p(B|A)$ is the conditional probability of event B occurring given that event A has occurred;

$p(A)$ is the probability of event A occurring;

$p(B)$ is the probability of event B occurring.

- The concept of conditionality probability considered that event A was dependent upon event B.
- This principle can be extended to event A being dependent on a number of mutually exclusive events B_1, B_2,…, B_n. We know:

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

$$p(A \cap B_1) = p(A|B_1) \times p(B_1)$$

$$p(A \cap B_2) = p(A|B_2) \times p(B_2)$$

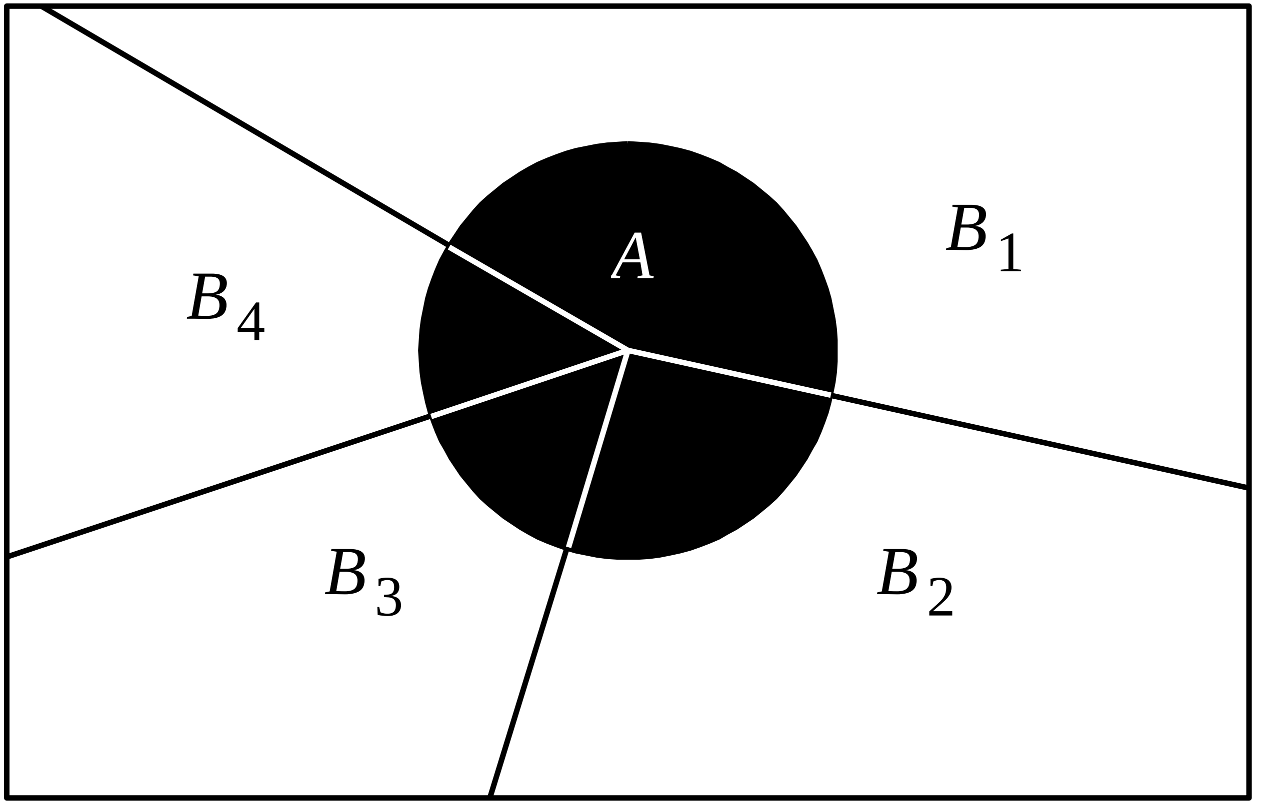
.

$$p(A \cap B_n) = p(A|B_n) \times p(B_n)$$

$$\sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$

The joint probability

$$\sum_{i=1}^{n} p\left(A \cap B_i\right) = \sum_{i=1}^{n} p\left(A \middle| B_i\right) \times p\left(B_i\right)$$



For all events B_i

$$\sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$

$$\sum_{i=1}^{n} p(A \cap B_i) = p(A)$$

Therefore

$$p(A) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$

If the occurrence of event A depends on only two mutually exclusive events, B and NOT B, we obtain:

$$\boxed{p(A) = p(A|B) \times p(B) + p(A|\neg B) \times p(\neg B)}$$

where $\neg$ is the logical function NOT.

