Bayesian Probability

28 October 2020
“All models are wrong, but some are useful.”

George E. P. Box
Logic

Logical representations are based on facts about the world – claims that are true or false – which can be combined with logical connectives.

Traditional logical inference is based on what we can conclude with certainty.
The real world involves uncertainty, and it’s hard to give general rules that always hold.

Models aren’t perfect, and learned models are especially imperfect.
∀ x . Fruit(x) ⇒ Tasty(x)

“All fruit is tasty.”
The Yale shooting problem

Fred, a turkey, is alive. \( \text{Alive}(0) \)

There is a gun, which is unloaded. \( \neg \text{Loaded}(0) \)

We load the gun, wait a moment, and shoot the gun at Fred.

\( \text{True} \Rightarrow \text{Loaded}(1) \)

\( \text{Loaded}(2) \Rightarrow \neg \text{Alive}(3) \)

Is Fred dead?
The Yale shooting problem

Fred, a turkey, is alive. \( \text{Alive}(0) \)

There is a gun, which is unloaded. \( \neg \text{Loaded}(0) \)

We load the gun, wait a moment, and shoot the gun at Fred.

\( \text{True} \Rightarrow \text{Loaded}(1) \)

\( \text{Loaded}(2) \Rightarrow \neg \text{Alive}(3) \)

Is Fred dead?

\emph{Maybe the gun became unloaded at time 2 when we weren’t looking. Then Fred’s alive!}
The Yale shooting problem

Fred, a turkey, is alive. \( \text{Alive}(0) \)

There is a gun, which is unloaded. \( \neg \text{Loaded}(0) \)

We load the gun, wait a moment, and shoot the gun at Fred.

\( \text{True} \Rightarrow \text{Loaded}(1) \)

\( \text{Loaded}(2) \Rightarrow \neg \text{Alive}(3) \)

Is Fred dead? \boxed{\text{But even if the gun was unloaded, maybe Fred died of fright at time 1!}}
There are many approaches, including

*Default (non-monotonic) logic*, where conclusions are tentative and can be withdrawn.

*Circumscription* (McCarthy): Adding predicates saying “and everything else is normal”, e.g.,

- If $x$ is a fruit and everything about it is normal, it’s tasty.
- If $y$ is a bird and everything about it is normal, it can fly.

(So, penguins are “abnormal” birds?)
Or we might decide we want to associate numeric weights with our rules, e.g.,

\[ \text{Sprinkler} \rightarrow_{0.99} \text{WetGrass} \]
\[ \text{WetGrass} \rightarrow_{0.7} \text{Rain} \]

This can lead to problems with combination:

Seeing the sprinkler running leads us to conclude it rained?
Introduction: Bayesian probability
“Life is intrinsically, well, boring and dangerous at the same time. At any given moment, the floor may open up. Of course, it almost never does; that's what makes it so boring.”

Edward Gorey
Probabilities are a powerful tool for reasoning about uncertainty, i.e., how likely events are to happen.

Probabilistic assertions summarize the effects of

*Laziness*: Failure to enumerate exceptions, qualifications, etc.

*Ignorance*: Lack of relevant facts, initial conditions, etc.
What do probabilities mean, anyway?
Relative frequencies

\[ P(A) \]: The probability a random event falls in \( A \) rather than \( \text{Not } A \).

This works well for dice and coin flips. (See CMPU 145!)
What's the probability of the US women's field hockey team winning gold medals at the 2020 2021 olympics in Tokyo?
What's the probability of the US women's field hockey team winning gold medals at the 2020 2021 Olympics in Tokyo?

It seems low. They've never won the gold. In 2016 they came in fifth.
What’s the probability of the US women’s field hockey team winning gold medals at the 2020 2021 olympics in Tokyo?

