Unit-6

Probabilistic Reasoning

Uncertainty in AI: Motivation

- Now, moving towards Traditional AI to Modern AI
- Think about what is it that we have resolved so far in the field?
- I give you a new puzzle, I give you a new setting, I put you in a different setting you will not train for it - you should be able to solve it. You should be able to reason about it.

✔ let us come up with a very general algorithm – search problem

✔ uninformed systematic search to heuristic function in form systematic search to local search and to backtracking search and so on.

✔ But sometimes the problem has some characteristics which we can do the reason which to make some inferences.

✔ So we said okay search and inference together is even better than search right.

Cont’d

- ✔ Earlier in search our language was just a program blackbox but now we said let us go inside the black box let us define the language.
- ✔ we define the language of constraint satisfaction problems
- ✔ We said let us define a more interesting language we define the language of logic.
- ✔ we learned about some algorithms for inference and logic.
- ✔ then finally we said that all of this for single agent problems but we often have multi agent problems.
- ✔ and we talked about the minimax algorithm
- ✔ and we talked about the alpha-beta pruning algorithm

Cont’d

- All these traditional AI considered only deterministic problem.
- There is no notion of randomness, there is no notion of uncertainty
- Think for a robot picking up a cup, what’s possible outcome?
    - Heavy, stuck to the bottom, not cup at all and so on…
- In logic, we have to define a variable for each such possibility.

- Suppose, cuff and cold if these are the only symptoms I give you, which disease you will say?
- you say common cold. Why did not you say throat cancer?
- It is less probable? you have some intuition?

Cont'd

- Can logic deal with all of that?
- Can logic say that you could have throat cancer but the probability of throat cancer is ridiculously small and you probably have common cold because that is the most frequent.
- And unfortunately the traditional AI could not do this.
- And therefore logic went out of fashion and what came into fashion at the time is **probabilistic modeling** which is what we are going to start with.

Probabilistic reasoning in Artificial intelligence

- Till now, we have learned knowledge representation using first-order logic and propositional logic with certainty, which means we were sure about the predicates.
- With this knowledge representation, we might write $A \rightarrow B$, which means if A is true then B is true, but consider a situation where we are not sure about whether A is true or not then we cannot express this statement, this situation is called **uncertainty**.
- So to represent uncertain knowledge, where we are not sure about the predicates, we need **uncertain reasoning** or **probabilistic reasoning**.
- Causes of uncertainty in real world:
  - Information occurred from unreliable sources.
  - Experimental Errors
  - Equipment fault
  - Temperature variation
  - Climate change

Probabilistic reasoning

- Probabilistic reasoning is a way of knowledge representation where we apply the concept of probability to indicate the uncertainty in knowledge.
- In probabilistic reasoning, we combine probability theory with logic to handle the uncertainty.
- We use probability in probabilistic reasoning because it provides a way to handle the uncertainty that is the result of someone's laziness and ignorance.
- In the real world, there are lots of scenarios, where the certainty of something is not confirmed, such as
  - "It will rain today,"
  - "behavior of someone for some situations,"
  - "A match between two teams or two players."
- These are probable sentences for which we can assume that it will happen but not sure about it, so here we use probabilistic reasoning.

Need of probabilistic reasoning in AI

- When there are unpredictable outcomes.
- When specifications or possibilities of predicates becomes too large to handle.
- When an unknown error occurs during an experiment.

In probabilistic reasoning, there are two ways to solve problems with uncertain knowledge:

- Bayes' rule
- Bayesian Statistics

[Note] As probabilistic reasoning uses probability and related terms, so before understanding probabilistic reasoning, let's understand some common terms.

Probability

- Probability can be defined as a chance that an uncertain event will occur.
- It is the numerical measure of the likelihood that an event will occur.
- The value of probability always remains between 0 and 1 that represent ideal uncertainties.

  $0 \leq P(A) \leq 1$, where $P(A)$ is the probability of an event A.

  $P(A) = 0$, indicates total uncertainty in an event A.

  $P(A) = 1$, indicates total certainty in an event A.

- We can find the probability of an uncertain event by using the below formula

$$\textbf{Probability of occurrence} = \frac{\text{Number of desired outcomes}}{\text{Total number of outcomes}}$$