Similarly,

$$\boxed{p(B) = p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)}$$

Substituting this equation into the ***Bayesian rule*** yields:

$$\boxed{p(A|B) = \frac{p(B|A) \times p(A)}{p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)}}$$

3. Bayesian reasoning

Suppose all rules in the knowledge base are represented in the following form:

IF	*E* is true
THEN	*H* is true {with probability *p*}

This rule implies that if event *E* occurs, then the probability that event *H* will occur is *p*.

In expert systems, *H* usually represents a **hypothesis** and *E* denotes **evidence** to support this hypothesis.

The Bayesian rule expressed in terms of hypotheses and evidence looks like this:

$$p(H|E) = \frac{p(E|H)\times p(H)}{p(E|H)\times p(H) + p(E|\neg H)\times p(\neg H)}$$

where:

$p(H)$ is the prior probability of hypothesis H being true;

$p(E|H)$ is the probability that hypothesis H being true will result in evidence E;

$p(\neg H)$ is the prior probability of hypothesis H being false;

$p(E|\neg H)$ is the probability of finding evidence E even when hypothesis H is false.

- In expert systems, the probabilities required to solve a problem are provided by experts. An expert determines the **prior probabilities** for possible hypotheses $p(H)$ and $p(\neg H)$, and also the **conditional probabilities** for observing evidence E if hypothesis H is true, $p(E|H)$, and if hypothesis H is false, $p(E|\neg H)$.
- Users provide information about the evidence observed and the expert system computes $p(H|E)$ for hypothesis H in light of the user-supplied evidence E. Probability $p(H|E)$ is called the **posterior probability** of hypothesis H upon observing evidence E.

- We can take into account both multiple hypotheses H_1, H_2,..., H_m and multiple evidences E_1, E_2,..., E_n. The hypotheses as well as the evidences must be mutually exclusive and exhaustive.
- Single evidence E and multiple hypotheses follow:

$$\boxed{p(H_i|E) = \frac{p(E|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E|H_k) \times p(H_k)}}$$

- Multiple evidences and multiple hypotheses follow:

$$\boxed{p(H_i|E_1\,E_2\,\ldots\,E_n) = \frac{p(E_1\,E_2\,\ldots\,E_n|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E_1\,E_2\,\ldots\,E_n|H_k) \times p(H_k)}}$$

- This requires to obtain the conditional probabilities of all possible combinations of evidences for all hypotheses, and thus places an enormous burden on the expert.
- Therefore, in expert systems, conditional independence among different evidences assumed. Thus, instead of the unworkable *equation*, we attain:

$$p\left(H_i \middle| E_1\, E_2 \ldots E_n\right) = \frac{p\left(E_1 \middle| H_i\right) \times p\left(E_2 \middle| H_i\right) \times \ldots \times p\left(E_n \middle| H_i\right) \times p\left(H_i\right)}{\sum_{k=1}^{m} p\left(E_1 \middle| H_k\right) \times p\left(E_2 \middle| H_k\right) \times \ldots \times p\left(E_n \middle| H_k\right) \times p\left(H_k\right)}$$

Ranking potentially true hypotheses

Let us consider a simple example.

Suppose an expert, given three conditionally independent evidences E_1, E_2 and E_3, creates three mutually exclusive and exhaustive hypotheses H_1, H_2 and H_3, and provides prior probabilities for these hypotheses – $p(H_1)$, $p(H_2)$ and $p(H_3)$, respectively. The expert also determines the conditional probabilities of observing each evidence for all possible hypotheses.

The prior and conditional probabilities

Probability	*Hypothesis*		
	$i = 1$	$i = 2$	$i = 3$
$p(H_i)$	0.40	0.35	0.25
$p(E_1 \mid H_i)$	0.3	0.8	0.5
$p(E_2 \mid H_i)$	0.9	0.0	0.7
$p(E_3 \mid H_i)$	0.6	0.7	0.9

Assume that we first observe evidence E_3. The expert system computes the posterior probabilities for all hypotheses as

$$p\left(H_i \middle| E_3\right) = \frac{p\left(E_3 \middle| H_i\right) \times p\left(H_i\right)}{\sum_{k=1}^{3} p\left(E_3 \middle| H_k\right) \times p\left(H_k\right)}, \qquad i = 1, 2, 3$$