It seems low. They’ve never won the gold. In 2016 they came in fifth. However, it’s a one-time event, so there’s no way to make statements like “they will win x out of y times at the 2021 Summer Olympics”.
The relative idea of probability suggests what will happen over a repeating series of 2021 Summer Olympics – which is nonsensical.
How each candidate’s chances of winning more than half of pledged delegates have changed over time

Our latest forecast for how many pledged delegates each candidate will win after all states have voted. FiveThirtyEight is feeding into this in real time, and if you’d like to find the full data set for each candidate, you can click on their name.
Probabilities and beliefs

Suppose I flip a coin and hide the outcome.

What’s $P(\text{Heads})$?

- This is a statement about a belief, not about the world.
- The world is in exactly one state; the event’s already happened.

Assigning truth values to probabilities is tricky – we need to reference the speaker’s state of knowledge.
Views of probability:

Frequentist: Probabilities come from the relative frequencies of events, that is, the tendency for events like the one being considered to take place in similar contexts.
Views of probability:

Frequentist: Probabilities come from the relative frequencies of events, that is, the tendency for events like the one being considered to take place in similar contexts.

What does it mean for an event to be “like” the one being considered?
Views of probability:

**Frequentist:** Probabilities come from the relative frequencies of events, that is, the tendency for events like the one being considered to take place in similar contexts.

What does it mean for an event to be “like” the one being considered?

What counts as a similar context?
Views of probability:

*Frequentist:* Probabilities come from the relative frequencies of events, that is, the tendency for events like the one being considered to take place in similar contexts.

What does it mean for an event to be “like” the one being considered?

What counts as a similar context?

It’s questions all the way down.
Views of probability:

**Frequentist:** Probabilities come from the relative frequencies of events, that is, the tendency for events like the one being considered to take place in similar contexts.

**Subjectivist:** Probabilities are the degrees of belief a rational agent would have about possible events.
No two events are identical – or completely unique.

We can treat probabilities as beliefs but allow data (relative frequencies) to influence these beliefs.

That is, *in AI, probabilities reflect degrees of belief given observed evidence.*
This framework, where we assess probabilities in terms of expectation and belief, is sometimes called Bayesian probability after Thomas Bayes.
We'll use Bayes' rule to combine prior beliefs with new data.

Before we get to that, let's establish some basics.
Probabilities and distributions
Probabilities talk about *random variables*, i.e., aspects of the world about which we have uncertainty.

Like variables in a constraint-satisfaction problem, random variables have *domains*.

The domain of random variable $X$ can be written as $d(X)$ or we can say $X \in \{\ldots\}$

Domains may be discrete or continuous.

$X = x$ means the random variable $X$ has taken value $x$, and $P(x)$ is short for $P(X = x)$. 
Examples

\begin{align*}
R &= \text{Is it raining?} \\
T &= \text{Is it hot or cold?} \\
D &= \text{How long will it take to drive to work?}
\end{align*}

\begin{align*}
R &\in \{\text{True, False}\} \\
&\text{sometimes written as \{+r, \, -r\}} \\
T &\in \{\text{hot, cold}\} \\
D &\in [0, \infty)
\end{align*}
A *distribution* (for a discrete variable) is a table of the probabilities of the values.

Must have $\forall x \ . \ P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

<table>
<thead>
<tr>
<th></th>
<th>$P(T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$P$</td>
</tr>
<tr>
<td>hot</td>
<td>0.5</td>
</tr>
<tr>
<td>cold</td>
<td>0.5</td>
</tr>
</tbody>
</table>

A *probability* (lower case value) is a single number. $P(T = \text{cold}) = 0.5$
Examples

Let $X$ be a random variable indicating who wins the popular vote:

$$d(X) = \{\text{Biden, Trump, tie}\}.$$ 

A probability is associated with each event in the domain:

$$P(X = \text{Biden}) = 0.96$$
$$P(X = \text{Trump}) = 0.04$$
$$P(X = \text{tie}) < 0.01$$

Probabilities over the entire event space must sum to 1.
Expectation

Common use of probabilities: Each event has a *numerical value*.