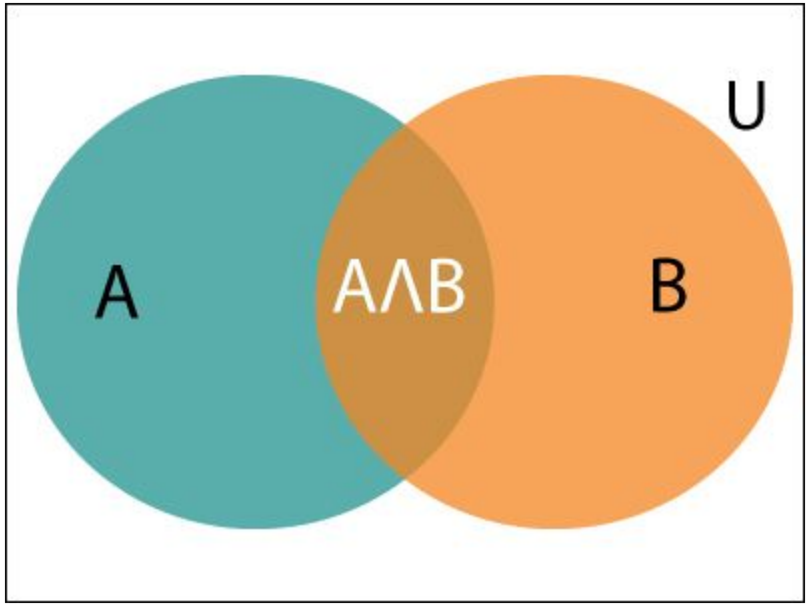
- **Event:** Each possible outcome of a variable is called an event
- **Sample space:** The collection of all possible events is called sample space.
- **Random variables:** Random variables are used to represent the events and objects in the real world.
- **Prior probability:** The prior probability of an event is probability computed before observing new information.
- **Posterior Probability:** The probability that is calculated after all evidence or information has taken into account. It is a combination of prior probability and new information.

Conditional probability

- Conditional probability is a probability of occurring an event when another event has already happened.
- suppose, we want to calculate the event A when event B has already occurred, "the probability of A under the conditions of B", it can be written as:

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

- Where P(A ∧ B)= Joint probability of A and B
- P(B)= Marginal probability of B.



Example

- In a class, there are 70% of the students who like English and 40% of the students who likes English and Mathematics.
- What is the percent of students those who like English also like Mathematics?

- Let, A is an event that a student likes Mathematics
- B is an event that a student likes English.

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{0.4}{0.7} = 57\%$$

- Hence, 57% are the students who like English also like Mathematics.

BAYE'S THEOREM: Describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

↳ In Probability theory it relates the Conditional probability & marginal probabilities of two random events.

↳ $P(H|E) = \dfrac{\text{no. of time H and E}}{\text{no. of times E}}$

↳ Calculate $P(B|A)$ with Knowledge of $P(A|B)$.

$P(H|E) = \dfrac{P(H \cap E)}{P(E)}$ } Prob. of H when E is true.

$$\left.\begin{aligned} P(A \cap B) &= P(A|B).P(B) \quad \text{--- (i)} \\ P(A \cap B) &= P(B|A).P(A) \quad \text{--- (ii)} \end{aligned}\right\} \text{from (i) and (ii) L.H.S are equal.}$$

↳ $P(A|B).P(B) = P(B|A).P(A)$

→ Posterior (Prob. of A when B is true)

→ marginal Prob. (Prob. of evidence)

So, $\boxed{P(B|A) = \dfrac{P(A|B).P(B)}{P(A)}}$ Baye's theorem formula.

Likeli hood (Prob. of evidence)

→ Prior Prob (Prob. of hypothesis)

Applying Bayes' rule

- Bayes' rule allows us to compute the single term P(B|A) in terms of P(A|B), P(B), and P(A).
- This is very useful in cases where we have a good probability of these three terms and want to determine the fourth one.
- Suppose we want to perceive the effect of some unknown cause, and want to compute that cause, then the Bayes' rule becomes:

$$P(\text{cause} \mid \text{effect}) = \frac{P(\text{effect} \mid \text{cause})\, P(\text{cause})}{P(\text{effect})}$$

Bayes' Theorem Example #1

- Find a patient's <u>probability of having liver disease if they are an alcoholic</u>. "Being an alcoholic" is the test (kind of like a litmus test) for liver disease.
- **A** could mean the event "Patient has liver disease." Past data tells you that 10% of patients entering your clinic have liver disease. **P(A) = 0.10**.
- **B** could mean the litmus test that "Patient is an alcoholic." Five percent of the clinic's patients are alcoholics. **P(B) = 0.05**.
- You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your **B|A**: the probability that a patient is alcoholic, given that they have liver disease, is **7%**.

Bayes' theorem tells you:

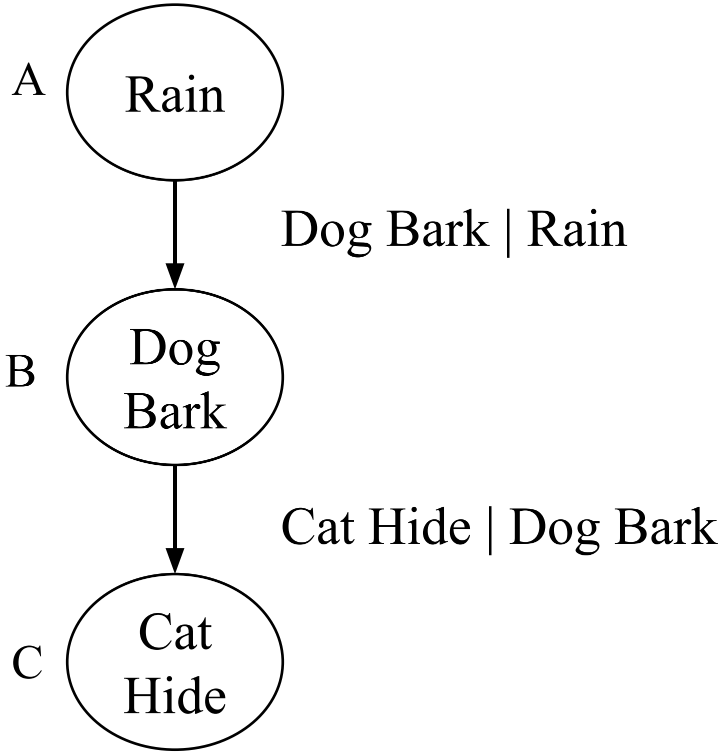
- **P(A|B) = (0.07 * 0.1)/0.05 = 0.14**
- In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is a large increase from the 10% suggested by past data. But it's still unlikely that any particular patient has liver disease.

Bayesian Belief Network in artificial intelligence

- A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph.
- It is also called a **Bayes network, belief network, decision network**, or **Bayesian model**.
- Bayesian networks are probabilistic, because these networks are built from a **probability distribution**, and also use probability theory for prediction and anomaly detection.
- Bayesian Network can be used for building models from data and experts opinions, and it consists of two parts:
    - **Directed Acyclic Graph**
    - **Table of conditional probabilities.**

Bayesian Belief Network

Directed Acyclic Graph (DAG)



If C is conditionally independent of A given B, then we can
Write P(A,B,C) = P(A)P(B|A)P(C|B)

Conditional Probability Table

	R	**~R**
B	9/48	18/48
~B	3/48	18/48

B = T & R = T □ 0.19

B = T & R = F □ 0.375

B = F & R = T □ 0.375

B = F & R = F □ 0.375

Convenient for representing probabilistic relation between multiple events

Example

- **Problem:**

Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry.

A, ~B, ~E, D, S

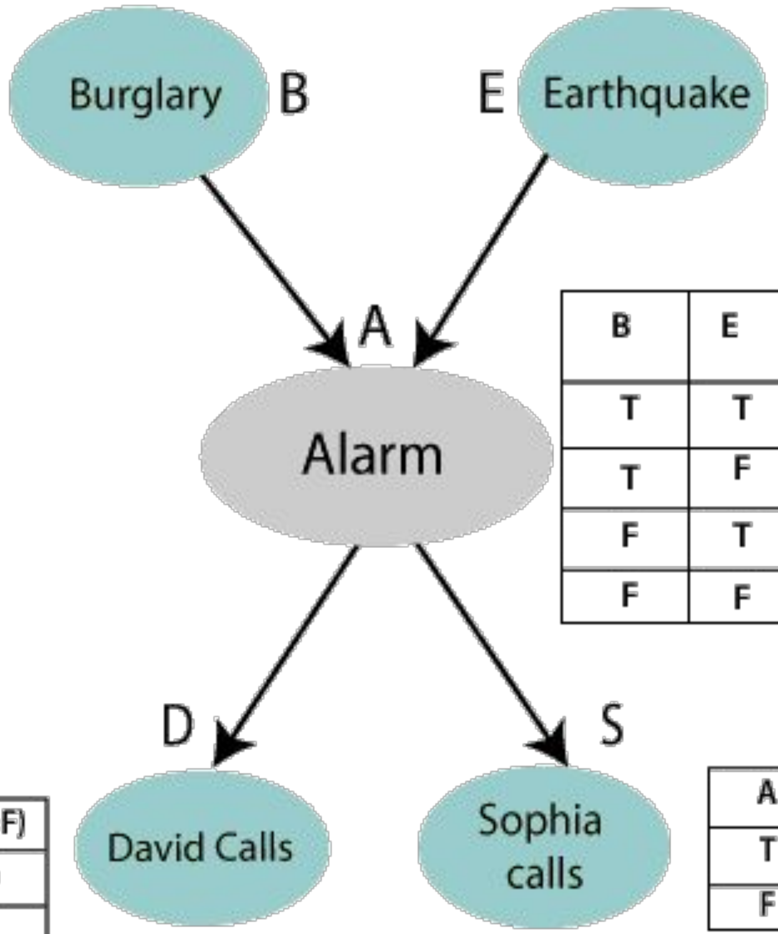
P(S, D, A, ¬B, ¬E)

= P (S|A) *P (D|A)*P (A|¬B , ¬E) *P (¬B) *P (¬E)

= 0.75 * 0.91 * 0.001 * 0.998 * 0.999

= 0.00068045.

