Reasoning System - Symbolic , Statistical : AI Course Lecture 23 – 28, notes, slides

www.myreaders.info/ , RC Chakraborty, e-mail rcchak@gmail.com , June 01, 2010

www.myreaders.info/html/artificial_intelligence.html

Reasoning System

Symbolic , Statistical

Artificial Intelligence

Reasoning system, topics : Definitions, formal and informal logic, uncertainty, monotonic logic, non-monotonic Logic; Methods of reasoning and examples - deductive, inductive, abductive, analogy; Sources of uncertainty; Reasoning and knowledge representation; Approaches to reasoning - symbolic, statistical and fuzzy. Symbolic reasoning : Non-monotonic reasoning - default reasoning, circumscription, truth maintenance systems; Implementation issues. Statistical Reasoning : Glossary of terms; Probability and Bayes' theorem; Certainty factors in rule based systems; Bayesian networks, Dempster Shafer theory - model, belief and plausibility, calculus, and combining beliefs; Fuzzy logic - description, membership function.

Reasoning System

Symbolic , Statistical

Artificial Intelligence

Topics

(Lectures 23, 24, 25, 26, 27, 28, 6 hours)

Reasoning System

Symbolic , Statistical

What is Reasoning ?

- Reasoning is the act of deriving a conclusion from certain premises using a given methodology.
- Reasoning is a process of thinking; reasoning is logically arguing; reasoning is drawing inference.
- When a system is required to do something, that it has not been explicitly told how to do, it must reason. It must figure out what it needs to know from what it already knows.
- Many types of Reasoning have long been identified and recognized, but many questions regarding their logical and computational properties still remain controversial.
- The popular methods of Reasoning include abduction, induction, model-based, explanation and confirmation. All of them are intimately related to problems of belief revision and theory development, knowledge assimilation, discovery and learning.

1. Reasoning

Any knowledge system to do something, if it has not been explicitly told how to do it then it must reason.

The system must figure out what it needs to know from what it already knows.

Example

If we know : Robins are birds. All birds have wings.

Then if we ask : Do robins have wings?

Some reasoning (although very simple) has to go on answering the question.

1.1 Definitions :

- **Reasoning** is the act of deriving a conclusion from certain premises using a given methodology.
  - Any knowledge system must reason, if it is required to do something which has not been told explicitly .
  - For reasoning, the system must find out what it needs to know from what it already knows.
  - **Example :**

    If we know : *Robins are birds*.
    All birds have wings

    Then if we ask: *Do robins have wings?*

  To answer this question - some reasoning must go.

- Human reasoning capabilities are divided into three areas:
  - ‡ **Mathematical Reasoning –** axioms, definitions, theorems, proofs
  - ‡ **Logical Reasoning –** deductive, inductive, abductive
  - ‡ **Non-Logical Reasoning –** linguistic , language

  These three areas of reasoning, are in every human being, but the ability level depends on education, environment and genetics.

  The **IQ** (Intelligence quotient) is the summation of mathematical reasoning skill and the logical reasoning.

  The **EQ** (Emotional Quotient) depends mostly on non-logical reasoning capabilities.

  Note : The Logical Reasoning is of our concern in AI

- **Logical Reasoning**

Logic is a language for reasoning. It is a collection of rules called Logic arguments, we use when doing logical reasoning.

Logic reasoning is the process of drawing conclusions from premises using rules of inference.
The study of logic is divided into formal and informal logic.
The formal logic is sometimes called symbolic logic.

Symbolic logic is the study of symbolic abstractions (construct) that capture the formal features of logical inference by a formal *system*.

Formal system consists of two components, a formal language plus a set of inference rules. The formal system has axioms.

Axiom is a sentence that is always true within the system.

Sentences are derived using the system's axioms and rules of derivation are called theorems.

■ Formal Logic

The Formal logic is the study of inference with purely formal content, ie. where content is made explicit.

Examples - Propositional logic and Predicate logic.

- ‡ Here the logical arguments are a set of rules for manipulating symbols. The rules are of two types
  - ◇ Syntax rules : say how to build meaningful expressions.
  - ◇ Inference rules : say how to obtain true formulas from other true formulas.
- ‡ Logic also needs semantics, which says how to assign meaning to expressions.

- **Informal Logic**

  The Informal logic is the study of natural language arguments.

  ‡ The analysis of the argument structures in ordinary language is part of informal logic.

  ‡ The focus lies in distinguishing good arguments (valid) from bad arguments or fallacies (invalid).

■ Formal Systems

Formal systems can have following three properties :

- ‡ **Consistency :** System's theorems do not contradict.
- ‡ **Soundness :** System's rules of derivation will never infer anything false, so long as start is with only true premises.
- ‡ **Completeness :** There are no true sentences in the system that cannot be proved using the derivation rules of the system.

System Elements

Formal systems consist of following elements :

- ‡ A finite set of symbols for constructing formulae.
- ‡ A grammar, is a way of constructing well-formed formulae (wff).
- ‡ A set of axioms; each axiom has to be a wff.
- ‡ A set of inference rules.
- ‡ A set of theorems.

A well-formed formulae, **wff**, is any string generated by a grammar. e.g., the sequence of symbols $((\alpha \rightarrow \beta) \rightarrow (\neg \beta \rightarrow \neg \alpha))$ is a WFF because it is grammatically correct in propositional logic.

■ Formal Language

A formal language may be viewed as being analogous to a collection of words or a collection of sentences.

- In computer science, a formal language is defined by precise mathematical or machine process able formulas.
- A formal language **L** is characterized as a set **F** of finite-length sequences of elements drawn from a specified finite set **A** of symbols.
- Mathematically, it is an unordered pair **L = { A, F }**
- If **A** is words
  then the set **A** is called alphabet of **L**, and
  the elements of **F** are called words.
- If **A** is sentence
  then the set **A** is called the lexicon or vocabulary of **F**, and
  the elements of **F** are then called sentences.
- The mathematical theory that treats formal languages in general is known as **formal language theory**.

- **Uncertainty in Reasoning**

  - The world is an uncertain place; often the Knowledge is imperfect which causes uncertainty. Therefore reasoning must be able to operate under uncertainty.

  - AI systems must have ability to reason under conditions of uncertainty.

Uncertainties		**Desired action**
‡ Incompleteness Knowledge	:	Compensate for lack of knowledge
‡ Inconsistencies Knowledge	:	Resolve ambiguities and contradictions
‡ Changing Knowledge	:	Update the knowledge base over time

- **Monotonic Logic**

Formal logic is a set of rules for making deductions that seem self evident. A Mathematical logic formalizes such deductions with rules precise enough to program a computer to decide if an argument is valid, representing objects and relationships symbolically.

Examples

- Predicate logic and the inferences we perform on it.
- All humans are mortal. Socrates is a human.
  Therefore Socrates is mortal.

In monotonic reasoning if we enlarge at set of axioms we cannot retract any existing assertions or axioms.

- ‡ Most formal logics have a monotonic consequence relation, meaning that adding a formula to a theory never produces a reduction of its set of consequences. In other words, a logic is monotonic if the truth of a proposition does not change when new information (axioms) are added. The traditional logic is monotonic.

- ‡ In mid 1970s, Marvin Minsky and John McCarthy pointed out that pure classical logic is not adequate to represent the commonsense nature of human reasoning. The reason is, the human reasoning is non-monotonic in nature. This means, we reach to conclusions from certain premises that we would not reach if certain other sentences are included in our premises.

- ‡ The non-monotonic human reasoning is caused by the fact that our knowledge about the world is always incomplete and therefore we are forced to reason in the absence of complete information. Therefore we often revise our conclusions, when new information becomes available.

- ‡ Thus, the need for non-monotonic reasoning in AI was recognized, and several formalizations of non-monotonic reasoning.

Only the non-monotonic logic reasoning is presented in next few slides.

- ## Non-Monotonic Logic

Inadequacy of monotonic logic for reasoning is said in the previous slide.

A monotonic logic cannot handle :

- Reasoning by default : because consequences may be derived only because of lack of evidence of the contrary.
- Abductive reasoning : because consequences are only deduced as most likely explanations.
- Belief revision : because new knowledge may contradict old beliefs.

A **non-monotonic** logic is a formal logic whose consequence relation is not monotonic. A logic is non-monotonic if the truth of a proposition may change when new information (axioms) are added.

‡ Allows a statement to be retracted.

‡ Used to formalize plausible (believable) reasoning.

Example 1 :

Birds typically fly.
Tweety is a bird.

Tweety (presumably) flies.

‡ Conclusion of non-monotonic argument may not be correct.

Example-2 : (Ref. Example-1)

If Tweety is a penguin, it is incorrect to conclude that Tweety flies. (Incorrect because, in example-1, default rules were applied when case-specific information was not available.)

‡ All non-monotonic reasoning are concerned with consistency. Inconsistency is resolved, by removing the relevant conclusion(s) derived by default rules, as shown in the example below.

Example -3 :

The truth value (true or false), of propositions such as "Tweety is a bird" accepts default that is normally true, such as "Birds typically fly". Conclusions derived was "Tweety flies". When an inconsistency is recognized, only the truth value of the last type is changed.

