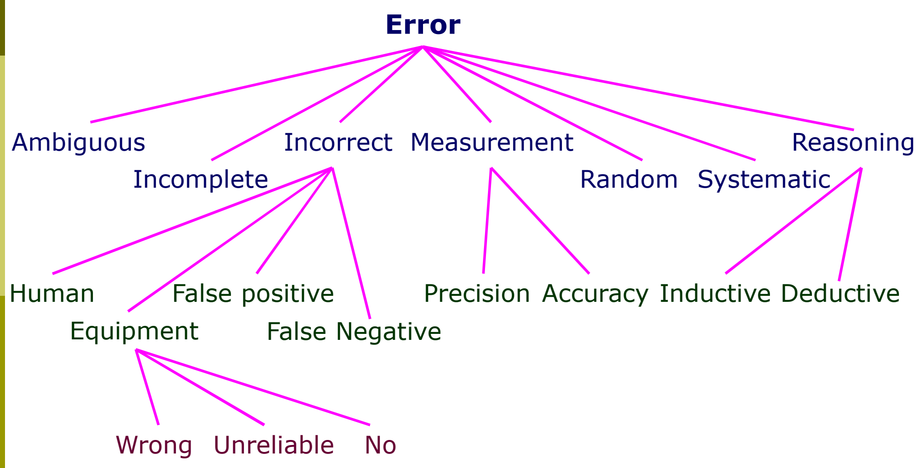
智能系统原理与开发
第05章 不确定性推理

李斌
复旦大学 计算机科学技术学院

Uncertainty

- Uncertainty can be considered as the ***lack of adequate*** information to make a decision.
- Uncertainty is a problem because it may prevent us from making the ***best*** decision and may even cause a ***bad*** decision to be made.
- All living creatures are ***experts*** at dealing with uncertainty or they could not survive in the real world.

Types of error



Examples of common type of error

Example	Error	Reason
Turn the valve off	Ambiguous	What valve? 哪个
Turn valve-1	Incomplete	Which way? 关还是开
Turn valve-1 off	Incorrect	Correct is on 应该开
Valve is stuck	False positive (Type I)	Valve is not stuck
Valve is not stuck	False negative (Type II)	Valve is stuck
Turn valve-1 to 5	Imprecise	Correct is 5.4 欠精度
Turn valve-1 to 5.4	Inaccurate	Correct is 9.2 不正确
Turn valve-1 to 5.4 or 6 or 0	Unreliable	Equipment error
Valve-1 setting is 5.4 or 5.5 or 5.1	Random error	Statistical fluctuation
Valve-1 is not stuck because it's never been stuck before	Invalid induction 无效归纳	Valve is stuck
Output is normal and so valve-1 is in good condition	Invalid deduction 无效演绎	Valve is stuck in open position

Errors and induction

All men are mortal
Socrates is man

Socrates is mortal

Deduction
Always right

My disk has never crashed

My disk will never crash

Induction
degree of confidence

演绎推理（Deduction）是从知识推导出事实的过程；
归纳推理（Induction）是从事实推导出知识的过程。

Inductive argument

Rules 1

The fire alarm goes off

There is a fire

Rules 2

The fire alarm goes off
I smell smoke

There is a fire

Rules 3

The fire alarm goes off
I smell smoke
My clothes are burning

There is a fire

Fallacy（谬论）

$p \rightarrow q$	**If** *the valve is in good condition,* **then** *the output is normal*
q	*The output is normal*
p	*The valve is in good condition*

The valve may be stuck in the open position. If it is necessary to close the valve, the problem will show up.

Three Statistical Experiments

- A lady, who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured into the cup first. In all of ten trials conducted to test this, she correctly determines which was poured first.
- A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In ten trials conducted to test this, he makes a correct determination each time.
- A drunken friend says he can predict the outcome of a flip of a fair coin. In ten trials conducted to test this, he is correct each time.