T	0.002
F	0.998



T	0.001
F	0.999

B	E	P(A=T)	P(A=F)
T	T	0.94	0.06
T	F	0.95	0.04
F	T	0.69	0.69
F	F	0.999	0.999

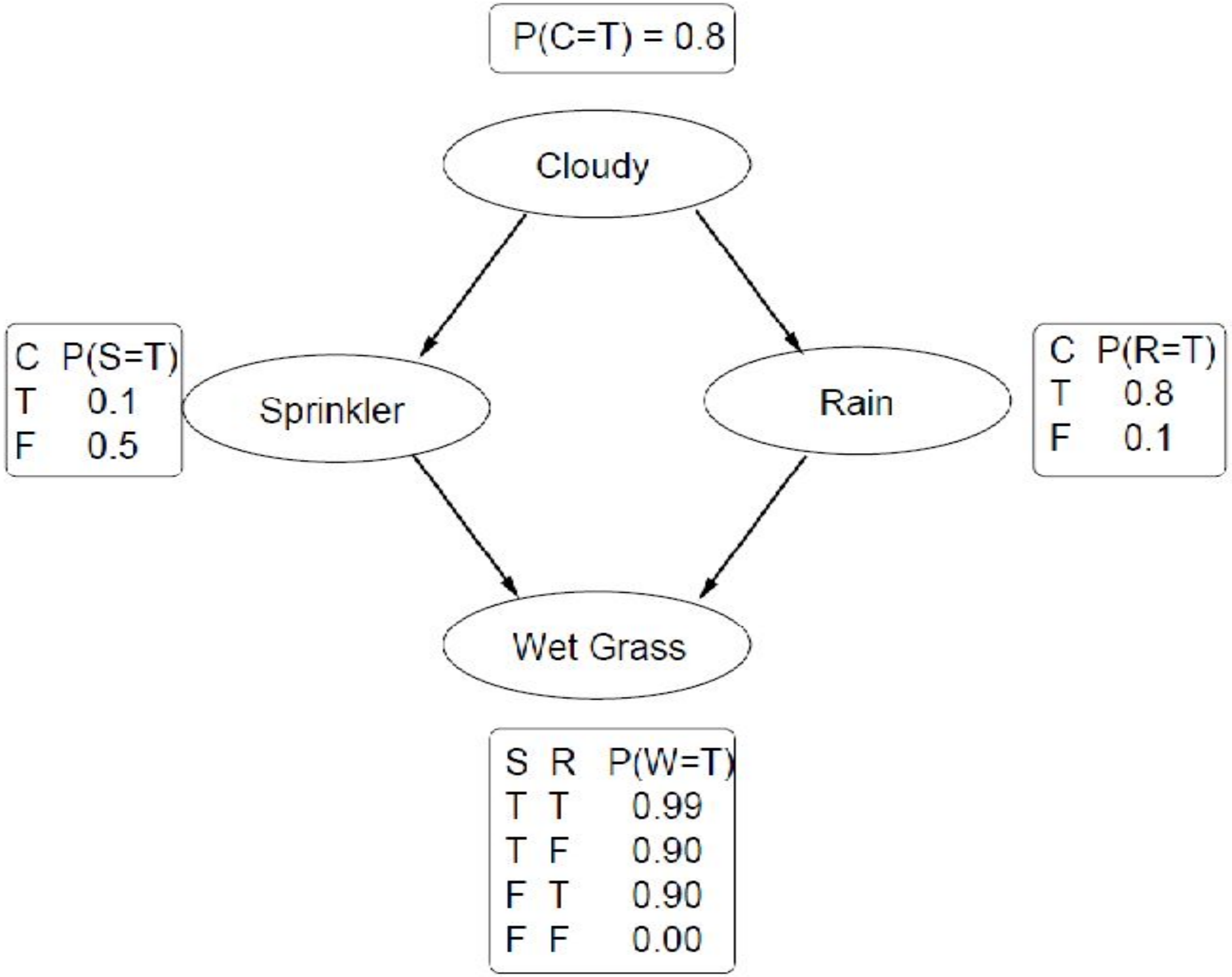
A	P (D=T)	P (D=F)
T	0.91	0.09
F	0.05	0.95

A	P (S=T)	P (S=F)
T	0.75	0.25
F	0.02	0.98

Another Example

- Whether the grass is wet, **W**, depends on whether the sprinkler has been used, **S**, or whether it has rained, **R**.
- Whether the sprinkler is used depends on whether it is cloudy, similarly for whether it has rained.
- The probability of the grass being wet **(W)** is conditionally independent of it being cloudy, given information about the sprinklers and whether it has rained.
- This joint probability may be expressed as

P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S,R)