1.2 Different Methods of Reasoning

Mostly three kinds of logical reasoning: Deduction, Induction, Abduction.

- **Deduction**
  - ‡ Example: "When it rains, the grass gets wet. It rains. Thus, the grass is wet."

    This means in determining the conclusion; it is using rule and its precondition to make a conclusion.
  - ‡ Applying a general principle to a special case.
  - ‡ Using theory to make predictions
  - ‡ Usage: Inference engines, Theorem provers, Planning.

- **Induction**
  - ‡ Example: "The grass has been wet every time it has rained. Thus, when it rains, the grass gets wet."

    This means in determining the rule; it is learning the rule after numerous examples of conclusion following the precondition.
  - ‡ Deriving a general principle from special cases
  - ‡ From observations to generalizations to knowledge
  - ‡ Usage: Neural nets, Bayesian nets, Pattern recognition

- **Abduction**
  - Example: "When it rains, the grass gets wet. The grass is wet, it must have rained."

    Means determining the precondition; it is using the conclusion and the rule to support that the precondition could explain the conclusion.
  - Guessing that some general principle can relate a given pattern of cases
  - Extract hypotheses to form a tentative theory
  - Usage: Knowledge discovery, Statistical methods, Data mining.

- **Analogy**
  - Example: "An atom, with its nucleus and electrons, is like the solar system, with its sun and planets."

    Means analogous; it is illustration of an idea by means of a more familiar idea that is similar to it in some significant features. and thus said to be analogous to it.
  - finding a common pattern in different cases
  - usage: Matching labels, Matching sub-graphs, Matching transformations.

Note: Deductive reasoning and Inductive reasoning are the two most commonly used explicit methods of reasoning to reach a conclusion.

- **More about different methods of Reasoning**

  - **Deduction Example**

    Reason from facts and general principles to other facts.

    Guarantees that the conclusion is true.

    - **Modus Ponens** : a valid form of argument affirming the antecedent**.**

      - If it is rainy, John carries an umbrella

        It is rainy
        ----------------- (doted line read as "therefore")
        John carries an umbrella.

      - If p then q

        p
        ----
        q

    - **Modus Tollens :** a valid form of argument denying the consequent.

      - If it is rainy, John carries an umbrella

        John does not carry an umbrella
        ----------------- (doted line read as "because")
        It is not rainy

      - If p then q

        not q
        ----
        not p

- **Induction Example**

  Reasoning from many instances to all instances.

  - **Good Movie**

    Fact — You have liked all movies starring Mery.

    Inference — You will like her next movie.

  - **Birds**

    Facts: — Woodpeckers, swifts, eagles, finches have four toes on each foot.

    Inductive Inference — All birds have 4 toes on each foot.

    (Note: partridges have only 3).

  - **Objects**

    Facts — Cars, bottles, blocks fall if not held up.

    Inductive Inference — If not supported, an object will fall.

    (Note: an unsupported helium balloon will rise.)

  - **Medicine**

    Noted — People who had cowpox did not get smallpox.

    Induction: — Cowpox prevents smallpox.

  Problem : Sometime inference is correct, sometimes not correct.

  Advantage : Inductive inference may be useful even if not correct. It generates a proposition which may be validated deductively.

- **Abduction Example**

Common form of human reasoning– "Inference to the best explanation".

In Abductive reasoning you make an assumption which, if true, together with your general knowledge, will explain the facts.

- **Dating**

Fact:	Mary asks John to a party.
Abductive Inferences	Mary likes John. John is Mary's last choice. Mary wants to make someone else jealous.

- **Smoking house**

Fact:	A large amount of black smoke is coming from a home.
Abduction1:	the house is on fire.
Abduction2:	bad cook.

- **Diagnosis**

Facts:	A thirteen year-old boy has a sharp pain in his right side, a fever, and a high white blood count.
Abductive inference	Appendicitis.

Problem: Not always correct; many explanations possible.

Advantage : Understandable conclusions.

■ Analogy Example

Analogical Reasoning yields conjectures, possibilities.

If A is like B in some ways, then infer A is like B in other ways.

‡ Atom and Solar System

Statements:	An atom, with its nucleus and electrons, is like the solar system, with its sun and planets.
Inferences:	Electrons travel around the nucleus.
	Orbits are circular.
?	Orbits are all in one plane.
?	Electrons have little people living on them.
Idea:	Transfer information from known (source) to unknown (target).

‡ Sun and Girl

Statement:	She is like the sun to me.
Inferences:	She lights up my life.
	She gives me warmth.
?	She is gaseous.
?	She is spherical.

‡ Sale man Logic

Statement:	John has a fancy car and a pretty girlfriend.
Inferences:	If Peter buys a fancy car,
	Then Peter will have a pretty girlfriend.

Problems : Few analogical inferences are correct

Advantage : Suggests novel possibilities. Helps to organize information.

1.3 Sources of Uncertainty in Reasoning

In many problem domains it is not possible to create complete, consistent models of the world. Therefore agents (and people) must act in uncertain worlds (which the real world is). We want an agent to make rational decisions even when there is not enough information to prove that an action will work.

- Uncertainty is omnipresent because of
  - ‡ Incompleteness
  - ‡ Incorrectness
- Uncertainty in Data or Expert Knowledge
  - ‡ Data derived from defaults/assumptions
  - ‡ Inconsistency between knowledge from different experts.
  - ‡ "Best Guesses"
- Uncertainty in Knowledge Representation
  - ‡ Restricted model of the real system.
  - ‡ Limited expressiveness of the representation mechanism.
- Uncertainty in Rules or Inference Process
  - ‡ Incomplete because too many conditions to be explicitly enumerated
  - ‡ Incomplete because some conditions are unknown
  - ‡ Conflict Resolution

1.4 Reasoning and KR

To certain extent, the reasoning depends on the way the knowledge is represented or chosen.

- A good knowledge representation scheme allows easy, natural and plausible (credible) reasoning.

- Reasoning methods are broadly identified as :

  - ‡ Formal reasoning : Using basic rules of inference with logic knowledge representations.
  - ‡ Procedural reasoning : Uses procedures that specify how to perhaps solve sub problems.
  - ‡ Reasoning by analogy : This is as Human do, but more difficult for AI systems.
  - ‡ Generalization and abstraction : This is also as Human do; are basically learning and understanding methods.
  - ‡ Meta-level reasoning : Uses knowledge about what we know and ordering them as per importance.

- Note : What ever may be the reasoning method, the AI model must be able to reason under conditions of uncertainty mentioned before.

1.5 Approaches to Reasoning

There are three different approaches to reasoning under uncertainties.

- ‡ Symbolic reasoning
- ‡ Statistical reasoning
- ‡ Fuzzy logic reasoning

The first two approaches are presented in the subsequent slides.

2. Symbolic Reasoning

The basis for intelligent mathematical software is the integration of the "power of symbolic mathematical tools" with the suitable "proof technology".

Mathematical reasoning enjoys a property called **monotonicity**, that says,

"If a conclusion follows from given premises **A, B, C, ...** then it also follows from any larger set of premises, as long as the original premises **A, B, C, ...** are included."

Human reasoning is not monotonic.

People arrive to conclusions only tentatively, based on partial or incomplete information, reserve the right to retract those conclusions while they learn new facts. Such reasoning is **non-monotonic**, precisely because the set of accepted conclusions have become smaller when the set of premises is expanded.

2.1 Non-Monotonic Reasoning

Non-Monotonic reasoning is a generic name to a class or a specific theory of reasoning. Non-monotonic reasoning attempts to formalize reasoning with incomplete information by classical logic systems.

The Non-Monotonic reasoning are of the type

- Default reasoning
- Circumscription
- Truth Maintenance Systems

- **Default Reasoning**

  This is a very common from of non-monotonic reasoning. The conclusions are drawn based on what is most likely to be true.
  There are two approaches, both are logic type, to Default reasoning : one is Non-monotonic logic and the other is Default logic.

  - **Non-monotonic logic**

    It has already been defined. It says, "the truth of a proposition may change when new information (axioms) are added and a logic may be build to allows the statement to be retracted."

    Non-monotonic logic is predicate logic with one extension called modal operator **M** which means "consistent with everything we know". The purpose of **M** is to allow consistency.

    A way to define consistency with PROLOG notation is :
    To show that fact **P** is true, we attempt to prove **¬P**.
    If we fail we may say that **P** is consistent since **¬P** is false.

    Example :

    $\forall$**x : plays_instrument(x)** $\wedge$ **M manage(x)** $\rightarrow$ **jazz_musician(x)**

    States that for all **x**, the **x** plays an instrument and if the fact that **x** can manage is *consistent* with all other knowledge then we can conclude that **x** is a jazz musician.

Default Logic

Default logic initiates a new inference rule: $\frac{A : B}{C}$ where

- **A** is known as the prerequisite,
- **B** as the justification, and
- **C** as the consequent.

- Read the above inference rule as:
  " if **A**, and if it is consistent with the rest of what is known to assume that **B**, then conclude that **C** ".

- The rule says that given the prerequisite, the consequent can be inferred, provided it is consistent with the rest of the data.

- Example : Rule that "*birds typically fly*" would be represented as

  $$\frac{\text{bird}(x) : \text{flies}(x)}{\text{flies}(x)}$$ which says

  " If **x** is a bird and the claim that **x** flies is consistent with what we know, then infer that **x** flies".

