CPSC 322: Introduction to Artificial Intelligence

Reasoning under Uncertainty: Introduction to Probability

Textbook reference: [8.1]

Instructor: Varada Kolhatkar
University of British Columbia

Credit: These slides are adapted from the slides of the previous offerings of the course. Thanks to all instructors for creating and improving the teaching material and making it available!

Announcements

- Assignment 4 has been released.
    - Due date: **Nov 29th, 11:59 PM**
- Final exam scheduled: **Dec 9 at 7:00pm**

A rough CPSC 322 overview

Representation and reasoning

Environment

Problem

		Deterministic	Stochastic
Static	Constraint satisfaction	Variables + constraints Arc consistency Search	
	Query	Logics Search	Belief networks Variable elimination
Sequential	Planning	STRIPS Search	Decision networks Variable elimination Markov decision processes Value iteration

Lecture outline

- Random variables and possible world semantics
- Probability distributions and marginalization
- Conditional probability
- Product Rule, chain Rule, Bayes Rule
- Class activity

Today: Learning outcomes

From this lecture, students are expected to be able to:

- Define and give examples of random variables, their domains and probability distributions.
- Calculate the probability of a proposition f given $\mu(w)$ for the set of possible worlds.
- Define a joint probability distribution.
- Prove the formula to compute conditional probability $P(h \mid e)$
- Derive and use Bayes Rule
- Derive and use Chain Rule and Product Rule

Introduction to probability (Motivation)

To act in the real world, we almost always have to handle **uncertainty**.

Two main sources of uncertainty

Sensing Uncertainty: The agent cannot fully observe a state of interest. For example:

- Right now, how many people are in this room? In this building?
- What disease does this patient have?

Effect Uncertainty: The agent cannot be certain about the effects of its actions. For example:

- If I work hard, will I get an A+?
- Will this drug work for this patient?

Introduction to probability (Motivation)

To act in the real world, we almost always have to handle uncertainty (both effect and sensing uncertainty)

- Deterministic domains are an abstraction
  - Sometimes this abstraction enables more powerful inference
- Now we don't make this abstraction anymore
  - Our representation becomes more expressive and general

Introduction to probability (Motivation)

- AI main focus shifted from logic to probability in the 1980s
- The language of probability is very **expressive** and **general**
- New representations enable **efficient reasoning**
  - We will see some of these, in particular **Bayesian networks**
- **Reasoning under uncertainty** is part of the 'new' AI
- This is **not a dichotomy**: framework for probability is logical!
- New frontier: combine logic and probability

Introduction to probability (Motivation)

"Dealing with uncertainty turned out to be more important than thinking with logical precision. We think of a clever argument or solution to a problem as one that contains a series of irrefutable logical steps and are impressed when someone can come up with such a sequence. But this is exactly what computers do well. The hard part is dealing with uncertainty, and choosing a good answer from among many possibilities. The fundamental tools of A.I. shifted from Logic to Probability in the late 1980s, and fundamental progress in the theory of uncertain reasoning underlies many of the recent practical advances."…

"**Reasoning under uncertainty (and lots of data) are key to progress** "

–Peter Norvig

Source: https://nypost.com/2011/02/13/the-machine-age/

Probability as a formal measure of uncertainty (ignorance)

Probability measures **an agent's degree of belief** in propositions about states of the world.

- It does not measure how true a proposition is.
- Propositions are true or false. We simply may not know exactly which.
- Belief in a proposition f can be measured in terms of a number between 0 and 1 — this is the probability of f.
- P("roll of fair die came out as a 6") = 1/6 ≈ 16.7% = 0.167
- Using probabilities between 0 and 1 is purely a convention.

Probability as a formal measure of uncertainty (ignorance)

Probability measures **an agent's degree of belief** in propositions about states of the world.

<u>Examples</u>:

I roll a fair dice. What is the probability that the result is a '6'?

- It is 1/6 = 16.7%.
- The result is either 6 or not but I don't know which one.

I now look at the dice. What is 'the' (my) probability now?

- My probability is now either 1 or 0, depending on what I observed.
- Your probability hasn't changed: $1/6 \approx 16.7\%$

What if I tell some of you the result is even?

- Their probability increases to $1/3 \approx 33.3\%$ (assuming they believe I speak the truth)

Different agents can have different degrees of belief in (probabilities for) a proposition conditioned on the evidence they have.

Probability as a formal measure of uncertainty (ignorance)

iclicker

Belief in a proposition f can be measured in terms of a number between 0 and 1 — this is the probability of f.

