Probability

Yuntian Deng

Lecture 6

Readings: RN 12.2 - 12.3. PM 8.1.

Outline

Learning Goals

- Calculate prior, posterior, and joint probabilities using the sum rule, the product rule, the chain rule and Bayes' rule.

Why handle uncertainty?

Why does an agent need to handle uncertainty?

- An agent may not observe everything in the world. Does not know what state it is in.
- An action may not have its intended consequences. Does not know what state it will be in after a sequence of actions.

An agent needs to

- Reason about their uncertainty.
- Make a decision based on their uncertainty.

Why handle uncertainty?

Why does an agent need to handle uncertainty?

- An agent may not observe everything in the world. Does not know what state it is in.
- An action may not have its intended consequences. Does not know what state it will be in after a sequence of actions.

An agent needs to

- Reason about their uncertainty.
- Make a decision based on their uncertainty.

→ An agent does not know everything, but needs to make a decision anyway.

Decisions are made in the absence of information or in the presence of noisy information.

Best it can do is know how uncertain it is and act accordingly.

Probability

- Probability is the formal measure of uncertainty.
- There are two camps: *Frequentists* and *Bayesians*.

Probability

- Probability is the formal measure of uncertainty.
- There are two camps: *Frequentists* and *Bayesians*.
- **Frequentists' view of probability:**
  - Frequentists view probability as something *objective*.
  - Compute probabilities by counting the frequencies of events.

$\rightarrow$ Prob of heads for this coin = prob of heads in history.

Cannot make decision without observation

Probability

- Probability is the formal measure of uncertainty.
- There are two camps: *Frequentists* and *Bayesians*.
- **Frequentists' view of probability:**
  - Frequentists view probability as something *objective*.
  - Compute probabilities by counting the frequencies of events.

  $\rightarrow$ Prob of heads for this coin = prob of heads in history.
  Cannot make decision without observation
- **Bayesians' view of probability:**
  - Bayesians view probability as something *subjective*.
  - Probabilities are degrees of belief.
  - We start with **prior** beliefs and **update** beliefs based on new evidence.

  $\rightarrow$ Prob of heads for this coin = prob of heads in agent's previous experience. Different agents may have different beliefs. With no data, can make decision based on uninformed prior.

Frequentists vs. Bayesian Example:

- Bayesians assume that the initial probability of seeing coin head is $\frac{1}{1+1}$. Whenever it sees n heads and m tail, Bayesians will update the belief to $\frac{1+n}{1+n+1+m}$.
- Frequentists only believe observation, therefore, the belief is always $\frac{n}{n+m}$.
- After 2 heads and 4 tails
- Bayesians' belief of coin head is $\frac{3}{8}$.
- Frequentists' belief of coin head is $\frac{2}{6}$.

Real-life example

Your degree of belief that a bird can fly is your measure of belief in the flying ability of an individual based only on the knowledge that the individual is a bird.

Other agents may have different probabilities, as they may have had different experiences with birds or different knowledge about this particular bird.

An agent's belief in a bird's flying ability is affected by what the agent knows about that bird.

An agent can update its beliefs as it receives evidence. For example, if the agent sees a penguin for the first time, it may initially believe that the penguin can fly!

Random variable

A **random variable**

- Has a **domain** of possible values
- Has an associated **probability distribution**, which is a function from the domain of the random variable to $[0, 1]$.

Example:

- random variable: The alarm is sounding.
- domain: $\{$True, False$\}$
- P(The alarm is sounding = True) = 0.1
  P(The alarm is sounding = False) = 0.9

Shorthand notation for Boolean random variables

Let A be a Boolean random variable.

- $P(A)$ denotes $P(A = true)$.
- $P(\neg A)$ denotes $P(A = false)$.

Axioms of Probability

Let A and B be Boolean random variables.

- Every probability is between 0 and 1.

$$0 \leq P(A) \leq 1$$

- Necessarily true propositions have probability 1.
  Necessarily false propositions have probability 0.

$$P(true) = 1, P(false) = 0$$

- The inclusion-exclusion principle:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

These axioms limit the functions that can be considered as probability functions.

Axioms of Probability

Note:

- Probability between 0-1 is purely convention.
- $0 < P(a) < 1$ means you have belief about the truth of a. It does not meant that a is true to some degree, just that you are ignorant of its truth value. Probability = measure of ignorance.