Thus,

$$p\left(H_1 \middle| E_3\right) = \frac{0.6 \cdot 0.40}{0.6 \cdot 0.40 + 0.7 \cdot 0.35 + 0.9 \cdot 0.25} = 0.34$$

$$p\left(H_2 \middle| E_3\right) = \frac{0.7 \cdot 0.35}{0.6 \cdot 0.40 + 0.7 \cdot 0.35 + 0.9 \cdot 0.25} = 0.34$$

$$p\left(H_3 \middle| E_3\right) = \frac{0.9 \cdot 0.25}{0.6 \cdot 0.40 + 0.7 \cdot 0.35 + 0.9 \cdot 0.25} = 0.32$$

After evidence E_3 is observed, belief in hypothesis H_1 decreases and becomes equal to belief in hypothesis H_2. Belief in hypothesis H_3 also increases and even nearly reaches beliefs in hypotheses H_1 and H_2.

Suppose now that we observe evidence E_1. The posterior probabilities are calculated as

$$p\left(H_i \middle| E_1 E_3\right) = \frac{p\left(E_1 \middle| H_i\right) \times p\left(E_3 \middle| H_i\right) \times p\left(H_i\right)}{\sum_{k=1}^{3} p\left(E_1 \middle| H_k\right) \times p\left(E_3 \middle| H_k\right) \times p\left(H_k\right)}, \qquad i = 1, 2, 3$$

Hence,

$$p\left(H_1 \middle| E_1 E_3\right) = \frac{0.3 \cdot 0.6 \cdot 0.40}{0.3 \cdot 0.6 \cdot 0.40 + 0.8 \cdot 0.7 \cdot 0.35 + 0.5 \cdot 0.9 \cdot 0.25} = 0.19$$

$$p\left(H_2 \middle| E_1 E_3\right) = \frac{0.8 \cdot 0.7 \cdot 0.35}{0.3 \cdot 0.6 \cdot 0.40 + 0.8 \cdot 0.7 \cdot 0.35 + 0.5 \cdot 0.9 \cdot 0.25} = 0.52$$

$$p\left(H_3 \middle| E_1 E_3\right) = \frac{0.5 \cdot 0.9 \cdot 0.25}{0.3 \cdot 0.6 \cdot 0.40 + 0.8 \cdot 0.7 \cdot 0.35 + 0.5 \cdot 0.9 \cdot 0.25} = 0.29$$

Hypothesis H_2 has now become the most likely one.

After observing evidence E_2, the final posterior probabilities for all hypotheses are calculated:

$$p\left(H_i \middle| E_1E_2E_3\right)=\frac{p\left(E_1 \middle| H_i\right)\times p\left(E_2 \middle| H_i\right)\times p\left(E_3 \middle| H_i\right)\times p\left(H_i\right)}{\sum_{k=1}^{3} p\left(E_1 \middle| H_k\right)\times p\left(E_2 \middle| H_k\right)\times p\left(E_3 \middle| H_k\right)\times p\left(H_k\right)}, \qquad i=1,2,3$$

$$p\left(H_1 \middle| E_1E_2E_3\right)=\frac{0.3\cdot 0.9\cdot 0.6\cdot 0.40}{0.3\cdot 0.9\cdot 0.6\cdot 0.40+0.8\cdot 0.0\cdot 0.7\cdot 0.35+0.5\cdot 0.7\cdot 0.9\cdot 0.25}=0.45$$

$$p\left(H_2 \middle| E_1E_2E_3\right)=\frac{0.8\cdot 0.0\cdot 0.7\cdot 0.35}{0.3\cdot 0.9\cdot 0.6\cdot 0.40+0.8\cdot 0.0\cdot 0.7\cdot 0.35+0.5\cdot 0.7\cdot 0.9\cdot 0.25}=0$$

$$p\left(H_3 \middle| E_1E_2E_3\right)=\frac{0.5\cdot 0.7\cdot 0.9\cdot 0.25}{0.3\cdot 0.9\cdot 0.6\cdot 0.40+0.8\cdot 0.0\cdot 0.7\cdot 0.35+0.5\cdot 0.7\cdot 0.9\cdot 0.25}=0.55$$

Although the initial ranking was H_1, H_2 and H_3, only hypotheses H_1 and H_3 remain under consideration after all evidences (E_1, E_2 and E_3) were observed.