$P(f) = 0$ means that f is believed to be

A. Probably true

B. Probably false

C. Definitely false ✅

D. Definitely true

Probability theory and random variables

- **Probability Theory**: system of **logical** axioms and formal operations for sound reasoning under uncertainty
- **Basic element:** random variable X
  - X is a variable like the ones we have seen in CSP/ Planning/Logic, but the agent can be uncertain about the value of X
  - As usual, the domain of a random variable X, written dom(X), is the set of values X can take

Random variables

Types of variables

- Boolean:
  E.g., Cancer (does the patient have cancer or not?)

- Categorical:
  E.g., CancerType could be one of
  {breastCancer, lungCancer, skinMelanomas}

- Numeric:
  E.g., Temperature (integer or real)

- We will focus on **Boolean** and **categorical** variables

Random variables

A tuple of random variables $< X_1, X_2, \ldots, X_n >$ is a complex random variable with domain $dom(X_1) \times dom(X_2) \times \ldots \times dom(X_n)$

Assignment: $X = x$ means X has value x

A **proposition** is a Boolean formula made from assignments of values to variables
Example: $raining\_outside = T \wedge \#people\_in\_room = 30$

Possible worlds semantics

A possible world w specifies an assignment to each random variable.

<u>Example</u>: If we model only 2 Boolean variables *Smoking* and *Cancer*, there are $2^2 = 4$ possible worlds.

$w_1 : smoking = T \wedge Cancer = T$
$w_2 : smoking = T \wedge Cancer = F$
$w_3 : smoking = F \wedge Cancer = T$
$w_4 : smoking = F \wedge Cancer = F$

Smoking	*Cancer*
T	T
T	F
F	T
F	F

$w_4 \models smoking = F$

$w_2 \not\models Cancer = T$

$w \models X = x$ means variable X is assigned value x in world w.

(Related but not identical to its meaning in logic)

Semantics of probability

- The belief of being in each possible world w can be expressed as probability $P = \mu(w)$.

- For sure I must be in one of them and so

$$\sum_{w \in W} \mu(w) = 1, \text{ where } W \text{ is the set of all possible worlds}$$

<u>Example</u>: Vancouver weather modelled as one categorical variable with domain: $\{sunny, cloudy\}$

$w_1 : Weather = sunny$

$w_2 : Weather = cloudy$

Weather	*P*
sunny	0.4
cloudy	?

Semantics of probability

Now we have an additional variable: Temperature, modelled as a categorical variable with domain {hot, mild, cold}

There are now 6 possible worlds

What's the probability of it being sunny and cold?

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	?
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Probability of a proposition

The probability of proposition f is defined as:

$$P(f) = \sum_{w \models f} \mu(w)$$

Example 1: $f : T = cold$

only $w_3 \models f$ and $w_6 \models f$

So $P(f) = \mu(w_3) + \mu(w_6)$

= 0.20 + 0.10 = 0.30

Example 2:

$g : W = sunny \wedge T = cold$

only $w_3 \models g$

So $P(g) = \mu(w_3) = 0.10$

Possible world	Weather (W)	Temperature (T)	μ(w)
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
w_4	cloudy	hot	0.05
w_5	cloudy	mild	0.35
w_6	cloudy	cold	0.20

Lecture outline

- Random variables and possible world semantics
- Probability distributions and marginalization 👈🏻
- Conditional probability
- Product Rule, chain Rule, Bayes Rule
- Class activity

Probability distributions and marginalization

Consider the case where possible worlds are simply assignments to one random variable.

Definition: **A probability distribution** P on a random variable X is a function from $dom(X) \rightarrow [0,1]$ such that $P(x)$ is the probability of the proposition $X = x$.

When $dom(X)$ is infinite, we need a probability density function. In this class, we will focus on the finite case.

Note: we use the notations $P(f)$ and $p(f)$ interchangeably.

Joint probability distribution (JPD)

The joint distribution over random variables $X_1, X_2, \ldots, X_n$ is a probability distribution over the joint random variable with domain $dom(X_1) \times dom(X_2) \times \ldots \times dom(X_n)$ (the Cartesian product)

The table shows a joint probability distribution over random variables *Weather* and Temperature.

Each row corresponds to an assignment of values to these variables, and the probability of this joint assignment.

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Joint probability distribution (JPD)

- In general, each row corresponds to an assignment $X_1 = x_1, \ldots, X_n = x_n$ and its probability $P(X_1 = x_1, \ldots, X_n = x_n)$.

- We also write $P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n)$.