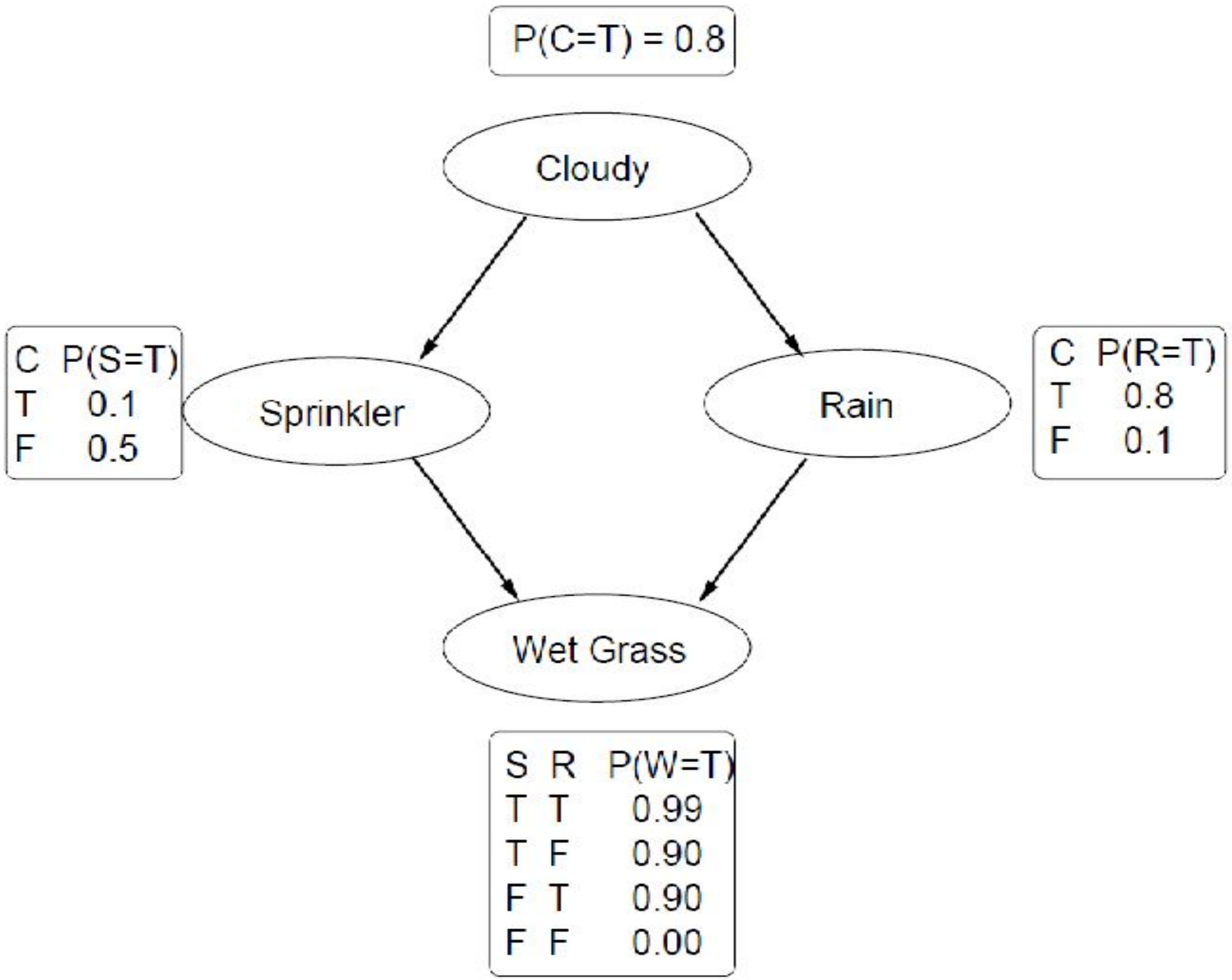


It is cloudy, what's the probability that the grass is wet?

The basic task is given some observation infer the probability of an event. So a question may be

- It is cloudy, what's the probability that the grass is wet?
- so we want to compute P(W = T|C = T). (Note: for simplicity of notation $P(W_T|C_T)$ will be used for P(W = T|C = T).)
- Re-expressing this request in terms of the joint probability