Three Statistical Experiments

- In all three situations, the unknown quantity θ is the probability of the person answering correctly.
- A classical significance test of the various claims consider the null hypotheses (H_0) that $\theta = 0.5$
- In all three situations this hypothesis would be rejected with a (one-tailed) significance level 2^{-10}.

Problems of hypothesis test

- **A point null hypothesis**
  The point null hypothesis is almost certainly not exactly true, and that this will always be confirmed by *a large enough sample*.
- **Tests of fit**
  It is virtually certain that the model is not exactly correct, so *a large enough sample* will almost always reject the model.

$$X_1, X_2, \ldots, X_n \sim N(\theta, 1) \qquad \frac{\bar{X} - \theta_0}{1/\sqrt{n}}$$

Conclusion

- A theory should be proposed to solve the above problems.
- A link should be established between epistemic reasoning (reasoning about what to believe) and practical reasoning (reasoning about what to do).
- Individual preferences should be considered properly when dealing with the above issues.

Bayes' theorem

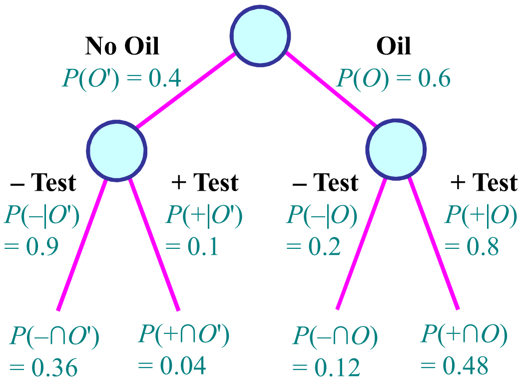
The method of Bayesian decision making is used in the PROSPECTOR experts system to decide favorable sites for mineral exploration.

$P(O) = 0.6 \quad P(O') = 0.4$

$P(+|O) = 0.8 \quad P(-|O) = 0.2$

$P(+|O') = 0.1 \quad P(-|O') = 0.9$

Bayesian decision making



Probabilities

Prior

Subjective Opinion of site

$P(H_i)$

Conditional

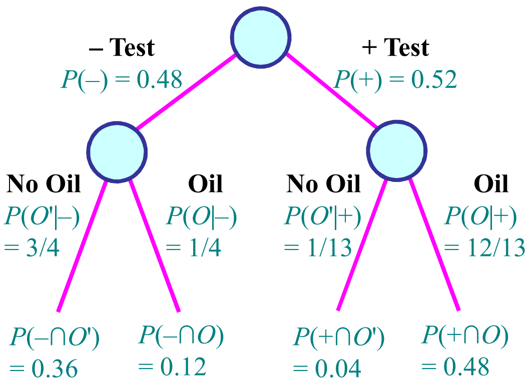
Seismic Test Result

$P(E|H_i)$

Joint

$P(E\cap H_i) = P(E|H_i)P(H_i)$

Bayesian decision making



Probabilities

Unconditional

$P(E)$

Posterior

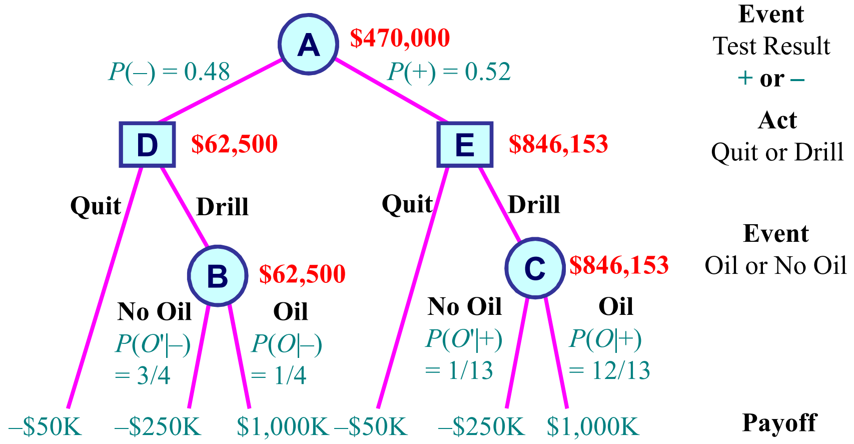
$P(H_i|E)$

$= P(E|H_i)\cdot P(H_i)/P(E)$

Joint

$P(E\cap H) = P(H_i|E)P(E)$

Bayesian decision making



Oil lease, if successful	$1,250,000
Drilling expense	−$200,000
Seismic survey	−$50,000