4. Naïve Bayes Classifiers

- Bayesian Inference calculate the posterior distribution based on priories

$$p(H_i|E_1\,E_2\,\ldots E_n)=\frac{p(E_1|H_i)\times p(E_2|H_i)\times\ldots\times p(E_n|H_i)\times p(H_i)}{\sum_{k=1}^{m}p(E_1|H_k)\times p(E_2|H_k)\times\ldots\times p(E_n|H_k)\times p(H_k)}$$

- But if we don't know the priories, and we only have some historical data? (Learning from the data)
- We can use naïve Bayes classifier to select the hypothesis or classes.

Naïve Bayes Classifiers

- Bayes Rules

$$\boxed{p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}}$$

where:

$p(A|B)$ is the conditional probability that event A occurs given that event B has occurred;

$p(B|A)$ is the conditional probability of event B occurring given that event A has occurred;

$p(A)$ is the probability of event A occurring;

$p(B)$ is the probability of event B occurring.

- Bayes Rules from previous slide can be represented based on hypothesis and evidence as :

$$\boxed{p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}}$$

Maximum A Posteriori (MAP)

- For *a set of all events E*, we compute the maximum a posteriori hypothesis:

$$\boxed{\begin{aligned} h_{MAP} &= \underset{h\in H}{\arg\max}\, P(h \mid E) \\ h_{MAP} &= \underset{h\in H}{\arg\max}\, \frac{P(E \mid h)P(h)}{P(E)} \\ h_{MAP} &= \underset{h\in H}{\arg\max}\, P(E \mid h)P(h) \end{aligned}}$$

- We can omit P(E) since it is constant and independent of the hypothesis
- This the only difference between MAP and Bayesian Inference

Naïve Bayes Classifiers

- Suppose that a set of training examples includes records with conjunctive attributes values *(a_1, a_2, .. a_n)* and target function is based on finite set of classes *V*.

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(v_j \mid a_1, a_2, .., a_n)$$

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, \frac{P(a_1, a_2, .., a_n \mid v_j) P(v_j)}{P(a_1, a_2, .., a_n)}$$

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(a_1, a_2, .., a_n \mid v_j) P(v_j)$$

Estimation

$$\boxed{v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(a_1, a_2, .., a_n \mid v_j) P(v_j)}$$

- We can easily estimate $P(v_j)$ by computing relative frequency of each target class in the training set.
- But, estimating $P(a_1,a_2,..a_n|v_j)$ is difficult.

- So, <u>we assume that attributes values are conditionally independent</u> given the target value:

$$P(a_1, a_2, .., a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

What we know

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, .., a_n \mid v_j) P(v_j)$$

Therefore

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

- Estimating $P(a_i|v_j)$ instead of $P(a_1,a_2,..,a_n|v_j)$ greatly reduces the number of parameters.
- The learning step in Naïve Bayes consists of estimating $P(a_i|v_j)$ and $P(v_j)$ based on the frequencies in the training data. (The priories are embedded in the training data)
- There is no explicit search during training.
- An unseen instance is classified by computing the class that maximizes the posterior.

Example

Day	*Outlook*	*Temp*	*Humidity*	*Wind*	*Play Tennis*
Dl	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**

Outlook			Temperature			Humidity			Wind			Play	
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
Sunny	2	3	**Hot**	2	2	**High**	3	4	**Weak**	6	2	9	5
Overcast	4	0	**Mild**	4	2	**Normal**	6	1	**Strong**	3	3		
Rainy	3	2	**Cool**	3	1								
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
Sunny	2/9	3/5	**Hot**	2/9	2/5	**High**	3/9	4/5	**Weak**	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	**Mild**	4/9	2/5	**Normal**	6/9	1/5	**Strong**	3/9	3/5		
Rainy	3/9	2/5	**Cool**	3/9	1/5								

- Suppose that we want classify following new instance:
- Outlook = sunny, temp = cool, humidity = high, wind = strong.

$$v_{NB} = \underset{v_j \in \{yes, no\}}{\arg\max}\ P(v_j) \prod_i P(a_i \mid v_j)$$

$$v_{NB} = \underset{v_j \in \{yes, no\}}{\arg\max}\ P(v_j) P(outlook = sunny \mid v_j) P(temp = cool \mid v_j) P(humidity = high \mid v_j) P(wind = strong \mid v_j)$$