- Note : Since, all we know about Tweety is that :
  Tweety is a bird, we therefore inferred that Tweety flies.

- The idea behind non-monotonic reasoning is to reason with first order logic, and if an inference can not be obtained then use the set of default rules available within the first order formulation.

[continuing default logic]

- Applying Default Rules :

  While applying default rules, it is necessary to check their justifications for consistency, not only with initial data, but also with the consequents of any other default rules that may be applied. The application of one rule may thus block the application of another. To solve this problem, the concept of default theory was extended.

- Default Theory

  It consists of a set of premises **W** and a set of default rules **D**. An extension for a default theory is a set of sentences **E** which can be derived from **W** by applying as many rules of **D** as possible (together with the rules of deductive inference) without generating inconsistency.

  Note : **D** the set of default rules has a unique syntax of the form

$$\frac{\alpha(\vec{x}) \; : \; E\,\beta(\vec{x})}{\gamma(\vec{x})} \quad \text{where}$$

$\alpha(\vec{x})$	is the prerequisite of the default rule
$E\,\beta(\vec{x})$	is the consistency test of the default rule
$\gamma(\vec{x})$	is the consequent of the default rule

The rule can be read as

For all individual $x_1 \; . \; . \; . \; . \; x_m$

If $\alpha(\vec{x})$ is believed and

If each of $\beta(\vec{x})$ is consistent with our beliefs,

Then $\gamma(\vec{x})$ may be believed.

[continuing default logic]

Example :

A Default Rule says " *Typically an American adult owns a car* ".

$$\frac{\mathbf{American(x) \wedge Adult(x) : M((\exists y) \,.\, car(y) \wedge owns(x,y))}}{\mathbf{((\exists y) \,.\, car(y) \wedge owns(x,y))}}$$

The rule is explained below :

The rule is only accessed if we wish to know whether or not John owns a car then an answer can not be deduced from our current beliefs.

This default rule is applicable if we can prove from our beliefs that John is an American and an adult, and believing that there is some car that is owned by John does not lead to an inconsistency.

If these two sets of premises are satisfied, then the rule states that we can conclude that John owns a car.

- **Circumscription**

  Circumscription is a non-monotonic logic to formalize the common sense assumption. Circumscription is a formalized rule of conjecture (guess) that can be used along with the rules of inference of first order logic.

  Circumscription involves formulating rules of thumb with "*abnormality*" predicates and then restricting the extension of these predicates, circumscribing them, so that they apply to only those things to which they are currently known.

  - Example : Take the case of Bird Tweety

    The rule of thumb is that "birds typically fly" is conditional. The predicate "Abnormal" signifies abnormality with respect to flying ability.

    Observe that the rule $\forall$**x(Bird(x) & ¬ Abnormal(x) → Flies))** does not allow us to infer that *"Tweety flies"*, since we do not know that he is abnormal with respect to flying ability.

    But if we add axioms which circumscribe the *abnormality predicate* to which they are currently known say *"Bird Tweety"* then the inference can be drawn. This inference is non-monotonic.

- **Truth Maintenance Systems**

Reasoning Maintenance System (RMS) is a critical part of a reasoning system. Its purpose is to assure that inferences made by the reasoning system (RS) are valid.

The RS provides the RMS with information about each inference it performs, and in return the RMS provides the RS with information about the whole set of inferences.

Several implementations of RMS have been proposed for non-monotonic reasoning. The important ones are the :

- Truth Maintenance Systems (TMS) and
- Assumption-based Truth Maintenance Systems (ATMS).

The TMS maintains the consistency of a knowledge base as soon as new knowledge is added. It considers only one state at a time so it is not possible to manipulate environment.

The ATMS is intended to maintain multiple environments.

The typical functions of TMS are presented in the next slide.

Truth Maintenance Systems (TMS)

A truth maintenance system maintains consistency in knowledge representation of a knowledge base.

The functions of TMS are to :

- **Provide justifications for conclusions**

  When a problem solving system gives an answer to a user's query, an explanation of that answer is required;

  Example : An advice to a stockbroker is supported by an explanation of the reasons for that advice. This is constructed by the Inference Engine (IE) by tracing the justification of the assertion.

- **Recognize inconsistencies**

  The Inference Engine (IE) may tell the TMS that some sentences are contradictory. Then, TMS may find that all those sentences are believed true, and reports to the IE which can eliminate the inconsistencies by determining the assumptions used and changing them appropriately.

  Example : A statement that either Abbott, or Babbitt, or Cabot is guilty together with other statements that Abbott is not guilty, Babbitt is not guilty, and Cabot is not guilty, form a contradiction.

- **Support default reasoning**

  In the absence of any firm knowledge, in many situations we want to reason from default assumptions.

  Example : If "Tweety is a bird", then until told otherwise, assume that "Tweety flies" and for justification use the fact that "Tweety is a bird" and the assumption that "birds fly".

2.2 Implementation Issues

The issues and weaknesses related to implementation of non-monotonic reasoning in problem solving are :

- How to derive exactly those non-monotonic conclusion that are relevant to solving the problem at hand while not wasting time on those that are not necessary.
- How to update our knowledge incrementally as problem solving progresses.
- How to over come the problem where more than one interpretation of the known facts is qualified or approved by the available inference rules.
- In general the theories are not computationally effective, decidable or semi decidable.

The solutions offered, considering the reasoning processes into two parts :

- one, a problem solver that uses whatever mechanism it happens to have to draw conclusions as necessary, and
- second, a truth maintenance system whose job is to maintain consistency in knowledge representation of a knowledge base.

3. Statistical Reasoning :

In the logic based approaches described, we have assumed that everything is either believed false or believed true.

However, it is often useful to represent the fact that we believe such that something is probably true, or true with probability (say) 0.65.

This is useful for dealing with problems where there is randomness and unpredictability (such as in games of chance) and also for dealing with problems where we could, if we had sufficient information, work out exactly what is true.

To do all this in a principled way requires techniques for probabilistic reasoning. In this section, the Bayesian Probability Theory is first described and then discussed how uncertainties are treated.

- **Recall glossary of terms**

  - **Probabilities :**

    Usually, are descriptions of the likelihood of some event occurring (ranging from **0** to **1**).

  - **Event :**

    One or more outcomes of a probability experiment .

  - **Probability Experiment :**

    Process which leads to well-defined results call outcomes.

  - **Sample Space :**

    Set of all possible outcomes of a probability experiment.

  - **Independent Events :**

    Two events, E_1 and E_2, are independent if the fact that E_1 occurs does not affect the probability of E_2 occurring.

  - **Mutually Exclusive Events :**

    Events E_1, E_2, ..., E_n are said to be mutually exclusive if the occurrence of any one of them automatically implies the non-occurrence of the remaining $n - 1$ events.

  - **Disjoint Events :**

    Another name for mutually exclusive events.

- **Classical Probability :**

  Also called a priori theory of probability.

  The probability of event **A** = no of possible outcomes **f** divided by the total no of possible outcomes **n** ; ie., **P(A) = f / n.**

  Assumption: All possible outcomes are equal likely.

- **Empirical Probability :**

  Determined analytically, using knowledge about the nature of the experiment rather than through actual experimentation.

- **Conditional Probability :**

  The probability of some event **A**, given the occurrence of some other event **B**. Conditional probability is written **P(A|B)**, and read as "the probability of **A**, given **B** ".

- **Joint probability :**

  The probability of two events in conjunction. It is the probability of both events together. The joint probability of **A** and **B** is written **P(A ∩ B)** ; also written as **P(A, B)**.

- **Marginal Probability :**

  The probability of one event, regardless of the other event. The marginal probability of **A** is written **P(A)**, and the marginal probability of **B** is written **P(B)**.

- **Examples**

  - **Example 1**

Sample Space - Rolling two dice

The sums can be **{ 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 }.**

Note that each of these are not equally likely. The only way to get a sum **2** is to roll a **1** on both dice, but can get a sum **4** by rolling out comes as **(1,3), (2,2), or (3,1)**.

Table below illustrates a sample space for the sum obtain.

First dice	Second Dice					
	1	2	3	4	5	6
1	**2**	**3**	**4**	**5**	**6**	**7**
2	**3**	**4**	**5**	**6**	**7**	**8**
3	**4**	**5**	**6**	**7**	**8**	**9**
4	**5**	**6**	**7**	**8**	**9**	**10**
5	**6**	**7**	**8**	**9**	**10**	**11**
6	**7**	**8**	**9**	**10**	**11**	**12**

Classical Probability

Table below illustrates frequency and distribution for the above sums.

Sum	2	3	4	5	**6**	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Relative frequency	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	1/36	1/36

The classical probability is the relative frequency of each event.

Classical probability **P(E) = n(E) / n(S)**; **P(6) = 5 / 36, P(8) = 5 / 36**

Empirical Probability

The empirical probability of an event is the relative frequency of a frequency distribution based upon observation **P(E) = f / n**

- **Example 2**

Mutually Exclusive Events (disjoint) : means nothing in common

Two events are mutually exclusive if they cannot occur at the same time.