Bayesian analysis

A farmer try to decide which crop to grow next year.

a_1: drought-resistant crop（抗旱作物）

a_2: high-yielding crop（高产作物）

θ : the precipitation（降水）of next year

Loss function（损失函数）

$$l(\theta,a)=\begin{cases}200-2\theta, & \text{if } a=a_1\\ 3000-10\theta, & \text{if } a=a_2\end{cases}$$

Bayesian analysis

A farmer collects the information about ***precipitation*** by listening to ***weather forecast***. If next year's precipitation is less than 400mm, he will grow ***drought-resistant crop***, otherwise, he will choose to grow ***high-yielding crop***.

Is it wise to make the decision like this?

Bayesian analysis

According to history record, the accuracy of weather forecast has the Cauchy distribution **C**(400, θ) .

$$f(x\mid\theta)=\begin{cases}\dfrac{1}{\pi}\dfrac{400}{400^2+(x-\theta)^2}, & \text{if } x\ge 0\\ 0, & \text{if } x<0\end{cases}$$

Bayesian analysis

The risk of decision is:

$$R(\theta,\delta)=\int_0^{400}\frac{1}{\pi}\frac{400}{400^2+(x-\theta)^2}(200-2\theta)dx$$
$$+\int_{400}^{\infty}\frac{1}{\pi}\frac{400}{400^2+(x-\theta)^2}(3000-10\theta)dx$$

$$=\frac{200-2\theta}{\pi}[arctg(\frac{400-\theta}{400})-arctg(\frac{-\theta}{400})]$$
$$+\frac{3000-10\theta}{\pi}[\frac{\pi}{2}-arctg(\frac{400-\theta}{400})]$$

Bayesian analysis

According to the farmer's experience, the next year's precipitation has a prior density as follows.

θ	0	100	200	300	400	500	600	700	800
$\pi(\theta)$	0	0.052	0.104	0.153	0.178	0.204	0.153	0.104	0.052

The Bayes risk of decision is:

$$r(\pi,\delta)=\sum R(\theta,\delta)\pi(\theta)=-1176$$

The farmer is smart!

Bayesian analysis

- 在进行推断时，除了样本信息之外，还有两类相关的信息特别重要：对决策后果的认识和先验信息
  - 对样本进行分析后采取下一步行为或决策带来可能后果的认识，这种认识常需要量化为每一个可能的行为和决策所造成的损失。
  - 先验信息来自于过去的经验。
- 将先验信息纳入统计学中去并探索如何利用这种信息的方法被称为贝叶斯分析，而把损失函数加到统计分析中形成了决策论，两者结合在一起就是贝叶斯决策论

Difficulties with the Bayesian method

Bayes' Theorem is used to determine the probability of a specific disease, given certain symptoms as

$$P(D_i \mid E) = \frac{P(E \mid D_i)P(D_i)}{P(E)} = \frac{P(E \mid D_i)P(D_i)}{\sum_j P(E \mid D_j)P(D_j)}$$

D_i the ith disease,

E the evidence,

$P(E)$ the prior probability of the patient having the disease before any evidence is known

$P(E|D_i)$ the conditional probability that the patient will exhibit E, given that disease D_i is present.

Difficulties with the Bayesian method

A convenient form of Bayes' Theorem that expresses the accumulation of incremental evidence is:

$$P(D_i \mid E) = \frac{P(E_2 \mid D_i \wedge E_1) P(D_i \mid E_1)}{\sum_j P(E_2 \mid D_j \wedge E_1) P(D_j \mid E_1)}$$

where E_2 is the new evidence added to the existing body of evidence, E_1, to yield the new evidence.