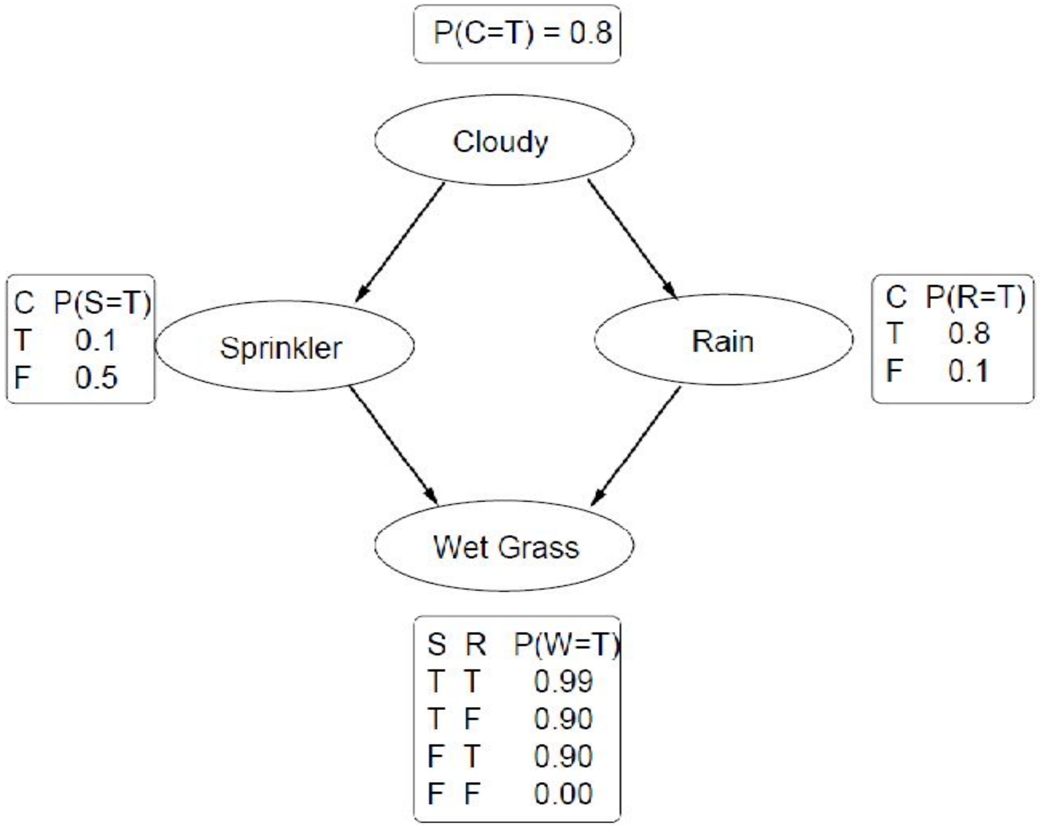
$$P(W_{\mathrm{T}}|C_{\mathrm{T}}) = \frac{P(W_{\mathrm{T}}, C_{\mathrm{T}})}{P(C_{\mathrm{T}})}$$



It is cloudy, what's the probability that the grass is wet?

The denominator is known (0.8). The numerator may expressed as a marginal distribution

$$P(W_\mathrm{T}, C_\mathrm{T}) = \sum_S \sum_R P(W_\mathrm{T}, S, R, C_\mathrm{T})$$
$$= \sum_S \sum_R P(W_\mathrm{T}|S, R) P(S|C_\mathrm{T}) P(R|C_\mathrm{T}) P(C_\mathrm{T})$$



- $P(A, B) = P(A \mid B)\, P(B)$

  (rule of conditional probability)

- $P(W_T, S, R, C_T)$

  $= P(W_T \mid S, R, C_T)\, P(S, R, C_T)$

  $= P(W_T \mid S, R)\, P(S \mid R, C_T)\, P(R, C_T)$

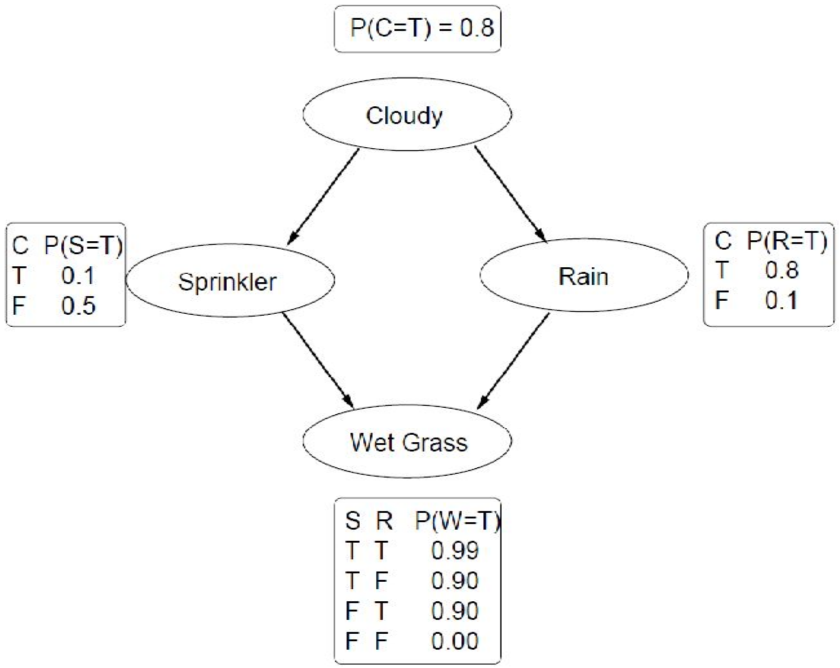
  $= P(W_T \mid S, R)\, P(S \mid C_T)\, P(R \mid C_T)\, P(C_T)$

$$P(W_T \mid S, R)\ P(S \mid C_T)\ P(R \mid C_T)$$

where the summation are over the variable being T, or F.