Joint Probability Distribution

- A **probabilistic model** contains a set of random variables.
- An **atomic event** assigns a value to every random variable in the model.
- A **joint probability distribution** assigns a probability to every atomic event.

Example of Joint Distribution

Consider the weather and temperature of each day.

Two random variables:

- $weather$, with domain {Sunny, Cloudy};
- $temperature$, with domain {Hot, Mild, Cold}.

The joint distribution $P(weather, temperature)$:

	Hot	Mild	Cold
Sunny	0.10	0.20	0.10
Cloudy	0.05	0.35	0.20

Prior and Posterior Probabilities

$P(X)$:

- **prior** or **unconditional** probability
- Likelihood of X in the absence of any other information
- Based on the background information

$P(X|Y)$

- **posterior** or **conditional** probability
- Likelihood of X given Y.
- Based on Y as evidence

The Holmes Scenario

Mr. Holmes lives in a high crime area and therefore has installed a burglar alarm. He relies on his neighbours to phone him when they hear the alarm sound. Mr. Holmes has two neighbours, Dr. Watson and Mrs. Gibbon.

Unfortunately, his neighbours are not entirely reliable. Dr. Watson is known to be a tasteless practical joker and Mrs. Gibbon, while more reliable in general, has occasional drinking problems.

Mr. Holmes also knows from reading the instruction manual of his alarm system that the device is sensitive to earthquakes and can be triggered by one accidentally. He realizes that if an earthquake has occurred, it would surely be on the radio news.

Modelling the Holmes Scenario

What are the random variables?

Modelling the Holmes Scenario

What are the random variables?

$\rightarrow$

- B: A Burglary is happening.
- A: The alarm is going.
- W: Dr. Watson is calling.
- G: Mrs. Gibbon is calling.
- E: Earthquake is happening.
- R: A report of earthquake is on the radio news.

Modelling the Holmes Scenario

What are the random variables?

→

- B: A Burglary is happening.
- A: The alarm is going.
- W: Dr. Watson is calling.
- G: Mrs. Gibbon is calling.
- E: Earthquake is happening.
- R: A report of earthquake is on the radio news.

How many probabilities are there in the joint probability distribution?

Modelling the Holmes Scenario

What are the random variables?

→

- B: A Burglary is happening.
- A: The alarm is going.
- W: Dr. Watson is calling.
- G: Mrs. Gibbon is calling.
- E: Earthquake is happening.
- R: A report of earthquake is on the radio news.

How many probabilities are there in the joint probability distribution?

→ There are $2^6 = 64$ probabilities.

Probability over a Subset of the Variables

Given a joint probability distribution,
we can compute the probability over a subset of the variables using the *sum rule*.

We can sum out every variable that we do not care about.

Probability over a Subset of the Variables

Given a joint probability distribution,
we can compute the probability over a subset of the variables using the *sum rule*.

We can sum out every variable that we do not care about.

→ Sum out: Fix the variables that we do care about. Add up all probabilities while varying the value of the variables that we don't care about.

Start with $P(A, B, C)$. To calculate $P(A \wedge B)$, we can sum out C:

$$P(A \wedge B) = P(A \wedge B \wedge C) + P(A \wedge B \wedge \neg C).$$

Probability over a Subset of the Variables

Given a joint probability distribution,
we can compute the probability over a subset of the variables using the *sum rule*.

We can sum out every variable that we do not care about.

→ Sum out: Fix the variables that we do care about. Add up all probabilities while varying the value of the variables that we don't care about.

Start with $P(A, B, C)$. To calculate $P(A \land B)$, we can sum out C:

$$P(A \land B) = P(A \land B \land C) + P(A \land B \land \neg C).$$

To calculate $P(A)$, we can further sum out B:

$$P(A) = P(A \land B) + P(A \land \neg B).$$

Q #2: Probability over a subset of the variables

Q: What is probability that
the alarm is NOT going and Dr. Watson is calling?

(A) 0.36
(B) 0.46
(C) 0.56
(D) 0.66
(E) 0.76

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

Q #2: Probability over a subset of the variables

Q: What is probability that
the alarm is NOT going and Dr. Watson is calling?

(A) 0.36
(B) 0.46
(C) 0.56
(D) 0.66
(E) 0.76

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$\rightarrow$ Correct answer is (A) 0.36.

$$P(\neg A \land W) = P(\neg A \land W \land G) + P(\neg A \land W \land \neg G)$$

$$= 0.036 + 0.324 = 0.36$$

Q #3: Probability over a subset of the variables

Q: What is probability that
the alarm is going and Mrs. Gibbon is NOT calling?