The situation grows worse as more pieces of evidence accumulate and thus more probabilities are required.

Although this formula is exact, all these probabilities are not generally known

MYCIN rule

IF The stain of the organism is gram positive, and
the morphology of the organism is coccus, and
the growth conformation of the organism is chains

THEN There is suggestive evidence (0.7) that the identity of the organism is streptococcus（链球菌）.

This can be written in terms of posterior probability as

$P(H \mid E_1 \wedge E_2 \wedge E_3) = 0.7$

The experts would agree to above equation, they became uneasy and refused to agree with the probabilistic result

$P(H' \mid E_1 \wedge E_2 \wedge E_3) = 0.3$

Belief and disbelief

Assume your grad point average (GPA) has not been too good and you need an A in this course to bring up your GPA.

$$P(\text{ graduating} \mid \text{A in this course }) = 0.70$$

It somehow seems intuitively wrong by

$$P(\text{ not graduating} \mid \text{A in this course }) = 0.30$$

Is $P(H \mid E) = 1 - P(H' \mid E)$ reasonable?

Belief and disbelief

There could be problems due to a number of reasons that would still prevent your graduation, such as

- School catalog changes so that not all your courses counted toward the degree.
- You forgot to take a required course
- Rejection of transfer courses.
- Rejection of some elective courses you took.
- Tuition and library fines that you owe and were hoping would be forgotten weren't
- Your GPA was lower than you thought and an A still won't raise it up.
- "They" are out to get you.

Certainty factors

In MYCIN, the degree of confirmation was originally defined as the certainty factor the difference between belief and disbelief

$$CF(H, E) = MB(H, E) - MD(H, E)$$

CF the certainty factor in the hypothesis H due to evidence E
MB the measure of increased belief in H due to E
MD the measure of increased disbelief in H due to E

MB and MD

The measures of belief and disbelief were defined in terms of probabilities by

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

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

Characteristics of MB, MD and CF

Characteristics	Values
Ranges	$0 \le MB \le 1$ $0 \le MD \le 1$ $-1 \le CF \le 1$
Certain True Hypothesis $P(H\|E) = 1$	$MB = 1$ $MD = 0$ $CF = 1$
Certain False Hypothesis $P(H'\|E) = 0$	$MB = 0$ $MD = 1$ $CF = -1$
Lack of Evidence $P(H\|E) = P(H)$	$MB = 0$ $MD = 0$ $CF = 0$

Combine evidence

Evidence	Antecedent Certainty
E_1 AND E_2	$\min[CF(E_1, e), CF(E_2, e)]$
E_1 OR E_2	$\max[CF(E_1, e), CF(E_2, e)]$
NOT E	$-CF(E, e)$

For example, given a logical expression for combining evidence such as

$$E = (E_1 \text{ AND } E_2 \text{ AND } E_3) \text{ OR } (E_4 \text{ AND NOT } E_5)$$

$$= \max[\min(E_1, E_2, E_3), \min(E_4, -E_5)]$$

CF for the streptococcus rule

IF The stain of the organism is gram positive, and
The morphology of the organism is coccus, and
The growth conformation of the organism is chains

THEN There is suggestive evidence (0.7) that the identity of the organism is streptococcus.

where the certainty factor of the hypothesis under certain evidence is

$$CF(H,E) = CF(H, E_1 \cap E_2 \cap E_3) = 0.7$$

CF for the streptococcus rule

Assuming

$$CF(E_1, e) = 0.5,\ CF(E_2, e) = 0.6,\ CF(E_3, e) = 0.3$$

then

$$\begin{aligned} CF(E, e) &= CF(E_1 \cap E_2 \cap E_3, e) \\ &= \min[CF(E_1, e), CF(E_2, e), CF(E_3, e)] \\ &= \min[0.5, 0.6, 0.3] \end{aligned}$$

The certainty factor of the conclusion is

$$\begin{aligned} CF(H, e) &= CF(E, e)\ CF(H, E) \\ &= 0.3 \times 0.7 \\ &= 0.21 \end{aligned}$$