From the simple example this is (note $P(C_T)$ has simply been cancelled from the numerator and denominator)

$$= P(W_T \mid S_T, R_T)\ P(S_T \mid C_T)\ P(R_T \mid C_T) \qquad (S = T \ \&\ R = T)$$
$$+ P(W_T \mid S_F, R_T)\ P(S_F \mid C_T)\ P(R_T \mid C_T) \qquad (S = F \ \&\ R = T)$$
$$+ P(W_T \mid S_T, R_F)\ P(S_T \mid C_T)\ P(R_F \mid C_T) \qquad (S = T \ \&\ R = F)$$
$$+ P(W_T \mid S_F, R_F)\ P(S_F \mid C_T)\ P(R_F \mid C_T) \qquad (S = F \ \&\ R = F)$$

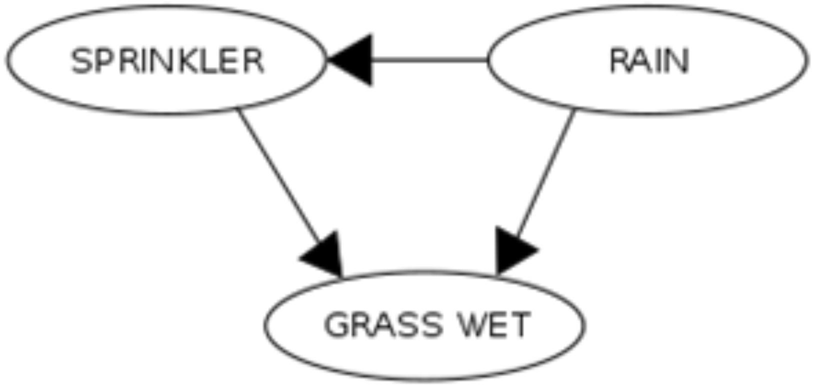


$$\boxed{\begin{aligned} P(W_T|C_T) &= 0.99 \times 0.1 \times 0.8 + 0.90 \times 0.1 \times 0.2 \\ &\quad + 0.90 \times 0.9 \times 0.8 + 0.00 \times 0.9 \times 0.2 \\ &= 0.7452 \end{aligned}}$$

Example

- The grass is wet, what is the probability of rain?

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Dempster-Shafer theory

- The **theory of belief functions**, also referred to as **evidence theory** or **Dempster–Shafer theory (DST)**, is a general framework for reasoning with uncertainty, with understood connections to other frameworks such as probability, possibility and imprecise probability theories.
- First introduced by **Arthur P. Dempster** in the context of statistical inference, the theory was later developed by **Glenn Shafer** into a general framework for modeling epistemic uncertainty—a mathematical theory of evidence.
- The theory allows one to **combine evidence** from different sources and arrive at a **degree of belief** (represented by a mathematical object called **belief function**) that takes into account all the available evidence.
- Dempster–Shafer theory is a generalization of the **Bayesian theory of subjective probability**.

Arthur P. Dempster

Basic

- Beside the probability, there is also a thing called belief.
- That is belief of a person when doing inference.

- **Example:** Suppose a person come to you with a coin and says that if you toss the coin then 90% of time it will be head.
- Now you do not **believe** that character.
- **X** is the event of coin to be head then my belief **Bel(X) = 0** and **Bel(not X) = 0** because I don't have any evidence to support the fact.
- Now, a very reliable person in which my confidence is 90%, says that the coin is fair.
- So my **Bel(X) = 0.5 * 0.9 = 0.45** and **Bel(not X) = 0.5 * 0.9 = 0.45**
- Total probability is **0.90** (**0.1** less than total probability of **1**)
- That show my **lack of confidence** on the person.

- **Probability theory limitation**
  - Assign a single number to measure any situation, no matter how it is complex
  - Cannot deal with missing evidence, heuristics, and limited knowledge
- **Dempster-Shafer theory**
  - Extend probability theory
  - Consider a set of propositions as a whole
  - Assign a set of propositions an interval [believe, plausibility] to constraint the degree of belief for each individual propositions in the set
  - The belief measure bel is in [0,1]
    - 0 – no support evidence for a set of propositions
    - 1 – full support evidence for a set of propositions
  - The plausibility of p,
- **pl(p) = 1 – bel(¬ (p))**
    - Reflect how evidence of ¬ (p) relates to the possibility for belief in p
    - Bel(not(p))=1: full support for not(p), no possibility for p
    - Bel(not(p))=0: no support for not(p), full possibility for p
    - Range is also in [0,1]

Need for Dempster-Shafer theory

- 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.
- DST is an evidence theory, it combines all possible outcomes of the problem. Hence it is used to solve problems where there may be a chance that different evidence will lead to some different result.
- For example, Let A represent the proposition "Moore is attractive". Then the axioms of probability insist that $P(A) + P(\neg A) = 1$.
- 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.

Planning

Planning in AI

- The planning in Artificial Intelligence is about the decision making tasks performed by the robots or computer programs to achieve a specific goal.
- The execution of planning is about choosing a sequence of actions with a high likelihood to complete the specific task.