(a) If two events are mutually exclusive,
then probability of both occurring at same time is **P(A and B) = 0**

(b) If two events are mutually exclusive ,
then the probability of either occurring is **P(A or B) = P(A) + P(B)**

Given **P(A)= 0.20, P(B)= 0.70**, where **A** and **B** are disjoint

then **P(A and B) = 0**

The table below indicates intersections ie **"and"** of each pair of events. "Marginal" means total; the values in bold means given; the rest of the values are obtained by addition and subtraction.

	B	B'	Marginal
A	**0.00**	0.20	**0.20**
A'	0.70	0.10	0.80
Marginal	**0.70**	0.30	1.00

Non-Mutually Exclusive Events

The non-mutually exclusive events have some overlap.

When **P(A)** and **P(B)** are added, the probability of the intersection (ie. "and") is added twice, so subtract once.

P(A or B) = P(A) + P(B) - P(A and B)

Given : **P(A) = 0.20, P(B) = 0.70, P(A and B) = 0.15**

	B	B'	Marginal
A	**0.15**	0.05	**0.20**
A'	0.55	0.25	0.80
Marginal	**0.70**	0.30	1.00

- **Example 3**

Factorial , Permutations and Combinations

Factorial

The factorial of an integer $n \geq 0$ is written as **n!** .

$n! = n \times n\text{-}1 \times . . . \times 2 \times 1$. and in particular, $0! = 1$.

It is, the number of permutations of **n** distinct objects;

e.g., no of ways to arrange 5 letters **A, B, C, D** and **E** into a word is **5!**

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

$N! = (N) \times (N-1) \times (N-2) \times . . . \times (1)$

$n! = n\,(n - 1)!\ ,\quad 0! = 1$

[continuing next slide]

[continuing example 3]

Permutation

The permutation is arranging elements (objects or symbols) into distinguishable sequences. The ordering of the elements is important. Each unique ordering is a permutation.

Number of permutations of ' **n** ' different things taken ' **r** ' at a time is given by $P^n_r = \frac{n!}{(n-r)!}$

(for convenience in writing, here after the symbol P^n_r is written as **nPr** or **P(n,r)**)

Example 1

Consider a total of **10** elements, say integers **{1, 2, ..., 10}.**

A permutation of **3** elements from this set is **(5, 3, 4)**.

Here **n = 10** and **r = 3**.

The number of such unique sequences are calculated as **P(10,3) = 720**.

Example 2

Find the number of ways to arrange the three letters in the word **CAT** in to two-letter groups like **CA** or **AC** and no repeated letters.

This means permutations are of size **r = 2** taken from a set of size **n = 3**. so **P(n, r) = P(3,2) = 6.**

The ways are listed as **CA CT AC AT TC TA.**

Similarly, permutations of size **r = 4**, taken from a set of size **n = 10**,

$$P(n, r) = P(10,4) = \frac{10!}{(10-4)!} = \frac{10!}{6!} = \frac{10x9x8x7x6x5x4x3x2x1}{6x5x4x3x2x1}$$

[continuing next slide]

[continuing example 3]

Combinations

Combination means selection of elements (objects or symbols).

The ordering of the elements has no importance.

Number of Combination of ' **n** ' different things, taken ' **r** ' at a time is

$$C^n_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$ here

r is the size of each combination of elements,

n is the total size of elements from which elements are permuted,

! is the factorial operator.

(for convenience in writing, here after the symbol C^n_r is written as **nCr** or **C(n,r)**)

Example

Find the number of combinations of size **2** without repeated letters that can be made from the three letters in the word **CAT**, order doesn't matter; **AT** is the same as **TA.**

This means combinations of size **r =2** taken from a set of size **n = 3**, so **C(n , r) = C(3 , 2) = 3** . The ways are listed as **CA CT CA** .

Using the formula for finding the number of combinations of **r** objects from a set of **n** objects is:

$$C(n, r) = C(3,2) = \frac{n!}{r!\,(n-r)!} = \frac{3!}{2! \times 1!} = \frac{3 \times 2 \times 1}{2 \times 1 \times (1!)} = \frac{6}{2} = 3$$

If **n** is large then finding **n!** becomes difficult. The alternate way is given below

Find combinations of size **r = 4**, taken from a set of size **n = 10**,

$$C(n, r) = C(10,4) = \frac{P(10,4)}{4!} = \frac{10!}{4! \times 6!} = \frac{10!}{4! \times (10-4)!}$$

$$= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1 \ (6 \times 5 \times 4 \times 3 \times 2 \times 1)}$$

3.1 Probability and Bayes' Theorem

In probability theory, Bayes' theorem relates the conditional and marginal probabilities of two random events.

- **Probability :** The Probabilities are numeric values between **0** and **1** (both inclusive) that represent ideal uncertainties (not beliefs).

  - **Probability of event A is P(A)**

$$P(A) = \frac{\text{instances of the event A}}{\text{total instances}}$$

P(A) = 0	indicates total uncertainty in **A**,
P(A) = 1	indicates total certainty and
0< P(A) < 1	values in between tells degree of uncertainty

Probability Rules :

- All probabilities are between **0** and **1** inclusive **0 <= P(E) <= 1**.
- The sum of all the probabilities in the sample space is **1**.
- The probability of an event which must occur is **1**.
- The probability of the sample space is **1**.
- The probability of any event which is not in the sample space is zero.
- The probability of an event not occurring is **P(E') = 1 - P(E)**

Example 1 : A single 6-sided die is rolled.

What is the probability of each outcome?

What is the probability of rolling an even number?

What is the probability of rolling an odd number?

The possible outcomes of this experiment are 1, 2, 3, 4, 5, 6.

The Probabilities are :

P(1)	=	No of ways to roll 1 / total no of sides	=	1/6
P(2)	=	No of ways to roll 2 / total no of sides	=	1/6
P(3)	=	No of ways to roll 3 / total no of sides	=	1/6
P(4)	=	No of ways to roll 4 / total no of sides	=	1/6
P(5)	=	No of ways to roll 5 / total no of sides	=	1/6
P(6)	=	No of ways to roll 6 / total no of sides	=	1/6
P(even)	=	ways to roll even no / total no of sides	=	3/6 = 1/2
P(odd)	=	ways to roll odd no / total no of sides	=	3/6 = 1/2

Example 2 : Roll two dices

Each dice shows one of 6 possible numbers;

Total unique rolls is 6 x 6 = 36;

List of the joint possibilities for the two dices are:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Roll two dices;

The rolls that add up to 4 are ((1,3), (2,2), (3,1)).

The probability of rolling dices such that total of 4 is 3/36 = 1/12 and the chance of it being true is (1/12) x 100 = 8.3%.

- **Conditional probability P(A|B)**

A conditional probability is the probability of an event given that another event has occurred.

Example : Roll two dices.

What is the probability that the total of two dice will be greater than 8 given that the first die is a 6 ?

First List of the **joint possibilities** for the two dices are:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	**(6, 3)**	**(6, 4)**	**(6, 5)**	**(6, 6)**

There are 6 outcomes for which the first die is a 6, and of these, there are 4 outcomes that total more than 8 are (6,3; 6,4; 6,5; 6,6).

The probability of a total > 8 given that first die is 6 is therefore 4/6 = 2/3 .

This probability is written as: $\underbrace{P(\text{total}>8}_{\textbf{event}} \mid \underbrace{\text{1st die} = 6)}_{\textbf{condition}} = 2/3$

Read as "The probability that the total is > 8 given that die one is 6 is 2/3."

Written as **P(A|B)** , is the probability of event A given that the event B has occurred.

- **Probability of A and B is P(A and B)**

The probability that events **A** and **B** both occur.

Note : Two events are independent if the occurrence of one is unrelated to the probability of the occurrence of the other.

‡ **If A and B are independent**

then probability that events **A** and **B** both occur is:

P(A and B) = P(A) x P(B)

ie product of probability of **A** and probability of **B**.

‡ **If A and B are not independent**

then probability that events **A** and **B** both occur is:

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

P(B|A) is **conditional probability** of **B** given **A**

Example 1: P(A and B) if events A and B are independent

- Draw a card from a deck , then replace it, draw another card.
- Find probability that 1st card is Ace of clubs (event A) and 2nd card is any Club (event B).
- Since there is only one Ace of Clubs, therefore probability P(A) = 1/52.
- Since there are 13 Clubs, the probability P(B) = 13/52 = 1/4.
- Therefore, **P(A and B) =** p(A) x p(B) = 1/52 x 1/4 = 1/208.

Example 2: P(A and B) if events A and B are not independent

- Draw a card from a deck, not replacing it, draw another card.
- Find probability that both cards are Aces ie the 1st card is Ace (event A) and the 2nd card is also Ace (event B).
- Since 4 of 52 cards are Aces, therefore probability P(A) = 4/52 = 1/13.
- Of the 51 remaining cards, 3 are aces. so, probability of 2nd card is also Ace (event B) is P(B|A) = 3/51 = 1/17.
- Therefore, **P(A and B)** = p(A) x p(B|A) = 1/13 x 1/17 = 1/221

- **Probability of A or B is P(A or B)**

  The probability of either event **A** or event **B** occur.

  Two events are mutually exclusive if they cannot occur at same time.