(A) 0.05
(B) 0.06
(C) 0.07
(D) 0.08
(E) 0.09

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

Q #3: Probability over a subset of the variables

Q: What is probability that
the alarm is going and Mrs. Gibbon is NOT calling?

(A) 0.05
(B) 0.06
(C) 0.07
(D) 0.08
(E) 0.09

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$\rightarrow$ Correct answer is (B) 0.06.

$$P(A \wedge \neg G) = P(A \wedge \neg W \wedge \neg G) + P(A \wedge W \wedge \neg G)$$

$$= 0.012 + 0.048 = 0.06$$

Q #4: Probability over a subset of the variables

Q: What is probability that **the alarm is NOT going**?

(A) 0.1
(B) 0.3
(C) 0.5
(D) 0.7
(E) 0.9

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

Q #4: Probability over a subset of the variables

Q: What is probability that **the alarm is NOT going**?

(A) 0.1
(B) 0.3
(C) 0.5
(D) 0.7
(E) 0.9

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$\rightarrow P(\neg A) = 0.036 + 0.054 + 0.324 + 0.486 = 0.9$

Correct answer is (E) 0.9.

Conditional Probability

Given a joint probability distribution,
how do we compute the probability one variable A
conditioned on knowing the value of another variable B?

We can use **the product rule**.

For example, how do we calculate $P(A|B)$
given a joint distribution over A, B, C?

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

$\rightarrow$

Conditional Probability

Given a joint probability distribution,
how do we compute the probability one variable A conditioned on knowing the value of another variable B?

We can use **the product rule**.

For example, how do we calculate $P(A|B)$ given a joint distribution over A, B, C?

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

$\rightarrow$ Only shows the case when A and B are both true. Convert between a prior/unconditional probability and a conditional probability.

Observing B = true rules out all possible worlds where B is false, leaving a set whose total probability is just P(B = true). Within that set, we want the worlds in which A is true.

Q #5: A conditional probability

Q: What is probability that
Dr. Watson is calling given that the alarm is NOT going?

(A) 0.2
(B) 0.4
(C) 0.6
(D) 0.8
(E) 1.0

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$P(\neg A \land W) = 0.36$,
$P(A \land \neg G) = 0.06$,
$P(\neg A) = 0.9$.

Q #5: A conditional probability

Q: What is probability that
Dr. Watson is calling given that the alarm is NOT going?

(A) 0.2
(B) 0.4
(C) 0.6
(D) 0.8
(E) 1.0

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$P(\neg A \wedge W) = 0.36,$
$P(A \wedge \neg G) = 0.06,$
$P(\neg A) = 0.9.$

$\rightarrow P(W|\neg A) = P(\neg A \wedge W)/P(\neg A) = 0.36/0.9 = 0.4$

Correct answer is (B) 0.4.

Q #6: A conditional probability

Q: What is probability that
Mrs. Gibbon is NOT calling given that the alarm is going?

(A) 0.2
(B) 0.4
(C) 0.6
(D) 0.8
(E) 1.0

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$P(\neg A \wedge W) = 0.36$,
$P(A \wedge \neg G) = 0.06$,
$P(\neg A) = 0.9$.

Q #6: A conditional probability

Q: What is probability that
Mrs. Gibbon is NOT calling given that the alarm is going?

(A) 0.2
(B) 0.4
(C) 0.6
(D) 0.8
(E) 1.0

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$P(\neg A \land W) = 0.36,$
$P(A \land \neg G) = 0.06,$
$P(\neg A) = 0.9.$

$\rightarrow P(A) = 1 - P(\neg A) = 0.1$

$P(\neg G|A) = P(\neg G \land A)/P(A) = 0.06/0.1 = 0.6$

Correct answer is (C) 0.6.

Inference Using the Prior and Conditional Probabilities

How do we

- calculate a probability over a subset of the variables?
- calculate a conditional probability?