- The sum of probabilities across the whole table is 1.

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Marginalization

- Suppose you have the joint probability distribution of n variables.
- Can you compute the probability distribution for each variable?
- Can you compute the probability distribution for any combination of variables?

Marginalization

Given the joint distribution, we can compute distributions over smaller sets of variables through marginalization:

$$P(X = x) = \sum_{z \in dom(Z)} P(X = x, Z = z)$$

This corresponds to summing out a dimension in the table.

The new table still sums to 1. It must, since it's a probability distribution!

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Weather	μ(w)
sunny	
cloudy	

Marginalization

Given the joint distribution, we can compute distributions over smaller sets of variables through marginalization:

$$P(X = x) = \sum_{z \in dom(Z)} P(X = x, Z = z)$$

This corresponds to summing out a dimension in the table.

The new table still sums to 1. It must, since it's a probability distribution!

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Weather	μ(w)
sunny	0.4
cloudy	0.6

P(Weather=sunny)
= P(Weather=sunny, Temperature = hot) + P(Weather=sunny, Temperature = mild) + P(Weather = sunny, Temperature = cold)
= 0.10 + 0.20 + 0.10 = 0.40

Marginalization (pair-share)

Given the joint distribution, we can compute distributions over smaller sets of variables through marginalization:

$$P(X = x) = \sum_{z \in dom(Z)} P(X = x, Z = z)$$

This corresponds to summing out a dimension in the table.

The new table still sums to 1. It must, since it's a probability distribution!

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Temperature	μ(w)
hot	?
mild	?
cold	?

Marginalization

Given the joint distribution, we can compute distributions over smaller sets of variables through marginalization:

$$P(X = x) = \sum_{z \in dom(Z)} P(X = x, Z = z)$$

This corresponds to summing out a dimension in the table.

The new table still sums to 1. It must, since it's a probability distribution!

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Temperature	μ(w)
hot	0.15
mild	0.55
cold	0.30

Marginalization

We can also get marginals for more than one variable.

$$P(X = x, Y = y) = \sum_{z_1 \in dom(Z_1), \ldots, z_n \in dom Z_n} P(X = x, Y = y, Z_1 = z_1, \ldots, Z_n = z_n)$$

Wind	Weather	Temperature	μ(w)
yes	sunny	hot	0.04
yes	sunny	mild	0.09
yes	sunny	cold	0.07
yes	cloudy	hot	0.01
yes	cloudy	mild	0.10
yes	cloudy	cold	0.12
no	sunny	hot	0.06
no	sunny	mild	0.11
no	sunny	cold	0.03
no	cloudy	hot	0.04
no	cloudy	mild	0.25
no	cloudy	cold	0.08

Weather	Temperature	μ(w)
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Lecture outline

- Random variables and possible world semantics
- Probability distributions and marginalization
- Conditional probability 👈
- Product Rule, chain Rule, Bayes Rule
- Class activity

Conditional probability: Motivation

- We model our environment with a set of random variables.
- Assuming we have the JPD, we can compute the probability of any formula.
- Are we done with reasoning under uncertainty?
- What can happen?
- Think of a patient showing up at the doctor's office. We are interested in knowing whether she has a disease.

Conditioning

- **Probabilistic conditioning** specifies how to **revise beliefs based on new information**.
- You build a probabilistic model (JPD) taking all background information into account. This gives the **prior probability**, $P(h)$, for the hypothesis.
- Observe new information about the world. Call all information we received subsequently the **evidence** e.
- Integrate the two sources of information to compute the conditional probability, $P(h \mid e)$. This is called the **posterior probability** of h given e.

Conditioning: Example

- **Prior probability** for having a disease (typically small)
- **Evidence**: a test for the disease comes out positive
  - But diagnostic tests have false positives
- **Posterior probability**: integrate prior and evidence

Conditioning: Example

You have a prior for the joint distribution of weather and temperature.

Possible world	Weather (W)	Temperature (T)	μ(w)
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
w_4	cloudy	hot	0.05
w_5	cloudy	mild	0.35
w_6	cloudy	cold	0.20

Conditioning: Example

You have a prior for the joint distribution of weather and temperature

Possible world	Weather (W)	Temperature (T)	μ(w)
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
~~w_4~~	~~cloudy~~	~~hot~~	~~0.05~~
~~w_5~~	~~cloudy~~	~~mild~~	~~0.35~~
~~w_6~~	~~cloudy~~	~~cold~~	~~0.20~~

What happens in terms of possible worlds if we know the value of a random variable (or a set of random variables)?