  ‡ **If A and B are mutually exclusive**

  then probability that events **A** or **B** occur is:

  P(A or B) = p(A) + p(B)

  ie sum of probability of **A** and probability of **B**

  ‡ **If A and B are not mutually exclusive**

  then probability that events **A** and **B** both occur is:

  P(A or B) = P(A) x P(B|A) – P(A and B) where

  P(A and B) is probability that events **A** and **B** both occur while events **A** and **B** are independent and **P(B|A)** is conditional probability of **B** given **A**.

  Example 1: P(A or B) if events A or B are mutually exclusive

  - Rolling a die.
  - Find probability of getting either, event A as 1 or event B as 6?
  - Since it is impossible to get both, the event A as 1 and event B as 6 in same roll, these two events are mutually exclusive.
  - The probability P(A) = P(1) = 1/6 and P(B) = P(6) = 1/6
  - Hence probability of either event A or event B is :
  - **P(A or B)** = p(A) + p(B) = 1/6 + 1/6 = 1/3

  Example 2: P(A or B) if events A or B are not mutually exclusive

  - Find probability that a card from a deck will be either an Ace or a Spade?
  - probability P(A) is P(Ace) = 4/52 and P(B) is P(spade) = 13/52.
  - Only way in a single draw to be Ace and Spade is Ace of Spade; which is only one, so probability P(A and B) is P(Ace and Spade) = 1/52.
  - Therefore, the probability of event A or B is :

  P(A or B) = P(A) + P(B) – P(A and B)

  = P(ace) + P(spade) - P(Ace and Spade)

  = 4/52 + 13/52 - 1/52 = 16/52 = 4/13

Summary of symbols & notations

A U B	(A union B)	'Either A or B occurs or both occur'
A ∩ B	(A intersection B)	'Both A and B occur'
A ⊆ B	(A is a subset of B)	'If A occurs, so does B'
A'	$\bar{A}$	'Event A does not occur'
Φ	(the empty set)	An impossible event
S	(the sample space)	An event that is certain to occur
A ∩ B = Φ		Mutually exclusive Events
P(A)		Probability that event A occurs
P(B)		Probability that event B occurs
P(A U B)		Probability that event A or event B occurs
P(A ∩ B)		Probability that event A and event B occur
P(A ∩ B) = P(A) . P(B)		Independent events
P(A ∩ B) = 0		Mutually exclusive Events
P(A U B) = P(A) + P(B) – P(AB)		Addition rule;
P(A U B) = P(A) + P(B) – P(A) . P(B)		Addition rule; independent events
P(A U B) = P(A) + P(B) – P(A ∩ B)		
P(A U B) = P(A) + P(B) – P(B|A).P(A)		
P(A U B) = P(A) + P(B)		Addition rule; mutually exclusive Events
A|B	(A given B)	"Event A will occur given that event B has occurred"
P(A|B)		Conditional probability that event A will occur given that event B has occurred already
P(B|A)		Conditional probability that event B will occur given that event A has occurred already
P(A ∩ B) = P(A|B).P(B) or P(A ∩ B) = P(B|A).P(A)		Multiplication rule
P(A ∩ B) = P(A) . P(B)		Multiplication rule; independent events; ie probability of joint events A and B
P(A|B) = P(A ∩ B) / P(B)		Rule to determine a conditional probability from unconditional probabilities.

- **Bayes' Theorem**

Bayesian view of probability is related to degree of belief.
It is a measure of the plausibility of an event given incomplete knowledge.

Bayes' theorem is also known as Bayes' rule or Bayes' law, or called Bayesian reasoning.

The probability of an event **A** conditional on another event **B** ie **P(A|B)** is generally different from probability of **B** conditional on **A** ie **P(B|A)**.

- There is a definite relationship between the two, P(A|B) and P(B|A), and Bayes' theorem is the statement of that relationship.
- Bayes theorem is a way to calculate P(A|B) from a knowledge of P(B|A).
- Bayes' Theorem is a result that allows new information to be used to update the conditional probability of an event.

[Continued in next slide]

[Continued from previous slide]

- **Bayes' Theorem**

Let **S** be a sample space.

Let **A1, A2, ... , An** be a set of mutually exclusive events from **S**.

Let **B** be any event from the same **S**, such that **P(B) > 0**.

Then Bayes' Theorem describes following two probabilities :

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(A_1 \cap B) + P(A_2 \cap B) + - - - - + P(A_n \cap B)} \quad \text{and}$$

by invoking the fact $P(A_k \cap B) = P(A_k).P(B|A_k)$ the probability

$$P(A_k|B) = \frac{P(A_k).P(B|A_k)}{P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + - - - - + P(A_n).P(B|A_n)}$$

Applying Bayes' Theorem :

Bayes' theorem is applied while following conditions exist.

- the sample space **S** is partitioned into a set of mutually exclusive events **{A1, A2, ... , An }.**
- within **S**, there exists an event **B**, for which **P(B) > 0.**
- the goal is to compute a conditional probability of the form : **P(Ak|B)**.
- you know at least one of the two sets of probabilities described below
  - $P(A_k \cap B)$ for each A_k
  - $P(A_k)$ and $P(B|A_k)$ for each A_k

The Bayes' theorem is best understood through an example below.

Example 1: Applying Bayes' Theorem

Problem : Marie's marriage is tomorrow.

- in recent years, each year it has rained only 5 days.
- the weatherman has predicted rain for tomorrow.
- when it actually rains, the weatherman correctly forecasts rain 90% of the time.
- when it doesn't rain, the weatherman incorrectly forecasts rain 10% of the time.

The question : What is the probability that it will rain on the day of Marie's wedding?

Solution : The sample space is defined by two mutually exclusive events – "it rains" or "it does not rain". Additionally, a third event occurs when the "weatherman predicts rain".

The events and probabilities are stated below.

- Event A1 : rains on Marie's wedding.
- Event A2 : does not rain on Marie's wedding
- Event B : weatherman predicts rain.
- P(A1) = 5/365 =0.0136985 [Rains 5 days in a year.]
- P(A2) = 360/365 = 0.9863014 [Does not rain 360 days in a year.]
- P(B|A1) = 0.9 [When it rains, the weatherman predicts rain 90% time.]
- P(B|A2) = 0.1 [When it does not rain, weatherman predicts rain 10% time.]

We want to know **P(A1|B)**, the probability that it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman.

The answer can be determined from Bayes' theorem, shown below.

$$P(A1|B) = \frac{P(A1).P(B|A1)}{P(A1).P(B|A1)+P(A2).P(B|A2)} = \frac{(0.014)(0.9)}{[(0.014)(0.9)+(0.986)(0.1)]}$$

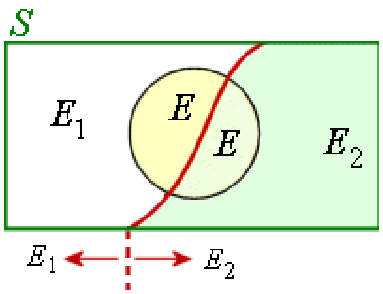
$$= 0.111$$

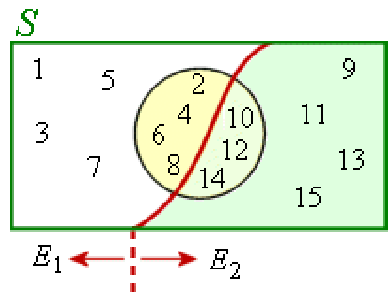
So, despite the weatherman's prediction, there is a good chance that Marie will not get rain on at her wedding.

Thus Bayes theorem is used to calculate conditional probabilities.

Example 2: Applying Bayes' Theorem

- Let **S** be a sample space.
- Let **E_1** and **E_2** be two mutually exclusive events forming a partition of the sample space **S**
- Let **E** be any event of the sample space such that **$P(E) \neq 0$**.





Recall from Conditional Probability

The notation **$P(E_1 \mid E)$** means "the probability of the event **E_1** given that **E** has already occurred".

- The sample space **S** is described as "the integers 1 to 15" and is partitioned into :

  E_1 = "the integers 1 to 8" and

  E_2 = "the integers 9 to 15".

- If **E** is the event "**even number**" then the probabilities for the situation described by Baye's Theorem can be calculated in two ways, both giving same results.

$$P(E_1|E) = \frac{P(E_1 \cap E)}{P(E_1 \cap E) + P(E_2 \cap E)} = \frac{4 / 15}{(4 / 15) + (3 / 15)} = 4 / 7$$

$$P(E_1|E) = \frac{P(E_1).P(E|E_1)}{P(E_1).P(E|E_1) + P(E_2).P(E|E_2)} = \frac{8 / 15 \times 4 / 8}{(8/15 \times 4/8) + (7/15 \times 3/15)} = 4 / 7$$

Thus Bayes' Theorem can be extended for Mutually Exclusive Events as :

$$P(E_i \mid E) = \frac{P(E_i \cap E)}{P(E_1 \cap E) + P(E_2 \cap E) + \ldots\ldots + P(E_k \cap E)}$$

Example 3 : Clinic Trial

In a clinic, the probability of the patients having HIV virus is **0.15**.

A blood test done on patients :

If patient has virus, then the test is **+ve** with probability **0.95**.