The prior probabilities

$P(A) = 0.1$

The conditional probabilities

$P(W|A) = 0.9$

$P(W|\neg A) = 0.4$

$P(W|A \land G) = 0.9$

$P(W|A \land \neg G) = 0.9$

$P(W|\neg A \land G) = 0.4$

$P(W|\neg A \land \neg G) = 0.4$

$P(G|A) = 0.3$

$P(G|\neg A) = 0.1$

$P(G|A \land W) = 0.3$

$P(G|A \land \neg W) = 0.3$

$P(G|\neg A \land W) = 0.1$

$P(G|\neg A \land \neg W) = 0.1$

Calculate a Joint Probability Using the Chain Rule

For two variables (a.k.a. the product rule):

$$P(A \wedge B) = P(A|B) * P(B)$$

For three variables:

$$P(A \wedge B \wedge C) = P(A|B \wedge C) * P(B|C) * P(C)$$

For any number of variables:

$$\begin{aligned}
&P(X_n \wedge X_{n-1} \wedge \cdots \wedge X_2 \wedge X_1) \\
&= \prod_{i=1}^{n} P(X_i | X_{i-1} \wedge \cdots \wedge X_1) \\
&= P(X_n | X_{n-1} \wedge \cdots \wedge X_2 \wedge X_1) * ... * P(X_2|X_1) * P(X_1)
\end{aligned}$$

Q #7: Calculate a joint probability

Q: What is probability that **the alarm is going, Dr. Watson is calling and Mrs. Gibbon is NOT calling**?

(A) 0.060
(B) 0.061
(C) 0.062
(D) 0.063
(E) 0.064

$P(A) = 0.1$
$P(W|A) = 0.9$
$P(W|A \land \neg G) = 0.9$
$P(G|A) = 0.3$
$P(G|A \land W) = 0.3$

Q #7: Calculate a joint probability

Q: What is probability that **the alarm is going, Dr. Watson is calling and Mrs. Gibbon is NOT calling**?

(A) 0.060
(B) 0.061
(C) 0.062
(D) 0.063
(E) 0.064

$P(A) = 0.1$
$P(W|A) = 0.9$
$P(W|A \land \neg G) = 0.9$
$P(G|A) = 0.3$
$P(G|A \land W) = 0.3$

$\rightarrow P(A \land W \land \neg G) = P(A) * P(W|A) * P(\neg G|A \land W) = 0.1 * 0.9 * 0.7 = 0.063$

Correct answer is (D) 0.063.

Q #8: Calculate a joint probability

Q: What is probability that **the alarm is NOT going, Dr. Watson is NOT calling and Mrs. Gibbon is NOT calling**?

(A) 0.486
(B) 0.586
(C) 0.686
(D) 0.786
(E) 0.886

$$P(A) = 0.1$$
$$P(W|\neg A) = 0.4$$
$$P(W|\neg A \land \neg G) = 0.4$$
$$P(G|\neg A) = 0.1$$
$$P(G|\neg A \land \neg W) = 0.1$$

Q #8: Calculate a joint probability

Q: What is probability that **the alarm is NOT going, Dr. Watson is NOT calling and Mrs. Gibbon is NOT calling**?

(A) 0.486
(B) 0.586
(C) 0.686
(D) 0.786
(E) 0.886

$P(A) = 0.1$
$P(W|\neg A) = 0.4$
$P(W|\neg A \land \neg G) = 0.4$
$P(G|\neg A) = 0.1$
$P(G|\neg A \land \neg W) = 0.1$

$\rightarrow P(\neg A \land \neg W \land \neg G) = P(\neg A) * P(\neg W|\neg A) * P(\neg G|\neg A \land \neg W)$

$= 0.9 * 0.6 * 0.9 = 0.486$

Correct answer is (A) 0.486.

Flipping a Conditional Probability

Often you have causal knowledge:

- $P(symptom \mid disease)$
- $P(alarm \mid fire)$

...and you want to do evidential reasoning:

- $P(disease \mid symptom)$
- $P(fire \mid alarm)$.

Flipping a Conditional Probability using the Bayes' Rule

Definition (Bayes' rule)

$$P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y)}.$$

Flipping a Conditional Probability using the Bayes' Rule

Definition (Bayes' rule)

$$P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y)}.$$

$\rightarrow$ You should not memorize the Bayes' rule.
You should be able to derive it using the product rule.

$$P(X \wedge Y) = P(X|Y) * P(Y) = P(Y|X) * P(X).$$

We do not need to know $P(Y)$ to calculate $P(X|Y)$. $P(Y)$ is simply a normalization constant. We can calculate $P(X|Y)$ and $P(\neg X|Y)$, and then normalize them to sum to 1.