For example, what is the average value of rolling a six-sided die?

\[
\frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5
\]

In general, given random variable \( X \) and function \( f(x) \):

\[
E[f(x)] = \sum_x P(x)f(x)
\]
Expectation

For example, in minimax search, we assumed the opposing player took the min-valued action (for us).

That assumed perfect play. If we have a probability distribution over the player’s actions, we can calculate their *expected value* (instead of min value) for each action.

Result: the expectimax algorithm.
Kolmogorov’s axioms of probability

\[ 0 \leq P(x) \leq 1 \]

\[ P(\text{True}) = 1 \]

\[ P(\text{False}) = 0 \]

\[ P(a \lor b) = P(a) + P(b) - P(a \land b) \]

These are sufficient to completely specify probability theory for discrete variables.
What should we do when several variables are involved?

Think about *atomic events*:

- Complete assignment of all variables
- All possible events
- Mutually exclusive
### Joint probability distribution

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>True</td>
<td>0.3</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>0.1</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>0.4</td>
</tr>
<tr>
<td>False</td>
<td>False</td>
<td>0.2</td>
</tr>
</tbody>
</table>

*Note:* Still adds up to 1
A **joint distribution** over a set of random variables, $X_1, X_2, \ldots, X_n$, specifies a real number for each assignment (or **outcome**), written as

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

or

$$P(x_1, x_2, \ldots, x_n).$$
Some analogies

\[ X \land Y \]

\[ X \lor Y \]

\[ \neg X \]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>1/3</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>1/3</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>1/3</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
</tr>
</tbody>
</table>
Probabilities for all possible atomic events:

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>True</td>
<td>0.3</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>0.1</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>0.4</td>
</tr>
<tr>
<td>False</td>
<td>False</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Can define individual probabilities in terms of the joint probability distribution, e.g.,

\[ P(\text{Raining}) = P(\text{Raining, Cold}) + P(\text{Raining, not Cold}) = 0.4 \]

\[ P(a) = \sum_{e_i \in e(a)} P(e_i) \]
Simplistic knowledge base:

Variables of interest, $X_1, \ldots, X_n$

Joint probability distribution over $X_1, \ldots, X_n$

This expresses all possible statistical information about relationships between the variables of interest.

Inference:

Queries over subsets of $X_1, \ldots, X_n$

E.g., $P(X_3)$

E.g., $P(X_3 \mid X_1)$
What if you have a joint probability distribution and you acquire new data?

E.g., I heard it’s cold. What’s the probability it’s also raining?
Conditioning
The *conditional probability* $P(B \mid A)$ is the probability of $B$ if $A$ is known ("fixed", "given").

Ways to think about this:

- $B$ is belief; $A$ is evidence affecting belief.
- $B$ is belief; $A$ is hypothetical.
- $B$ is unobserved; $A$ is observed.

Soft version of implication: $(A \Rightarrow B) \approx P(B \mid A) = 1$
We can write

\[ P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \]

This tells us the probability of \( A \) given \textit{only} knowledge \( B \).

This is a probability with regard to a \textit{state of knowledge}.

\[ P(\text{Disease} \mid \text{Symptom}) \]
\[ P(\text{Raining} \mid \text{Cold}) \]
\[ P(\text{Win} \mid \text{Injury}) \]
\[ P(\text{Raining} \mid \text{Cold}) = \frac{P(\text{Raining and Cold})}{P(\text{Cold})} \]

\[ = \frac{0.3}{0.7} \]

\[ \approx 0.43 \]

**Note:** \( P(\text{Raining} \mid \text{Cold}) + P(\text{not Raining} \mid \text{Cold}) = 1 \)
**Conditional distributions** are probability distributions over some variables given fixed values of others.

<table>
<thead>
<tr>
<th>Joint distribution</th>
<th>Conditional distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(T, W)$</td>
<td>$P(W \mid T = \text{hot})$</td>
</tr>
<tr>
<td>$T$</td>
<td>$W$</td>
</tr>
<tr>
<td>hot</td>
<td>sun</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
</tr>
<tr>
<td>$P(W \mid T = \text{cold})$</td>
<td></td>
</tr>
<tr>
<td>$W$</td>
<td>$P$</td>
</tr>
<tr>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Joint distributions are *everything*.