If the patient does not have the virus, then the test is **+ve** with probability **0.02**.

Assign labels to events : **H** = patient has virus; **P** = test +ve

Given : **$P(H) = 0.15$; $P(P|H) = 0.95$; $P(P|\neg H) = 0.02$**

Find :

If the test is **+ve** what are the probabilities that the patient

i) has the virus ie **$P(H|P)$** ; ii) does not have virus ie **$P(\neg H|P)$** ;

If the test is **-ve** what are the probabilities that the patient

iii) has the virus ie **$P(H|\neg P)$** ; iv) does not have virus ie **$P(\neg H|\neg P)$** ;

Calculations :

i) For **$P(H|P)$** we can write down Bayes Theorem as

$P(H|P) = [P(P|H)\ P(H)] / P(P)$

We know $P(P|H)$ and $P(H)$ but not $P(P)$ which is probability of a **+ve** result.

There are two cases, that a patient could have a **+ve** result, stated below :

1. Patient has virus and gets a **+ve** result : $H \cap P$
2. Patient does not have virus and gets a **+ve** result: $\neg H \cap P$

Find probabilities for the above two cases and then add

ie $P(P) = P(H \cap P) + P(\neg H \cap P)$.

But from the second axiom of probability we have :

$P(H \cap P) = P(P|H)\ P(H)$ and $P(\neg H \cap P) = P(P|\neg H)\ P(\neg H)$.

Therefore putting these we get :

$P(P) = P(P|H)\ P(H) + P(P|\neg H)\ P(\neg H) = 0.95 \times 0.15 + 0.02 \times 0.85 = 0.1595$

Now substitute this into Bayes Theorem and obtain $P(H|P)$

$$P(H|P) = \frac{P(P|H)\ P(H)}{P(P|H)\ P(H) + P(P|\neg H)\ P(\neg H)} = 0.95 \times 0.15 / 0.1595 = 0.8934$$

ii) Next is to work out $P(\neg H|P)$

$P(\neg H|P) = 1 - P(H|P) = 1 - 0.8934 = 0.1066$

iii) Next is to work out $P(H|\neg P)$; again we write down Bayes Theorem

$$P(H|\neg P) = \frac{P(\neg P|H)\ P(H)}{P(\neg P)}$$ here we need $P(\neg P)$ which is $1 - P(P)$

$= (0.05 \times 0.15)/(1-0.1595) = 0.008923$

iv) Finally, work out $P(\neg H|\neg P)$

It is just $1 - P(H|\neg P) = 1- 0.008923 = 0.99107$

3.2 Certainty Factors in Rule-Based Systems

The certainty-factor model was one of the most popular model for the representation and manipulation of uncertain knowledge in the early (1980s) Rule-based expert systems.

The model was criticized by researchers in artificial intelligence and statistics being ad-hoc-in nature. Researchers and developers have stopped using the model.

Its place has been taken by more expressive formalisms of Bayesian belief networks for the representation and manipulation of **uncertain knowledge**.

The manipulation of uncertain knowledge in the Rule-based expert systems is illustrated in the next three slide before moving to Bayesian Networks.

- **Rule Based Systems**

Rule based systems have been discussed in previous lectures.
Here it is recalled to explain uncertainty.

- A rule is an expression of the form **"if A then B"** where **A** is an assertion and **B** can be either an action or another assertion.

  Example : Trouble shooting of water pumps

  1. If pump failure then the pressure is low
  2. If pump failure then check oil level
  3. If power failure then pump failure

- Rule based system consists of a library of such rules.
- Rules reflect essential relationships within the domain.
- Rules reflect ways to reason about the domain.
- Rules draw conclusions and points to actions, when specific information about the domain comes in. This is called inference.
- The inference is a kind of chain reaction like :

  If there is a power failure then (see rules 1, 2, 3 mentioned above)

  Rule 3 states that there is a pump failure, and

  Rule 1 tells that the pressure is low, and

  Rule 2 gives a (useless) recommendation to check the oil level.

- It is very difficult to control such a mixture of inference back and forth in the same session and resolve such uncertainties.

  How to deal such uncertainties ?

[continued in the next slide]

[continued from the previous slide]

How to deal uncertainties in rule based system?

A problem with rule-based systems is that often the connections reflected by the rules are not absolutely certain (i.e. deterministic), and the gathered information is often subject to uncertainty.

In such cases, a certainty measure is added to the premises as well as the conclusions in the rules of the system.

A rule then provides a function that describes : how much a change in the certainty of the premise will change the certainty of the conclusion.

In its simplest form, this looks like :

If A (with certainty x) then B (with certainty f(x))

This is a new rule, say rule 4, added to earlier three rules.

[continued in the next slide]

[continued from the previous slide]

- There are many schemes for treating uncertainty in rule based systems.

  The most common are :

  ‡ Adding certainty factors.

  ‡ Adoptions of Dempster-Shafer belief functions.

  ‡ Inclusion of fuzzy logic.

  In these schemes, uncertainty is treated locally, means action is connected directly to incoming rules and uncertainty of their elements. Example : In addition to rule 4 , in previous slide, we have the rule

  If C (with certainty x) then B (with certainty g(x))

  Now If the information is that **A** holds with certainty **a** and **C** holds with certainty **c**, Then what is the certainty of **B** ?

  Note : Depending on the scheme, there are different algebras for such a combination of uncertainty. But all these algebras in many cases come to incorrect conclusions because combination of uncertainty is not a local phenomenon, but it is strongly dependent on the entire situation (in principle a global matter).

3.3 Bayesian Networks and Certainty Factors

A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. For example, a Bayesian network could represent the **probabilistic relationships** between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Bayesian Networks are also called : Bayes nets, Bayesian Belief Networks (BBNs) or simply Belief Networks. Causal Probabilistic Networks (CPNs).

A Bayesian network consists of :

- a set of nodes and a set of directed edges between nodes.
- the edges reflect cause-effect relations within the domain.
- The effects are not completely deterministic (e.g. disease -> symptom).
- the strength of an effect is modeled as a probability.

- **Bayesian Networks**

We have applied Bayesian probability theory, in earlier three examples (example 1, 2, and 3) , to relate two or more events. But this can be used to relate many events by tying them together in a network.

Consider the previous example 3 - Clinic trial

The trial says, the probability of the patients having HIV virus is **0.15**.

A blood test done on patients :

If patient has virus, the test is **+ve** with probability **0.95**.

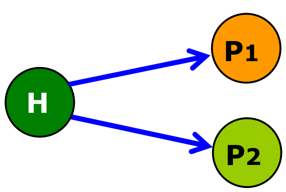
If the patient does not have the virus, the test is **+ve** with probability **0.02**.

This means given : **P(H) = 0.15 ; P(P|H) = 0.95 ; P(P|¬H) = 0.02**

Imagine, the patient is given a second test independently of the first; means the second test is done at a later date by a different person using different equipment. So, the error on the first test does not affect the probability of an error on the second test.

In other words the two tests are independent. This is depicted using the diagram below :

A simple example of a Bayesian Network.



Event **H** is the cause of the two events **P1** and **P2**.

The arrows represent the fact that **H** is driving **P1** and **P2**.

The network contained 3 nodes.

If both **P1** and **P2** are **+ve**

then find the probability that patient has the virus ?

In other words asked to find **P(H|P1 ∩ P2)** .

How to find ?

[continued in the next slide]

- **Bayes Theorem** *(Ref. previous previous slide example 3)*

$$P(H|P1 \cap P2) = \frac{P(P1 \cap P2|H) \cdot P(H)}{P(P1 \cap P2)}$$

Here there are two quantities which we do not know.

The first is $P(P1 \cap P2|H)$ and the second is $P(P1 \cap P2)$

‡ Find $P(P1 \cap P2|H)$

Since the two tests are independent, so

$$P(P1 \cap P2|H) = (P1|H)P(P2|H)$$

‡ Find $P(P1 \cap P2)$

As worked before for $P(P)$ which is the probability of a **+ve** result, here again break this into two separate cases:

- patient has virus and both tests are +ve
- patient not having virus and both tests are +ve

‡ As before use the second axiom of probability

$$P(P1 \cap P2) = P(P1 \cap P2\ |H)\ P(H) + P(P1 \cap P2\ |\neg H)\ P(\neg H)$$

‡ Because the two tests are independent given **H** we can write :

$$P(P1 \cap P2) = P(P1|H)\ P(P2|H)\ P(H) + P(P1|\neg H)\ P(P2|\neg H)\ P(\neg H)$$

$$= 0.95 \times 0.95 \times 0.15 + 0.02 \times 0.02 \times 0.85$$

$$= 0.135715$$

‡ Substitute this into Bayes Theorem above and obtain

$$P(H|P1 \cap P2) = \frac{P(P1 \cap P2|H) \cdot P(H)}{P(P1 \cap P2)}$$

$$= (0.95 \times 0.95 \times 0.15) / 0.135715 = 0.99749$$

‡ Note : The results while two independent HIV tests performed

- Previously we calculated the probability, that the patient had HIV given **one +ve test**, as **0.8934**.
- Later second HIV test was performed. After **two +ve tests**, we see that the probability has gone up to **0.99749**.
- So after two +ve tests it is more certain that the patient does have the HIV virus.