- First, we calculate the prior probabilities:

$$P(play = yes) = 9/14$$
$$P(play = no) = 5/14$$

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes)$$
$$= 9/14 \times 2/9 \times 3/9 \times 3/9 \times 3/9 = 0.0053$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no)$$
$$= 5/14 \times 3/5 \times 1/5 \times 4/5 \times 3/5 = 0.0206$$
$$v_{NB} = \underset{v_j \in \{yes,no\}}{\arg\max} P(v_j)P(sunny \mid v_j)P(cool \mid v_j)$$
$$P(high \mid v_j)P(strong \mid v_j) = no$$

5. Certainty factors theory and evidential reasoning

- Certainty factors theory is a popular alternative to Bayesian reasoning.
- A **certainty factor** (*cf*), a number to measure the expert's belief. The maximum value of the certainty factor is, say, +1.0 (definitely true) and the minimum –1.0 (definitely false). For example, if the expert states that some evidence is almost certainly true, a *cf* value of 0.8 would be assigned to this evidence.

Uncertain terms and their interpretation in MYCIN

Term	*Certainty Factor*
Definitely not	−1.0
Almost certainly not	−0.8
Probably not	−0.6
Maybe not	−0.4
Unknown	−0.2 to +0.2
Maybe	+0.4
Probably	+0.6
Almost certainly	+0.8
Definitely	+1.0

- In expert systems with certainty factors, the knowledge base consists of a set of rules that have the following syntax:

IF <evidence>
THEN <hypothesis> {*cf* }

- where *cf* represents belief in hypothesis *H* given that evidence *E* has occurred.

- The certainty factors theory is based on two functions: measure of belief $MB(H,E)$, and measure of disbelief $MD(H,E)$.

$$MB(H,E) = \begin{cases} 1 & \text{if } p(H) = 1 \\ \dfrac{max\left[p(H|E),\, p(H)\right] - p(H)}{max\left[1,\, 0\right] - p(H)} & \text{otherwise} \end{cases}$$

$$MD(H,E) = \begin{cases} 1 & \text{if } p(H) = 0 \\ \dfrac{min\left[p(H|E),\, p(H)\right] - p(H)}{min\left[1,\, 0\right] - p(H)} & \text{otherwise} \end{cases}$$

$p(H)$ is the prior probability of hypothesis H being true;
$p(H|E)$ is the probability that hypothesis H is true given evidence E.

- The values of $MB(H, E)$ and $MD(H, E)$ range between 0 and 1. The strength of belief or disbelief in hypothesis H depends on the kind of evidence E observed. Some facts may increase the strength of belief, but some increase the strength of disbelief.
- **The total strength of belief or disbelief in a hypothesis:**

$$\boxed{cf = \frac{MB(H,E) - MD(H,E)}{1 - min[MB(H,E), MD(H,E)]}}$$

Example:

Consider a simple rule:

IF $\quad A$ is X

THEN B is Y

An expert may not be absolutely certain that this rule holds. Also suppose it has been observed that in some cases, even when the IF part of the rule is satisfied and *object* A takes on *value* X, object B can acquire some different value Z.

IF $\quad A$ is X

THEN B is Y $\{cf\ 0.7\}$;

B is Z $\{cf\ 0.2\}$

The certainty factor assigned by a rule is propagated through the reasoning chain. This involves establishing the net certainty of the rule consequent when the evidence in the rule antecedent is uncertain:

$$\boldsymbol{cf\,(H,E) = cf\,(E) \times cf}$$

For example,

IF sky is clear

THEN the forecast is sunny {*cf* 0.8}

and the current certainty factor of *sky is clear* is 0.5, then

$$cf\,(H,E) = 0.5 \cdot 0.8 = 0.4$$

This result can be interpreted as *"It may be sunny"*.

- For conjunctive rules such as

IF	<evidence E_1>
	⋮
AND	<evidence E_n>
THEN	<hypothesis H> {*cf* }

the certainty of hypothesis H, is established as follows:

$$cf\,(H,E_1 \cap E_2 \cap \ldots \cap E_n) = min\,[cf\,(E_1), cf\,(E_2),\ldots, cf\,(E_n)] \times cf$$