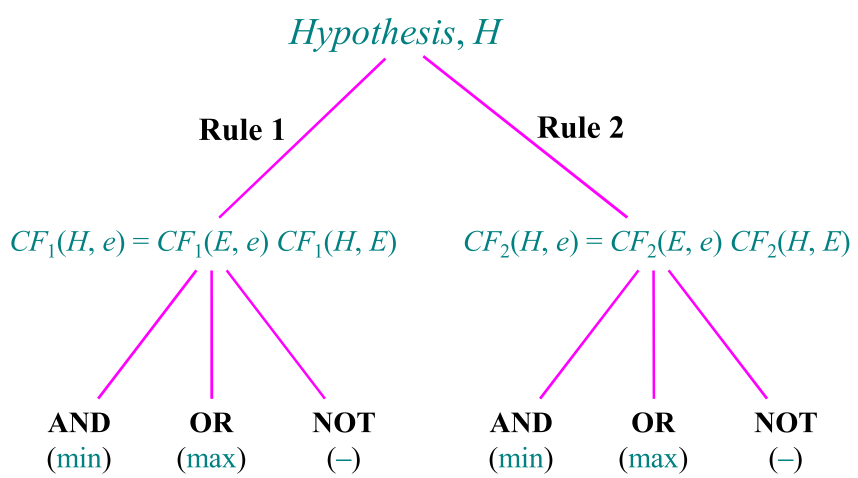
Combining function

The certainty factors of rules concluding the same hypothesis are calculated from the combining function for certainty factors defined as

$$CF_{COMB}(CF_1, CF_2) = \begin{cases} CF_1 + CF_2(1 - CF_1) & \text{both} > 0 \\ \dfrac{CF_1 + CF_2}{1 - \min(|CF_1|, |CF_2|)} & \text{one} < 0 \\ CF_1 + CF_2(1 + CF_1) & \text{both} < 0 \end{cases}$$

Note: $CF_{COMB}(X, Y) = CF_{COMB}(Y, X)$

Computing CF of two rules



Difficulties with certainty factors

One problem was that the CF values could be the opposite of conditional probabilities.

$P(H_1) = 0.8$ $\quad$ $P(H_2) = 0.2$

$P(H_1|E) = 0.9$ $\quad$ $P(H_2|E) = 0.8$

Then

$CF(H_1, E) = 0.5, CF(H_2, E) = 0.75$

Since one purpose of CF is to rank hypotheses in terms of likely diagnosis, it is a contradiction for a disease to have a higher conditional probability $P(H|E)$ and yet have a lower certainty factor, $CF(H, E)$

Difficulties with certainty factors

Second major problem with CF is that in general

$P(H|\mathrm{e}) \neq P(H|i)\ P(i|\mathrm{e})$

where i is some intermediate hypothesis based on evidence e.

Certainty factor of two rules in an inference chain is calculated as independent probabilities by

$CF(H, e) = CF(E, e)\ CF(H, E)$

Temporal reasoning

- Reasoning about events that depend on time is called ***temporal reasoning***.
- Expert systems that reason about temporal events such as aircraft traffic control could be very useful.
- Expert systems that reason over time have been developed in medicine.

Stochastic process

Transition matrix

$$\text{Present}\quad \begin{array}{c c} & \text{Future} \\ & \begin{array}{cc} S_1 & S_2 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \end{array} & \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \end{array} \qquad \text{Present}\quad \begin{array}{c c} & \text{Future} \\ & \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{array}{c} C_1 \\ C_2 \end{array} & \begin{bmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \end{bmatrix} \end{array}$$

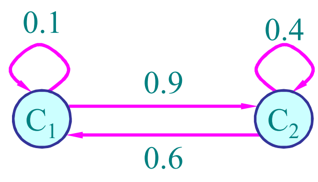
Where P_{mn} is the probability of a transition from state m to n.

C_1: buy **Lenovo** computer.
C_2: buy **Apple** computer.