Q #9: Flipping a conditional probability

Q: What is the probability that
the alarm is NOT going given that Dr. Watson is calling?

(A) 0.6
(B) 0.7
(C) 0.8
(D) 0.9
(E) 1.0

$$P(A) = 0.1$$
$$P(W|A) = 0.9$$
$$P(W|\neg A) = 0.4$$

Q #9: Flipping a conditional probability

Q: What is the probability that
the alarm is NOT going given that Dr. Watson is calling?

(A) 0.6
(B) 0.7
(C) 0.8
(D) 0.9
(E) 1.0

$$P(A) = 0.1$$
$$P(W|A) = 0.9$$
$$P(W|\neg A) = 0.4$$

$\rightarrow$

$$P(W) = P(A)P(W|A) + P(\neg A)P(W|\neg A)$$
$$= 0.1 * 0.9 + 0.9 * 0.4 = 0.45$$
$$P(\neg A|W) = P(\neg A)P(W|\neg A)/P(W) = 0.4 * 0.9/0.45 = 0.8$$

Correct answer is (C) 0.8.

Q #10: Flipping a conditional probability

Q: What is the probability that
the alarm is going given that Mrs. Gibbon is NOT calling?

(A) 0.04
(B) 0.05
(C) 0.06
(D) 0.07
(E) 0.08

$$P(A) = 0.1$$
$$P(G|A) = 0.3$$
$$P(G|\neg A) = 0.1$$

Q #10: Flipping a conditional probability

Q: What is the probability that
the alarm is going given that Mrs. Gibbon is NOT calling?

(A) 0.04
(B) 0.05
(C) 0.06
(D) 0.07
(E) 0.08

$$P(A) = 0.1$$
$$P(G|A) = 0.3$$
$$P(G|\neg A) = 0.1$$

$\rightarrow$

$$P(\neg G) = P(A)P(\neg G|A) + P(\neg A)P(\neg G|\neg A)$$
$$= 0.1 * 0.7 + 0.9 * 0.9 = 0.88$$
$$P(A|\neg G) = P(A)P(\neg G|A)/P(\neg G) = 0.1 * 0.7/0.88 = 0.08$$

Correct answer is (E) 0.08.

A universal approach

1. To calculate a conditional probability, convert it into a fraction of two joint probabilities using the product rule in reverse.

2. To calculate a joint probability (not involving all the variables), write it as a summation of joint probabilities (involving all the variables) by introducing the other variables using the sum rule in reverse.

3. Calculate every joint probability (involving all the variables) using the chain rule.

A universal approach: an example

Calculate $P(A|C)$, given $P(A)$, $P(B|A)$ and $P(C|A \wedge B)$.

You are given the following:

- $P(A) = 0.6$
- $P(B|A) = 0.4$, $P(\neg B|\neg A) = 0.2$
- $P(C|A \wedge B) = 0.1$, $P(C|\neg A \wedge B) = 0.2$, $P(C|A \wedge \neg B) = 0.5$, $P(C|\neg A \wedge \neg B) = 0.8$

Step 1: Conditional Probability to Joint Probability

Convert a conditional probability to joint probabilities.

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

Step 2: Convert Joint Probability to Involve All Missing Variables

Convert joint probability to involve all the variables

$$P(A|C) = \frac{P(A \wedge C)}{P(C)} = \frac{P(A \wedge C)}{P(A \wedge C) + P(\neg A \wedge C)}$$

$$P(A \wedge C) = P(A \wedge C \wedge B) + P(A \wedge C \wedge \neg B)$$

$$P(\neg A \wedge C) = P(\neg A \wedge C \wedge B) + P(\neg A \wedge C \wedge \neg B)$$

Step 3: Calculate Joint Probability with Chain Rule

Convert a joint probability into product of conditional distributions

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

$$P(A \wedge C \wedge B) = P(C|A \wedge B)P(B|A)P(A)$$
$$P(A \wedge C \wedge \neg B) = P(C|A \wedge \neg B)P(\neg B|A)P(A)$$
$$P(\neg A \wedge C \wedge B) = P(C|\neg A \wedge B)P(B|\neg A)P(\neg A)$$
$$P(\neg A \wedge C \wedge B) = P(C|\neg A \wedge \neg B)P(\neg B|\neg A)P(\neg A)$$