For example,

IF sky is clear
AND the forecast is sunny
THEN the action is 'wear sunglasses' {*cf* 0.8}

and the certainty of *sky is clear* is 0.9 and the certainty of the *forecast of sunny* is 0.7, then

$$cf\,(H,E_1 \cap E_2) = min\,[0.9, 0.7] \cdot 0.8 = 0.7 \cdot 0.8 = 0.56$$

- For disjunctive rules such as

IF	<evidence E_1>
	⋮
OR	<evidence E_n>
THEN	<hypothesis H> {cf }

the certainty of hypothesis H, is established as follows:

$$cf\,(H, E_1 \cup E_2 \cup \ldots \cup E_n) = max\,[cf\,(E_1), cf\,(E_2), \ldots, cf\,(E_n)] \times cf$$

For example,

IF sky is overcast
OR the forecast is rain
THEN the action is 'take an umbrella' {cf 0.9}

and the certainty of *sky is overcast* is 0.6 and the certainty of the *forecast of rain* is 0.8, then

$$cf\,(H, E_1 \cup E_2) = max\,[0.6, 0.8] \cdot 0.9 = 0.8 \cdot 0.9 = 0.72$$

- When the same consequent is obtained as a result of the execution of two or more rules, the individual certainty factors of these rules must be merged to give a combined certainty factor for a hypothesis.

Suppose the knowledge base consists of the following rules:

Rule 1: IF A is X
THEN C is Z {*cf* 0.8}

Rule 2: IF B is Y
THEN C is Z {*cf* 0.6}

What certainty should be assigned to object C having value Z if both *Rule* 1 and *Rule* 2 are fired?

Common sense suggests that, if we have two pieces of evidence (*A* is *X* and *B* is *Y*) from different sources (*Rule* 1 and *Rule* 2) supporting the same hypothesis (*C* is *Z*), then the confidence in this hypothesis should increase and become stronger than if only one piece of evidence had been obtained.

To calculate a combined certainty factor we can use the following equation:

$$cf(cf_1, cf_2) = \begin{cases} cf_1 + cf_2 \times (1 - cf_1) & \text{if } cf_1 > 0 \text{ and } cf_2 > 0 \\[2ex] \dfrac{cf_1 + cf_2}{1 - min\left[|cf_1|, |cf_2|\right]} & \text{if } cf_1 < 0 \text{ or } cf_2 < 0 \\[2ex] cf_1 + cf_2 \times (1 + cf_1) & \text{if } cf_1 < 0 \text{ and } cf_2 < 0 \end{cases}$$

where:

cf_1 is the confidence in hypothesis *H* established by *Rule* 1;

cf_2 is the confidence in hypothesis *H* established by *Rule* 2;

$|cf_1|$ and $|cf_2|$ are absolute magnitudes of cf_1 and cf_2, respectively.

The certainty factors theory provides a ***practical*** alternative to Bayesian reasoning. The heuristic manner of combining certainty factors is different from the manner in which they would be combined if they were probabilities. The certainty theory is not "mathematically pure" but does mimic the thinking process of a human expert.

Comparison of Bayesian reasoning and certainty factors

- Probability theory is the oldest and best-established technique to deal with inexact knowledge and random data. It works well in such areas as forecasting and planning, where statistical data is usually available and accurate probability statements can be made.

- However, in many areas of possible applications of expert systems, reliable statistical information is not available or we cannot assume the conditional independence of evidence. As a result, many researchers have found the Bayesian method unsuitable for their work. This dissatisfaction motivated the development of the certainty factors theory.
- Although the certainty factors approach lacks the mathematical correctness of the probability theory, it outperforms subjective Bayesian reasoning in such areas as diagnostics.

- Certainty factors are used in cases where the probabilities are not known or are too difficult or expensive to obtain. The evidential reasoning mechanism can manage incrementally acquired evidence, the conjunction and disjunction of hypotheses, as well as evidences with different degrees of belief.
- The certainty factors approach also provides better explanations of the control flow through a rule-based expert system.

- The Bayesian method is likely to be the most appropriate if reliable statistical data exists, the knowledge engineer is able to lead, and the expert is available for serious decision-analytical conversations.
- In the absence of any of the specified conditions, the Bayesian approach might be too arbitrary and even biased to produce meaningful results.
- The Bayesian belief propagation is of exponential complexity, and thus is impractical for large knowledge bases.

Want to know more?

- David E. Heckerman , Edward H. Shortliffe (1992) From certainty factors to belief networks
- http://research.microsoft.com/en-us/um/people/heckerman/HS91aim.pdf
- Rish, Irina. (2001). "An empirical study of the naive Bayes classifier".
- https://faculty.cc.gatech.edu/~isbell/reading/papers/Rish.pdf
- Andrew Gelman, et al. (2003) Bayesian Data Analysis