Some worlds are **ruled out** and others **become more likely**

Now you look outside and see that it's **sunny**. You are now certain that you are in one of the worlds w_1, w_2, w_3.

Conditioning: Example

You have a prior for the joint distribution of weather and temperature

Possible world	Weather (W)	Temperature (T)	$\mu(w)$
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
~~w_4~~	~~cloudy~~	~~hot~~	~~0.05~~
~~w_5~~	~~cloudy~~	~~mild~~	~~0.35~~
~~w_6~~	~~cloudy~~	~~cold~~	~~0.20~~

T	P(T|W = sunny)
hot	0.10/0.4 = 0.25
mild	?
cold	?

To get the conditional probability you simply renormalize to sum to 1

Now you look outside and see that it's **sunny**. You are now certain that you are in one of the worlds w_1, w_2, w_3.

Conditioning: Example

You have a prior for the joint distribution of weather and temperature

Possible world	Weather (W)	Temperature (T)	μ(w)
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
~~w_4~~	~~cloudy~~	~~hot~~	~~0.05~~
~~w_5~~	~~cloudy~~	~~mild~~	~~0.35~~
~~w_6~~	~~cloudy~~	~~cold~~	~~0.20~~

T	P(T|W = sunny)
hot	0.10/0.4 = 0.25
mild	0.20/0.4 = 0.50
cold	0.10/0.4 = 0.25

To get the conditional probability you simply renormalize to sum to 1

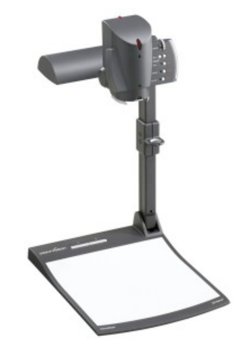
Now you look outside and see that it's **sunny**. You are now certain that you are in one of the worlds w_1, w_2, w_3.

Semantics of conditioning

Evidence e ("W=sunny") rules out possible worlds incompatible with e. Now we formalize what we did in the previous example.

Possible world	Weather (W)	Temperature (T)	$\mu(w)$	$\mu_e(w)$
w_1	sunny	hot	0.10	
w_2	sunny	mild	0.20	
w_3	sunny	cold	0.10	
~~w_4~~	~~cloudy~~	~~hot~~	~~0.05~~	
~~w_5~~	~~cloudy~~	~~mild~~	~~0.35~~	
~~w_6~~	~~cloudy~~	~~cold~~	~~0.20~~	

What is $P(e)$?



We represent the updated probability using a new measure, μ_e, over possible worlds.

$$\mu_e(w) = \begin{cases} \frac{1}{P(e)} \times \mu(w), & \textbf{if } w \models e \\ 0, & \textbf{if } w \not\models e \end{cases}$$

Semantics of conditioning

Evidence e ("W=sunny") rules out possible worlds incompatible with e. Now we formalize what we did in the previous example.

Possible world	Weather (W)	Temperature (T)	μ(w)	$\mu_e(w)$
w_1	sunny	hot	0.10	0.10/0.40 = 0.25
w_2	sunny	mild	0.20	0.20/0.40 = 0.50
w_3	sunny	cold	0.10	0.10/0.40 = 0.25
~~w_4~~	~~cloudy~~	~~hot~~	~~0.05~~	0
~~w_5~~	~~cloudy~~	~~mild~~	~~0.35~~	0
~~w_6~~	~~cloudy~~	~~cold~~	~~0.20~~	0

What is $P(e)$?

Marginalize out temperature, i.e.,
0.10 + 0.20 + 0.10 = 0.40

We represent the updated probability using a new measure, μ_e, over possible worlds.

$$\mu_e(w) = \begin{cases} \frac{1}{P(e)} \times \mu(w), & \textbf{if } w \models e \\ 0, & \textbf{if } w \not\models e \end{cases}$$

Conditional probability

$P(e)$: Sum of probability of all worlds in which e is true

$P(e \wedge h)$: Sum of probability of all worlds in which both h and e are true

$P(h \mid e) = \dfrac{p(h \wedge e)}{P(e)}$, only defined when $P(e) > 0$

$$\mu_e(w) = \begin{cases} \frac{1}{P(e)} \times \mu(w), & \text{if } w \models e \\ 0, & \text{if } w \not\models e \end{cases}$$

Conditional probability: The conditional probability formula h given evidence e is

$$P(h \mid e) = \sum_{w \models h} \mu_e(w) = \frac{1}{P(e)} \sum_{w \models h \wedge e} \mu(w) = \frac{P(h \wedge e)}{P(e)}$$