That is, they’re all you (statistically) need to know about $X_1, \ldots, X_n$.

Classification: $P(X_1 \mid X_2, \ldots, X_n)$

- **Thing you want to know**
- **Things you know**

Co-occurrence: $P(X_a, X_b)$

- *How likely are these two events together*

Rare event detection: $P(X_1, \ldots, X_n)$
Joint probability tables

- Grow very fast
- Need to sum out the other variables
- Might require lots of data
- Are not a function of $P(A)$ and $P(B)$.
Two events $A$ and $B$ are *independent* if and only if the fact that $A$ happened has no effect on the probability of $B$ and vice versa.

In other words,

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

and

$$P(B) = P(B \mid A).$$

E.g., two successive flips of a fair coin.
Independence is a simplifying *modeling assumption*

*Empirical* joint distributions are, at best, “close” to independent

If A and B are independent:

\[
P(A \text{ and } B) = P(A)P(B)
\]

\[
P(A \text{ or } B) = P(A) + P(B) - P(A)P(B)
\]
Are *Raining* and *Cold* independent?

\[ P(Raining = True) = 0.4 \]

\[ P(Cold = True) = 0.7 \]

\[ P(Raining = True, Cold = True) = ? \]
If the variables are independent, we can break the joint probability distribution into separate tables.

<table>
<thead>
<tr>
<th>Raining</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.6</td>
</tr>
<tr>
<td>False</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cold</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.75</td>
</tr>
<tr>
<td>False</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>True</td>
<td>0.45</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>0.15</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>0.3</td>
</tr>
<tr>
<td>False</td>
<td>False</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Often we have an intuition about independence, but always verify.

Dependence does not mean causation!
Much of probabilistic knowledge representation and machine learning is concerned with identifying and leveraging independence and mutual exclusivity.

    Independence is very rare.
    Is there a weaker type of structure we might be able to exploit?
A and B are *conditionally independent* given C if:

\[
P(A \mid B, C) = P(A \mid C)
\]

\[
P(A \mid B, C) = P(A \mid C) P(B \mid C)
\]

This means that if we know C, we can treat A and B as if they were independent.

A and B might not be independent otherwise!
Example

Consider three random variables:

*Temperature*

*Humidity*

*Season*

*Temperature* and *Humidity* are not independent, but they might be, given *Season*.

The season explains both, so they become independent of each other.
Bayes’ rule

A special piece of conditioning magic:

\[
P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}
\]

If we have condition \( P(A \mid B) \) and we receive new data for \( A \), we can compute a new distribution for \( B \), without needing the joint distribution.

*As new evidence comes in, we revise our beliefs.*
\[ P(B \mid A) = P(A \mid B) \cdot P(B) \] 

The “prior probability” – our assessment of the likelihood that B is true before we get a new piece of evidence.

Once we learn that A is also true, we can update our prior probability to get the “posterior probability”, \( P(B \mid A) \), that B is true in light of that new fact.
Bayesian knowledge bases

List of conditional and marginal probabilities…

\[ P(X_1) = 0.7 \]
\[ P(X_2) = 0.6 \]
\[ P(X_3 \mid X_2) = 0.57 \]

Queries:

\[ P(X_2 \mid X_1) \]
\[ P(X_3) \]

Less onerous than a joint probability distribution, but you may or may not be able to answer some questions.
Probabilistic inference is computing a desired probability from other known probabilities (e.g., conditional probability from joint probability).

We generally compute conditional probabilities

\[ P(\text{on time} \mid \text{no reported accidents}) = 0.90 \]
These represent the agent’s beliefs given the evidence

Probabilities change with new evidence:

\[ P(\text{on time} \mid \text{no accidents, 5 a.m.}) = 0.95 \]
\[ P(\text{on time} \mid \text{no accidents, 5 a.m., raining}) = 0.80 \]
Observing new evidence causes beliefs to be updated
A Bayesian narrative
Accident…or arson?
What’s the probability it was arson?
What’s the probability it was arson?