State diagram interpretation

Future

$$\text{Present}\quad \begin{array}{c|cc} & C_1 & C_2 \\ \hline C_1 & 0.1 & 0.9 \\ C_2 & 0.6 & 0.4 \end{array}$$



State diagram

C_1: buy **Lenovo** computer.
C_2: buy **Apple** computer.

State matrix

$S = [\, P_1, P_2, \ldots P_n \,]$

Initially, with 80 percent of the people owning lenovo, with 20 percent owing apple

$S_1 = [\, 0.8, 0.2 \,]$

$S_2 = S_1 \cdot$ Transition matrix

$$= [\, 0.8, 0.2 \,] \begin{bmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \end{bmatrix}$$

$$= [\, 0.2, 0.8]$$

Equilibrium

S_3 = [0.5, 0.5]

S_4 = [0.35, 0.65]

S_5 = [0.425, 0.575]

S_6 = [0.3875, 0.6125]

S_7 = [0.40625, 0.59375]

S_8 = [0.396875, 0.602125]

Notice that the states are ***converging*** on

S_n = [0.4, 0.6]

A transition matrix P is regular if some power of P has only positive entries, then a unique steady-state S_n exists.

Markov chain process

- A ***finite number*** of possible states.
- The process can be in ***one and only one*** state at a time.
- The process moves or ***steps*** successively from one state to another over time.
- The probability of a move depends only on the ***immediately preceding state***.

Solve steady-state matrix

$$[\,P_1 \;\; P_2\,] \begin{bmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \end{bmatrix} = [\,P_1 \;\; P_2\,]$$

We have

$$\left.\begin{aligned} 0.1P_1 + 0.6P_2 &= P_1 \\ 0.9P_1 + 0.4P_2 &= P_2 \end{aligned}\right\} \Longrightarrow P_1 = 2/3 \cdot P_2$$

$$\left.\begin{aligned} P_1 &= 2/3 \cdot P_2 \\ P_1 + P_2 &= 1 \end{aligned}\right\} \Longrightarrow$$

$$P_1 = 0.4,\ P_2 = 0.6$$

Fuzzy set

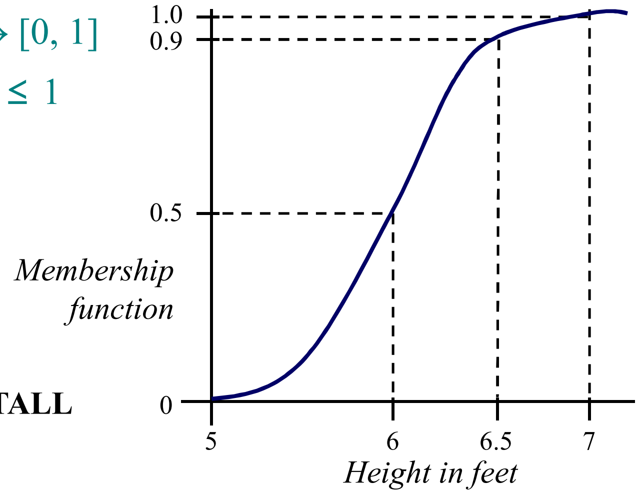
- Fuzzy set is primarily concerned with quantifying and reasoning using ***natural language*** in which many words have ambiguous meanings such as ***tall***, ***hot***, ***dangerous***, ***a little***, and so on.
- Classical concept that an object is either in a set or not in a set dates form the Aristotelian view of ***bivalent*** or ***two-valued logic***.
- The problem with this bivalent logic is that we live in an ***analog***, not a digital world.
- The development of analog theories of computation such as ***artificial neural systems*** and ***fuzzy theory*** more accurately represents the real world.