Lecture outline

- Random variables and possible world semantics
- Probability distributions and marginalization
- Conditional probability
- Product Rule, chain Rule, Bayes Rule 👈
- Class activity

Product rule

By definition, we know that $P(f_2 | f_1) = \dfrac{P(f_2 \wedge f_1)}{P(f_1)}$

We can rewrite this as: $P(f_2 \wedge f_1) = p(f_2 | f_1) \times P(f_1)$

In general,

Product Rule:
$P(f_n \wedge \ldots \wedge f_{i+1} \wedge f_i \wedge \ldots \wedge f_1)$
$= P(f_n \wedge \ldots \wedge f_{i+1} | f_i \wedge \ldots \wedge f_1) \times P(f_i \wedge \ldots \wedge f_1)$

Chain rule

By definition, we know that $P(f_2 \wedge f_1) = P(f_2 | f_1) \times P(f_1)$

In general,

$$P(f_n \wedge f_{n-1} \wedge \ldots \wedge f_1) = P(f_n | f_{n-1} \wedge \ldots \wedge f_1) \times P(f_{n-1} \wedge \ldots \wedge f_1)$$

$$= P(f_n | f_{n-1} \wedge \ldots \wedge f_1) \times P(f_{n-1} | f_{n-2} \wedge \ldots \wedge f_1) \times P(f_{n-2} \wedge \ldots \wedge f_1)$$

$$= \ldots = \prod_{i=1}^{n} P(f_i | f_{i-1} \wedge \ldots \wedge f_1)$$

Chain rule:

$$P(f_n \wedge f_{n-1} \wedge \ldots \wedge f_1) = \prod_{i=1}^{n} P(f_i | f_{i-1} \wedge \ldots \wedge f_1)$$

Why does chain rule help us?

- We can simplify some terms. For example, how about P(Weather | PriceOfMacbook)?
- Weather in Vancouver is independent of the price of oil:
- P(Weather | PriceOfMacbook) = P(Weather)
- Under independence, we gain compactness
- We can represent the JPD as a product of marginal distributions E.g., P(Weather, PriceOfMacbook) = P(Weather) x P(PriceOfMacbook)
- But not all variables are independent:
- P(Weather | Temperature) ≠ P(Weather)
- More about (conditional) independence later

Bayes rule

Often you have **causal knowledge** (forward from cause to evidence): P(evidence e | hypothesis h)
Examples:
P(symptom | disease)
P(alarm | fire)

… and you want to do **evidential reasoning** (backwards from evidence to cause): P(hypothesis h | evidence e)
Examples:
P(disease | symptom)
P(fire | alarm)

Bayes rule

1. By definition, we know that: $P(h\,|\,e) = \dfrac{P(h \wedge e)}{P(e)}$ and $P(e\,|\,h) = \dfrac{P(e \wedge h)}{P(h)}$

2. We can rearrange these terms and write:
   $P(h \wedge e) = P(h\,|\,e)P(e)$
   $P(e \wedge h) = P(e\,|\,h)P(h)$

3. But $P(h \wedge e) = P(e \wedge h)$

From 1., 2., 3. we can derive:

Bayes Rule: $P(h\,|\,e) = \dfrac{P(e\,|\,h)P(h)}{p(e)}$

Bayes rule: example

On average the alarm rings once a year: $P(alarm) = 1/365$

If there is a fire, the alarm will almost always ring:
$P(alarm \mid fire) = 0.999$

On average we have a fire every 10 years: $P(fire) = 1/3650$

The fire alarm rings. What is the probability there is fire?
$P(fire \mid alarm)?$

A. 0.9

B. 0.999

C. 0.0999 ✅

D. 0.01

Important note

Marginalization, conditioning and Bayes rule are crucial

They are core to reasoning under uncertainty

Be sure you understand them and be able to use them!

Lecture outline

- Random variables and possible world semantics
- Probability distributions and marginalization
- Conditional probability
- Product Rule, chain Rule, Bayes Rule
- Class activity 👈

Class activity (~15 mins)

World	Cavity	Toothache	Catch	$\mu(w)$
w_1	T	T	T	0.108
w_2	T	T	F	0.012
w_3	T	F	T	0.072
w_4	T	F	F	0.008
w_5	F	T	T	0.016
w_6	F	T	F	0.064
w_7	F	F	T	0.144
w_8	F	F	F	0.576

Coming up

Introduction to probability

8.2 Independence

8.3 Belief Networks