The next slide : a case where one tests is +ve and other is -ve.

Case where one tests is +ve and other is -ve.

This means, an error on one of the tests but we don't know which one; it may be any one.

The issue is - whether the patient has HIV virus or not ?

- We need to calculate **$P(H| P1 \cap \neg P2)$**.
  Following same steps for the case of two +ve tests, write Bayes Theorem

$$P(H| P1 \cap \neg P2) = \frac{P(P1 \cap \neg P2 \mid H)\ P(H)}{P(P1 \cap \neg P2)}$$

- Now work out **$P(P1 \cap \neg P2 \mid H)$** and **$P(P1 \cap \neg P2)$** using the fact that **P1** and **P2** are independent given **H** ,

$$P(P1 \cap \neg P2 \mid H) = P(P1|H)\ P(\neg P2|H) \quad \text{and}$$

$$\begin{aligned} P(P1 \cap \neg P2) &= P(P1 \cap \neg P2 \mid H)\ P(H) + P(P1 \cap \neg P2 \mid \neg H)\ P(\neg H) \\ &= P(P1|H)\ P(\neg P2 \mid H)\ P(H) + P(P1|\neg H)\ P(\neg P2|\neg H)\ P(\neg H) \\ &= 0.95 \times 0.05 \times 0.15 + 0.02 \times 0.98 \times 0.85 \\ &= 0.023785 \end{aligned}$$

- Substitute these values into Bayes Theorem, we obtain

$$P(H| P1 \cap \neg P2) = \frac{0.95 \times 0.05 \times 0.15}{0.023785} = 0.299$$

- Note :
  - Belief in **H**, the event that the patient has virus, has increased.
  - Prior belief was **0.15** but it has now gone up to **0.299**.
  - This appears strange because we have been given two contradictory pieces of data. But looking closely we see that probability of an error in each case is not equal.
- The probability of a +ve test when patient is actually -ve is **0.02**.
  The probability of a -ve test when patient is actually +ve is **0.05**.
  Therefore we are more inclined to believe an error on the second test and this slightly increases our belief that the patient is +ve.

- **More Complicated Bayesian Networks**

The previous network was simple contained three nodes. Let us look at a slightly more complicated one in the context of heart disease.

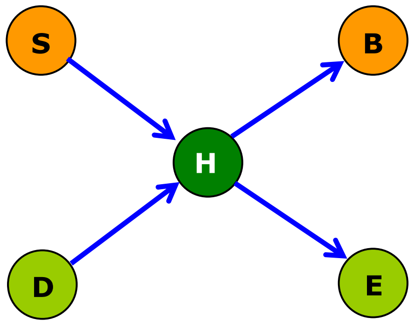
Given the following facts about heart disease.

- Either smoking or bad diet or both can make heart disease more likely.
- Heart disease can produce either or both of the following two symptoms:
  - ‡ high blood pressure
  - ‡ an abnormal electrocardiogram
- Here smoking and bad diet are regarded as causes of heart disease. The heart disease in turn is a cause of high blood pressure and an abnormal electrocardiogram.

[continued in the next slide]

[continued from the previous slide]

- An appropriate network for heart disease is represented as



The symbols define :

S = smoking,

D = bad diet,

H = heart disease,

B = high blood pressure,

E = abnormal electrocardiogram

Here **H** has two causes **S** and **D**.

Find probability of **H,** given each of the four possible combinations of **S** and **D**.

A medical survey gives us the following data :

$P(S) = 0.3$ $P(D) = 0.4$

$P(H | S \cap D) = 0.8$

$P(H | \neg S \cap D) = 0.5$

$P(H | S \cap \neg D) = 0.4$

$P(H | \neg S \cap \neg D) = 0.1$

$P(B | H) = 0.7$ $P(B | \neg H) = 0.1$

$P(E | H) = 0.8$ $P(E | \neg H) = 0.1$

Given these information, an answer to the question concerning this network :

what is the probability of heart disease ?

[Note : The interested students may try to the find answer.]

3.4 Dempster – Shafer Theory (DST)

DST is a mathematical **theory of evidence** based on belief functions and plausible reasoning. It is used to combine separate pieces of information (evidence) to calculate the probability of an event.

DST offers an alternative to traditional probabilistic theory for the mathematical representation of uncertainty.

DST can be regarded as, a more general approach to represent uncertainty than the Bayesian approach.
Bayesian methods are sometimes inappropriate

Example :

Let **A** represent the proposition **"Moore is attractive"**.

Then the axioms of probability insist that **P(A) + P(¬A) = 1**.

Now suppose that Andrew does not even know who **"Moore"** is, then

- ‡ We cannot say that Andrew believes the proposition if he has no idea what it means.
- ‡ Also, it is not fair to say that he disbelieves the proposition.
- ‡ It would therefore be meaningful to denote Andrew's belief **B** of **B(A)** and **B(¬A)** as both being **0**.
- ‡ Certainty factors do not allow this.

- **Dempster-Shafer Model**

The idea is to allocate a number between **0** and **1** to indicate a degree of belief on a proposal as in the probability framework.
However, it is not considered a probability but a belief mass.
The distribution of masses is called basic belief assignment.

In other words, in this formalism a degree of belief (referred as mass) is represented as a belief function rather than a Bayesian probability distribution.

Example: Belief assignment *(continued from previous slide)*

Suppose a system has five members, say five independent states, and exactly one of which is actual. If the original set is called **S, | *S* | = 5**, then the set of all subsets (the power set) is called $\mathbf{2^S}$.

- If each possible subset as a binary vector (describing any member is present or not by writing **1** or **0**), then $\mathbf{2^5}$ subsets are possible, ranging from the empty subset **(0, 0, 0, 0, 0)** to the "everything" subset **(1, 1, 1, 1, 1)**.
- The "empty" subset represents a "contradiction", which is not true in any state, and is thus assigned a mass of **one** ;
- The remaining masses are normalized so that their total is **1**.
- The "everything" subset is labeled as "unknown"; it represents the state where all elements are present **one** , in the sense that you cannot tell which is actual.

Note : Given a set S, the power set of S, written 2^S, is the set of all subsets of S, including the empty set and S.

- **Belief and Plausibility**

Shafer's framework allows for belief about propositions to be represented as intervals, bounded by two values, belief (or support) and plausibility:

belief ≤ plausibility

Belief in a hypothesis is constituted by the sum of the masses of all sets enclosed by it (i.e. the sum of the masses of all subsets of the hypothesis). It is the amount of belief that directly supports a given hypothesis at least in part, forming a lower bound.

Plausibility is 1 minus the sum of the masses of all sets whose intersection with the hypothesis is empty. It is an upper bound on the possibility that the hypothesis could possibly happen, up to that value, because there is only so much evidence that contradicts that hypothesis.

Example :

A proposition say **"the cat in the box is dead."**

Suppose we have **belief of 0.5** and **plausibility of 0.8** for the proposition.

[continued in the next slide]

[continued in the previous slide]

Example :

A proposition say **"the cat in the box is dead."**

Suppose we have **belief of 0.5** and **plausibility of 0.8** for the proposition.

- Evidence to state strongly, that proposition is true with confidence **0.5.**
- Evidence contrary to hypothesis ("the cat is alive") has confidence **0.2**.
- Remaining mass of **0.3** (the gap between the **0.5** supporting evidence and the **0.2** contrary evidence) is "**indeterminate**," meaning that the cat could either be dead or alive. This interval represents the level of uncertainty based on the evidence in the system.

Hypothesis	Mass	belief	plausibility
Null (neither alive nor dead)	0	0	0
Alive	0.2	0.2	0.5
Dead	0.5	0.5	0.8
Either (alive or dead)	0.3	1.0	1.0

- Null hypothesis is set to **zero** by definition, corresponds to "no solution".
- Orthogonal hypotheses "Alive" and "Dead" have probabilities of **0.2** and **0.5**, respectively. This could correspond to "Live/Dead Cat Detector" signals, which have respective reliabilities of **0.2** and **0.5**.
- All-encompassing "Either" hypothesis (simply acknowledges there is a cat in the box) picks up the slack so that the sum of the masses is **1.**
- Belief for the "Alive" and "Dead" hypotheses matches their corresponding masses because they have no subsets;
- Belief for "Either" consists of the sum of all three masses (Either, Alive, and Dead) because "Alive" and "Dead" are each subsets of "Either".
- "Alive" plausibility is **1- m** (Death) and "Dead" plausibility is **1- m** (Alive).
- "Either" plausibility sums **m(Alive) + m(Dead) + m(Either).**
- Universal hypothesis ("Either") will always have **100%** belief and plausibility; it acts as a checksum of sorts.

- **Dempster-Shafer Calculus**

In the previous slides, two specific examples of Belief and plausibility have been stated. It would now be easy to understand their generalization.

The Dempster-Shafer (DS) Theory, requires a Universe of Discourse **U** (or Frame of Judgment) consisting of mutually exclusive alternatives, corresponding to an attribute value domain. For instance, in satellite image classification the set **U** may consist of all possible classes of interest.