Membership function

Membership or compatibility function

$\mu_A(x): x \rightarrow [0, 1]$

$0 \leq \mu_A(x) \leq 1$



Fuzzy set **TALL**

Linguistic variable and typical value

Linguistic variable	Typical Values
height	dwarf, short, average, tall, giant
number	almost, none, several, few, many
stage of life	infant, toddler, child, teenager, adult
color	red, blue, green, yellow, orange
light	dim, faint, normal, bright, intense
dessert	pie, cake, ice cream, baked alaska

Fuzzy rules

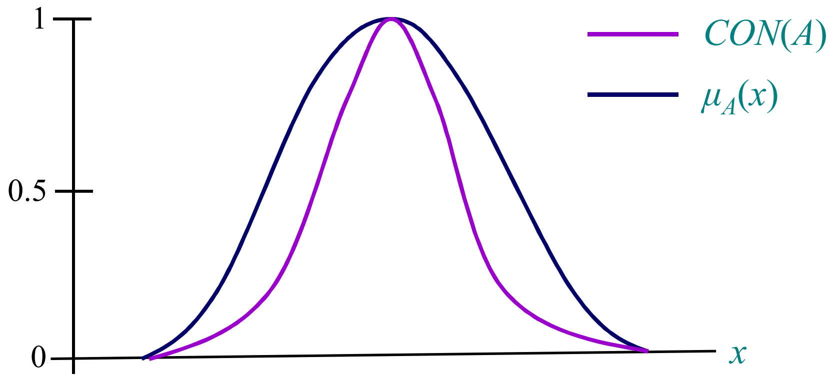
IF the TV is too dim **THEN** turn up the brightness
IF it is too hot **THEN** add some cold
IF the pressure is too high **THEN** open the relief value
IF interest rates are going up **THEN** buy bonds
IF interest rates are going down **THEN** buy stocks

In a fuzzy rule, the premise *x is A* and the consequent *y is B* can be true to a degree, instead of entirely true or entirely false

Premise: *x is A**
Implication: ***IF*** *x is A* ***THEN*** *y is B*
Consequent: *y is B**

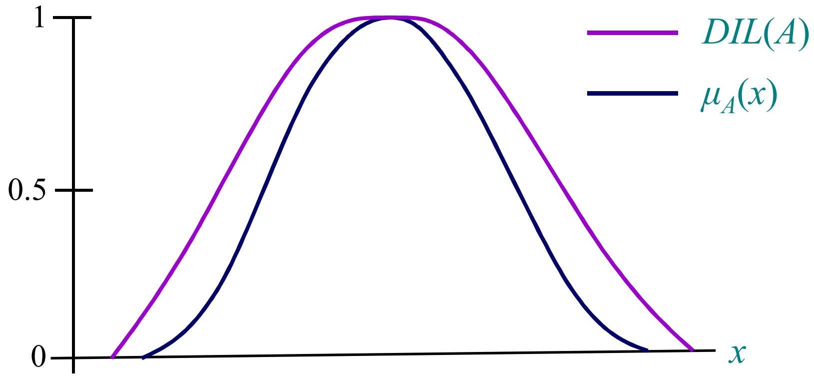
Concentration operation

Very $CON(A) = \mu_A(x)^2$



Dilation operation

More or Less $CON(A) = \mu_A(x)^{0.5}$



Linguistic hedges and operation

Hedge	Operator definition
Very F	$CON(F) = F^2$
More or Less F	$DIL(F) = F^{0.5}$
Plus F	$F^{1.25}$
Not F	$1 - F$
Not Very F	$1 - CON(F)$

TALL = { 0.125/5.5, 0.5/6, 0.875/6.5, 1/7 }

VERY TALL = { 0.0156/5.5, 0.25/6, 0.7656/6.5, 1/7 }

NOT TALL = { 0.875/5.5, 0.5/6, 0.125/6.5, 0/7 }

Fuzzy relation

		y			
x	120	130	140	150	160
120	1.0	0.7	0.4	0.2	0.0
130	0.7	1.0	0.6	0.5	0.2
140	0.4	0.6	1.0	0.8	0.5
150	0.2	0.5	0.8	1.0	0.8
160	0.0	0.2	0.5	0.8	1.0

$R(x, y)$ = **APPROXIMATELY EQUAL**
on the binary relation of people's weights.