Blocks-World planning problem

- The blocks-world problem is known as Sussman Anomaly.
- Noninterleaved planners of the early 1970s were unable to solve this problem, hence it is considered as anomalous.
- When two subgoals G1 and G2 are given, a noninterleaved planner produces either a plan for G1 concatenated with a plan for G2, or vice-versa.
- In blocks-world problem, three blocks labeled as 'A', 'B', 'C' are allowed to rest on the flat surface. The given condition is that only one block can be moved at a time to achieve the goal

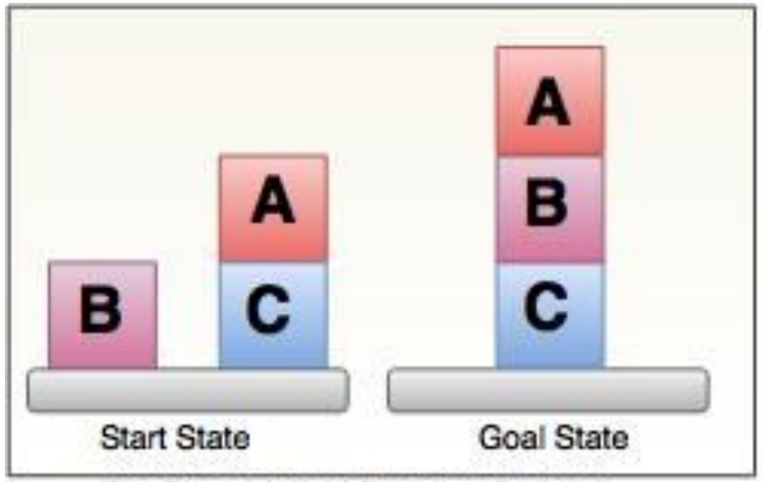


Fig: Blocks-World Planning Problem

Components of Planning System

The planning consists of following important steps:

- Choose the best rule for applying the next rule based on the best available heuristics.
- Apply the chosen rule for computing the new problem state.
- Detect when a solution has been found.
- Detect dead ends so that they can be abandoned and the system's effort is directed in more fruitful directions.
- Detect when an almost correct solution has been found.

AI Planners

- Well-known AI Planners:
  - STRIPS (Fikes and Nilsson, 1971): theorem proving system
  - ABSTRIPS (Sacerdoti, 1974): added hierarchy of abstractions
  - HACKER (Sussman, 1975): use library of procedures to plan
  - NOAH (Sacerdoti, 1975): problem decomposition and plan reordering

Goal stack planning

- This is one of the most important planning algorithms, which is specifically used by STRIPS.
- The stack is used in an algorithm to hold the action and satisfy the goal. A knowledge base is used to hold the current state, actions.
- Goal stack is similar to a node in a search tree, where the branches are created if there is a choice of an action.

- The important steps of the algorithm are as stated below:

i. Start by pushing the original goal on the stack. Repeat this until the stack becomes empty. If stack top is a compound goal, then push its unsatisfied subgoals on the stack.

ii. If stack top is a single unsatisfied goal then, replace it by an action and push the action's precondition on the stack to satisfy the condition.

iii. If stack top is an action, pop it from the stack, execute it and change the knowledge base by the effects of the action.

iv. If stack top is a satisfied goal, pop it from the stack.

Partial-Order Planning

Idea:

- works on several subgoals independently
- solves them with subplans
- combines the subplans
- flexibility in ordering the subplans
- least commitment strategy:
    - delaying a choice during search
    - Example, leave actions unordered, unless they must be sequential

Hierarchical Planning

- Hierarchical planning is a planning method based on Hierarchical Task Network (HTN) or HTN planning.
- It combines ideas from Partial Order Planning & HTN Planning.
- HTN planning is often formulated with a single "top level" action called Act, where the aim is to find an implementation of Act that achieves the goal.
- In HTN planning, the initial plan is viewed as a very high level description of what is to be done.
- This plan is refined by applying decomposition actions.
- Each action decomposition reduces a higher level action to a partially ordered set of lower-level actions.
- This decomposition continues until only the primitive actions remain in the plan.

Advantages of Hierarchical Planning:

- The key benefit of hierarchical structure is that, at each level of the hierarchy, plan is reduced to a small number of activities at the next lower level, so the computational cost of finding the correct way to arrange those activities for the current problem is small.
- HTN methods can create the very large plans required by many real-world applications.
- Hierarchical structure makes it easy to fix problems in case things go wrong.
- For complex problems hierarchical planning is much more efficient than single level planning.

Non-linear planning

- This planning is used to set a goal stack and is included in the search space of all possible subgoal orderings. It handles the goal interactions by interleaving method.
- **Advantage of non-Linear planning**

Non-linear planning may be an optimal solution with respect to plan length (depending on search strategy used).

- **Disadvantages of Nonlinear planning**

1. It takes larger search space, since all possible goal orderings are taken into consideration.
2. Complex algorithm to understand.

Algorithm

1. Choose a goal 'g' from the goalset

2. If 'g' does not match the state, then

   Choose an operator 'o' whose add-list matches goal g

   Push 'o' on the opstack

   Add the preconditions of 'o' to the goalset

3. While all preconditions of operator on top of opstack are met in state

   Pop operator o from top of opstack

   state = apply(o, state)

   plan = [plan; o]