Each subset $S \subseteq U$ is assigned a basic probability **m(S)**, a belief **Bel(S)**, and a plausibility **Pls(S)** so that

m(S), Bel(S), Pls(S) ∈ [0, 1] and **Pls(S) ≥ Bel(S)** where

- **m** represents the strength of an evidence, is the basic probability; e.g., a group of pixels belong to certain class, may be assigned value **m**.
- **Bel(S)** summarizes all the reasons to believe **S**.
- **Pls(S)** expresses how much one should believe in **S** if all currently unknown facts were to support **S**.

The true belief in **S** is somewhere in the belief interval **[Bel(S), Pls(S)]**.

The basic probability assignment **m** is defined as function $m : 2^U \rightarrow [0,1]$, where $m(\emptyset) = 0$ and sum of **m** over all subsets of **U** is **1** (i.e., $\sum_{S \subseteq U} m(s) = 1$).

For a given basic probability assignment **m**, the belief **Bel** of a subset **A** of **U** is the sum of **m(B)** for all subsets **B** of **A** , and the plausibility **Pls** of a subset **A** of **U** is **Pls(A) = 1 - Bel(A') (5)** where **A'** is complement of **A** in **U**.

[continued in the next slide]

[continued in the next slide]

Summarize :

The **confidence interval** is that interval of probabilities within which the true probability lies with a certain confidence based on the belief "B" and plausibility "PL" provided by some evidence "E" for a proposition "P".

The **belief** brings together all the evidence that would lead us to believe in the proposition **P** with some certainty.

The **plausibility** brings together the evidence that is compatible with the proposition **P** and is not inconsistent with it.

If **"Ω"** is the set of possible outcomes, then a mass probability "**M**" is defined for each member of the set 2^{Ω} and takes values in the range **[0,1]** . The Null set, **"ф"**, is also a member of 2^{Ω} .

Example

If **Ω** is the **set { Flu (F), Cold (C), Pneumonia (P) }**

Then 2^{Ω} is the **set {ф, {F}, {C}, {P}, {F, C}, {F, P}, {C, P}, {F, C, P}}**

Confidence interval is then defined as **[B(E), PL(E)]** where

$B(E) = \Sigma_A M$, where $A \subseteq E$ i.e., all evidence that makes us believe in the correctness of **P**, and

$PL(E) = 1 - B(\neg E) = \Sigma_{\neg A} M$, where $\neg A \subseteq \neg E$ i.e., all the evidence that contradicts **P**.

- **Combining Beliefs**

The Dempster-Shafer calculus combines the available evidences resulting in a belief and a plausibility in the combined evidence that represents a consensus on the correspondence. The model maximizes the belief in the combined evidences.

The rule of combination states that two basic probability assignments M_1 and M_2 are combined into a third basic probability assignment by the normalized orthogonal sum $m_1 \oplus m_2$ stated below.

Suppose M_1 and M_2 are two belief functions.
Let X be the set of subsets of Ω to which M_1 assigns a nonzero value and let Y be a similar set for M_2 ,
then a new belief function M_3 from the combination of beliefs in M_1 and M_2 is obtained as

$$M_3\,(Z) = \frac{\Sigma_{X \cap Y = Z}\; M_1(X)\, M_2(Y)}{1 - K}$$

where $\Sigma_{X \cap Y = \phi}\; M_1(X)\; M_2(Y)$, for $Z = \phi$

$M_3\,(\phi)$ is defined to be 0 so that the orthogonal sum remains a basic probability assignment.

3.5 Fuzzy Logic

We have discussed only binary valued logic and classical set theory like :

> A person belongs to a set of all human beings, and if given a specific subset, say all males, then one can say whether or not the particular person belongs to this set.

This is ok since it is the way human reason. e.g.,

IF person is male AND a parent THEN person is a father.

The rules are formed using operators.

Here, it is intersection operator "AND" which manipulates the sets.

However, not everything can be described using binary valued sets.

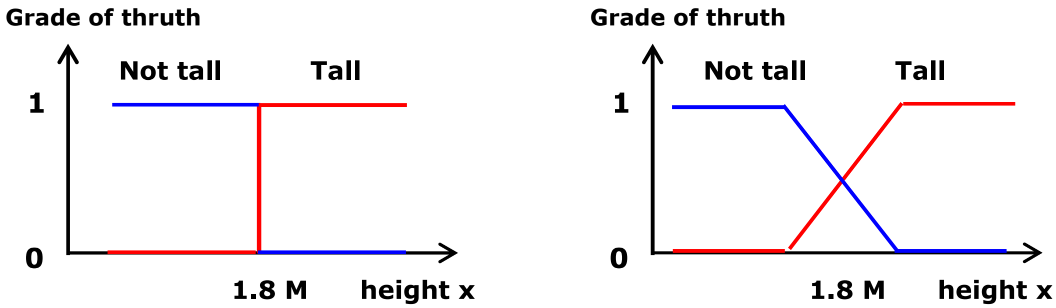
- The grouping of persons into "male" or "female" is easy, but as "tall" or "not tall" is problematic.
- A set of "tall" people is difficult to define, because there is no distinct cut-off point at which tall begins.

Fuzzy logic was suggested by Zadeh as a method for mimicking the ability of human reasoning using a small number of rules and still producing a smooth output via a process of interpolation.

- **Description of Fuzzy Logic**

With fuzzy logic an element could partially belong to a set represented by the **set membership**. Example, a person of height **1.79 m** would belong to both tall and not tall sets with a particular degree of membership.

Difference between binary logic and fuzzy logic



Binary valued logic {0, 1} **Fuzzy logic [0, 1]**

A fuzzy logic system is one that has at least one system component that uses fuzzy logic for its internal knowledge representation.

Fuzzy system communicate information using fuzzy sets.

Fuzzy logic is used purely for internal knowledge representation and externally it can be considered as any other system component.

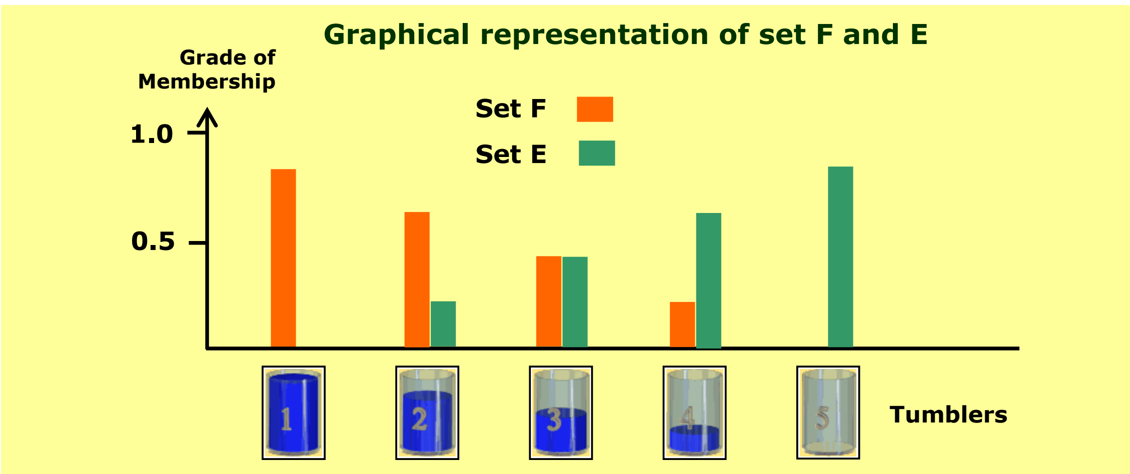
- **Fuzzy Membership**

Example : Five tumblers

- Consider two sets: **F** and **E**.
- **F** is set of all tumblers belong to the class full, and
- **E** is set of all tumblers belong to the class empty.

Definition of the set F and E

Tumblers	1	2	3	4	5
Grade of membership to set ***F***	**100%**	**75%**	**50%**	**25%**	**0%**
Grade of membership to set ***E***	**0%**	**25%**	**50%**	**75%**	**100%**



The sets **F** and **E** have some elements, having partial membership. Such kind of non-crisp sets are called **fuzzy sets.**
The set "all tumblers" here is the basis of the fuzzy sets **F** and **E**, is called the **base set.**

4. References : Textbooks

1. *"Artificial Intelligence", by Elaine Rich and Kevin Knight, (2006), McGraw Hill companies Inc., Chapter 7-8, page 195-248.*

2. *"Artificial Intelligence: A Modern Approach" by Stuart Russell and Peter Norvig, (2002), Prentice Hall, Chapter 7-10, page 195-369.*

3. *"Computational Intelligence: A Logical Approach", by David Poole, Alan Mackworth, and Randy Goebel, (1998), Oxford University Press, Chapter 2-3, page 23-104.*

4. *"Artificial Intelligence: Structures and Strategies for Complex Problem Solving", by George F. Luger, (2002), Addison-Wesley, Chapter 9, page 333-380.*

5. *"AI: A New Synthesis", by Nils J. Nilsson, (1998), Morgan Kaufmann Inc., Chapter 13-20, Page 217-358.*

6. *"Artificial Intelligence: Theory and Practice", by Thomas Dean, (1994), Addison-Wesley, Chapter 2-8, Page 23-355.*

7. *Related documents from open source, mainly internet. An exhaustive list is being prepared for inclusion at a later date.*