Composition operator

$R_1(x)$ = **HEAVY** = { 0.6/140, 0.8/150, 1/160 }

$R_2(x, y)$ = **APPROXIMATELY EQUAL**

$R_3(y)$ = **MORE OR LESS HEAVY**

$= R_1(x) \circ R_2(x, y)$ max $\min_x (R_1(x), R_2(x, y))$

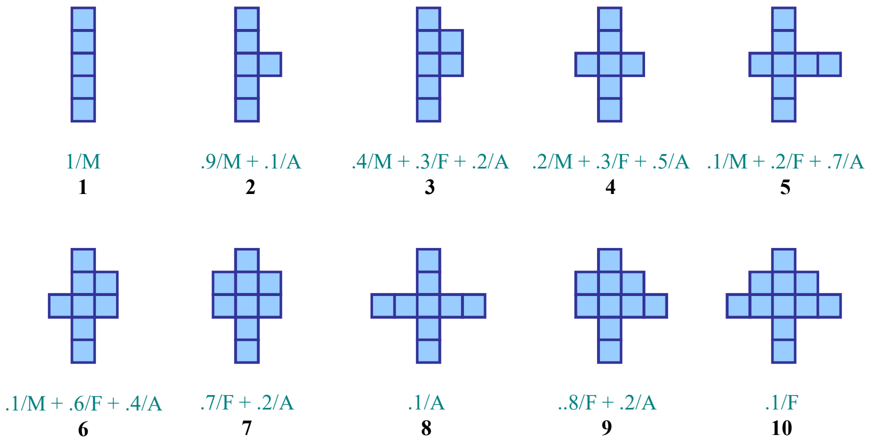
$$= [\, 0.0 \;\; 0.0 \;\; 0.6 \;\; 0.8 \;\; 1.0 \,] \circ \begin{pmatrix} 1.0 & 0.7 & 0.4 & 0.2 & 0.0 \\ 0.7 & 1.0 & 0.6 & 0.5 & 0.2 \\ 0.4 & 0.6 & 1.0 & 0.8 & 0.5 \\ 0.2 & 0.5 & 0.8 & 1.0 & 0.8 \\ 0.0 & 0.2 & 0.5 & 0.8 & 1.0 \end{pmatrix}$$

= { 0.4/120, 0.6/130, 0.8/140, 0.8/150, 1/160 }

Membership grades for images

	Membership Grade		
Image	**Missile**	**Fighter**	**Airliner**
1	1.0	0.0	0.0
2	0.9	0.0	0.1
3	0.4	0.3	0.2
4	0.2	0.3	0.2
5	0.1	0.2	0.7
6	0.1	0.6	0.4
7	0.0	0.7	0.2
8	0.0	0.0	1.0
9	0.0	0.8	0.2
10	0.0	1.0	0.0

Fuzzy sets for Aircraft identification



Fuzzy rules

IF IMAGE4 **THEN** TARGET4 (.2/M + .3/F + .5/A)

IF IMAGE6 **THEN** TARGET6 (.1/M + .6/F + .4/A)

TARGET = TARGET4 + TARGET6
= .2/M + .3/F + .5/A + .1/M + .6/F + .4/A
= .2/M + .6/F + .5/A

where only the ***maximum membership grades*** for each element are retained in the TARGET fuzzy set.

Fuzzy inference

In general, given n observations and rules

IF E_1 **THEN** H_1
IF E_2 **THEN** H_2
⋮
IF E_n **THEN** H_n

$$\mu_H = \max(\mu_{H_1}, \mu_{H_2}, \ldots, \mu_{H_n})$$
$$= \max[\ \min(\mu_{E_1}), \min(\mu_{E_2}), \ldots, \min(\mu_{E_n})\ ]$$

where each E_i may be some fuzzy expression.

For example, $E_1 = E_A$ AND (E_B OR NOT E_C)
Then, $\mu_{E_1} = \min(\mu_{E_A}, \max(\mu_{E_B}, 1 - \mu_{E_C})$

Thanks