In how many of the times that this house burned, was the fire started deliberately?
What’s the probability it was arson?

In how many of the times that this house burned, was the fire started deliberately?
What's the probability it was arson?

In how many of the times that this house burned, was the fire started deliberately?

How confident are we that it was arson?
The fire facts

1. *Prior probability*: About 10% of the fires in this town are arson.

2. There were strong signs this particular fire burned very hot – hotter than one would expect for a typical house fire, but consistent with the use of something like thermite to start it quickly. Only 8% of the fires in this town burn that hot.

3. About 50% of the arsons in this town use accelerants like thermite. In other words, the probability any given fire burns that hot *if* it were arson is about 0.5.
The fire facts

1. **Prior probability**: About 10% of the fires in this town are arson.

2. There were strong signs this particular fire burned very hot – hotter than one would expect for a typical house fire, but consistent with the use of something like thermite to start it quickly. Only 8% of the fires in this town burn that hot.

3. About 50% of the arsons in this town use accelerants like thermite. In other words, the probability any given fire burns that hot if it were arson is about 0.5.
The fire facts

1. *Prior probability*: About 10% of the fires in this town are arson.

2. There were strong signs this particular fire burned very hot – hotter than one would expect for a typical house fire, but consistent with the use of something like thermite to start it quickly. Only 8% of the fires in this town burn that hot.

3. About 50% of the arsons in this town use accelerants like thermite. In other words, the probability any given fire burns that hot *if* it were arson is about 0.5.
The fire facts

1. **Prior probability**: About 10% of the fires in this town are arson.

2. There were strong signs this particular fire burned very hot – hotter than one would expect for a typical house fire, of something like the thermite variety. Only 8% of the fires in this town burn that hot.

3. About 50% of the arsons in this town use accelerants like thermite. In other words, the probability any given fire burns that hot if it were arson is about 0.5.
$B = \text{“the fire was started deliberately”}$

$A = \text{“the fire burned very hot”}$

$P(B) = 0.10$

$P(A) = 0.08$

$P(A \mid B) = 0.50$

What’s the probability the fire was started deliberately given that it was a very hot fire?

$P(B \mid A) = ?$
$B = \text{“the fire was started deliberately”}$

$A = \text{“the fire burned very hot”}$

$P(B) = 0.10$

$P(A) = 0.08$

$P(A \mid B) = 0.50$

What’s the probability the fire was started deliberately given that it was a very hot fire?

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

$$= \frac{0.10 \cdot 0.50}{0.08} \approx 63\%$$
Accident…or arson?
Accident...or arson?

Very suspicious! Investigate further.
Based on the evidence, the investigator updates his *prior* estimate of the likelihood of arson (10%) to a higher *posterior* estimate (63%), reflecting the evidence from the heat of the fire.

As more evidence comes in, the posterior becomes the new prior and he can update this new prior even further, until all the evidence has been accounted for.
Bayesian perils

**Modified Bayes' Theorem:**

\[ P(H|X) = P(H) \times \left( 1 + P(C) \times \left( \frac{P(x|H)}{P(x)} - 1 \right) \right) \]

- **H**: Hypothesis
- **X**: Observation
- **P(H)**: Prior probability that H is true
- **P(x)**: Prior probability of observing X
- **P(C)**: Probability that you’re using Bayesian statistics correctly
Who will win the presidency?

**Chance of winning**

- **Hillary Clinton**: 71.4%
- **Donald Trump**: 28.6%
This was actually a pretty good forecast. But you get the point.
Acknowledgments

The lecture incorporates material from:

Patrick Juola and Stephen Ramsay, *Six Septembers: Mathematics for the Humanist*

Dan Klein and Pieter Abbeel, University of California, Berkeley; ai.berkeley.edu

George Konidaris, Brown University

Peter Norvig and Stuart Russell, *Artificial Intelligence: A Modern Approach*