Uncertainty

George Konidaris
gdk@cs.brown.edu

Spring 2017
Knowledge
Logic

Logical representations are based on:
  • *Facts* about the world.
  • Either *true* or *false*.
  • *We may not know which*.
  • Can be combined with logical connectives.

Logic inference is based on:
  • What we can conclude *with certainty*. 
Logic is Insufficient

The world is not deterministic.
There is no such thing as a fact.
Generalization is hard.
Sensors and actuators are noisy.
Plans fail.
Models are not perfect.
Learned models are especially imperfect.

\[ \forall x, \text{Fruit}(x) \implies \text{Tasty}(x) \]
Probability Theory
The Logic of Science
E. T. Jaynes
Probabilities

Powerful tool for reasoning about uncertainty.

Can prove that a person who holds a system of beliefs inconsistent with probability theory can be fooled.

But, they’re tricky:
  • Intuition often wrong or inconsistent.
  • Difficult to get.

What do probabilities really mean?
Relative Frequencies

Defined over events.

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

Works well for dice and coin flips!
Relative Frequencies

But this feels limiting.

What is the probability that the Red Sox win this year’s World Series?

• Meaningful question to ask.
• Can’t count frequencies (except naively).
• Only really happens once.

In general, all events only happen once.
Probabilities and Beliefs

Suppose I flip a coin and hide outcome.

- What is \( P(\text{Heads}) \)?

This is a statement about a belief, not the world. (the world is in exactly one state, with prob. 1)

Assigning truth values to probabilities is tricky - must reference speaker’s state of knowledge.

**Frequentists**: probabilities come from relative frequencies.

**Subjectivists**: probabilities are degrees of belief.
For Our Purposes

No two events are identical, or completely unique.

Use probabilities as beliefs, but allow data (relative frequencies) to influence these beliefs.

We use Bayes’ Rule to combine prior beliefs with new data.
To The Math

Probabilities talk about *random variables*:

- \(X, Y, Z\), with domains \(d(X), d(Y), d(Z)\).
- Domains may be *discrete* or *continuous*.
- \(X = x\): RV \(X\) has taken value \(x\).
  - Binary RVs: domain is \{true, false\}.
- \(P(x)\) is short for \(P(X = x)\).
Examples

X: RV indicating winner of Red Sox vs. Yankees game.

d(X) = \{\text{Red Sox, Yankees, tie}\}.

A probability is associated with each event in the domain:

- \( P(X = \text{Red Sox}) = 0.8 \)
- \( P(X = \text{Yankees}) = 0.19 \)
- \( P(X = \text{tie}) = 0.01 \)

Note: probabilities over the entire event space must sum to 1.
Kolmogorov’s Axioms of Probability

Sufficient to completely specify probability theory for discrete variables:

• $0 \leq P(x) \leq 1$
• $P(\text{true}) = 1, P(\text{false}) = 0$
• $P(a \text{ or } b) = P(a) + P(b) - P(a \text{ and } b)$
Multiple Events

What to do when several variables are involved?

Think about *atomic events*.
- Complete assignment of all variables.
- All possible events.
- Mutually exclusive.

RVs: Raining, Cold (both boolean):

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Prob.</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*.
Joint Probability Distribution

Probabilities to all possible atomic events (grows fast)

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Prob.</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 JPD:

\[
P(Raining) = P(Raining, Cold) + P(Raining, not Cold) = 0.4.
\]

\[
P(a) = \sum_{e_i \in e(a)} P(e_i)
\]
Independence

Critical property! But rare.

If A and B are independent:

• \( P(A \text{ and } B) = P(A)P(B) \)
• \( P(A \text{ or } B) = P(A) + P(B) - P(A)P(B) \)
Independence

Are *Raining* and *Cold* independent?

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Prob.</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>

\[
P(\text{Raining} = \text{True}) = 0.4
\]

\[
P(\text{Cold} = \text{True}) = 0.7
\]

\[
P(\text{Raining} = \text{True}, \text{Cold} = \text{True}) = \ ?
\]
Independence

If independent, can break JPD into separate tables.

<table>
<thead>
<tr>
<th>Raining</th>
<th>Prob.</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>Prob.</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>Prob.</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>
Independence is Critical

To compute $P(A \text{ and } B)$ we need a joint probability.

- This grows very fast.
- Need to sum out the other variables.
- Might require lots of data.
- NOT a function of $P(A)$ and $P(B)$.

If $A$ and $B$ are independent, then you can use separate, smaller tables.

_Much of machine learning and statistics is concerned with identifying and leveraging independence and mutual exclusivity._
Independence: Examples

*Independence: two events don’t effect each other.*

- Red Sox winning world series, Andy Murray winning Wimbledon.
- Two successive, fair, coin flips.
- It is raining, and winning the lottery.
- Poker hand and date.

Often we have an intuition about independence, but *always verify*. Dependence does not mean causation!
Mutual Exclusion

Two events are mutually exclusive when:

- $P(A \text{ or } B) = P(A) + P(B)$.
- $P(A \text{ and } B) = 0$.

This is *different from* independence.
Conditional Probabilities

What if you have a joint probability, and you acquire new data?

My iPhone tells me that it's cold. What is the probability that it is raining?

Write this as:

- $P(\text{Raining} \mid \text{Cold})$

<table>
<thead>
<tr>
<th>Raining</th>
<th>Cold</th>
<th>Prob.</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>
Conditional Probabilities

We can write:

\[ P(a|b) = \frac{P(a \text{ and } b)}{P(b)} \]

This tells us the probability of \( a \) given only knowledge \( b \).

This is a probability w.r.t a state of knowledge.

- \( P(\text{Disease} \mid \text{Symptom}) \)
- \( P(\text{Raining} \mid \text{Cold}) \)
- \( P(\text{Red Sox win} \mid \text{injury}) \)
Conditional Probabilities

\[
P(\text{Raining} \mid \text{Cold}) = \frac{P(\text{Raining and Cold})}{P(\text{Cold})}
\]

\[
\ldots \quad P(\text{Cold}) = 0.7
\]

\[
\ldots \quad P(\text{Raining and Cold}) = 0.3
\]

\[
P(\text{Raining} \mid \text{Cold}) \approx 0.43.
\]

Note!

\[
P(\text{Raining} \mid \text{Cold}) + P(\text{not Raining} \mid \text{Cold}) = 1!
\]
Bayes's Rule

Special piece of conditioning magic.

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

If we have conditional \( P(B \mid A) \) and we receive new data for \( B \), we can compute new distribution for \( A \). (Don’t need joint.)

As evidence comes in, revise belief.
Bayes Example

Suppose $P(\text{cold}) = 0.7$, $P(\text{headache}) = 0.6$. 
$P(\text{headache} \mid \text{cold}) = 0.57$

What is $P(\text{cold} \mid \text{headache})$?

$$P(\text{c} \mid \text{h}) = \frac{P(\text{h} \mid \text{c})P(\text{c})}{P(\text{h})}$$

$$P(\text{c} \mid \text{h}) = \frac{0.57 \times 0.7}{0.6} = 0.66$$

Not always symmetric!
Not always intuitive!
\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

**Bayes**

- **Evidence**
- **Sensor Model**
- **Prior**
Probability Distributions

If you have a discrete RV, probability distribution is a table:

<table>
<thead>
<tr>
<th>Flu</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.6</td>
</tr>
<tr>
<td>False</td>
<td>0.4</td>
</tr>
</tbody>
</table>

What if you have a real-valued random variable?

- Temperature tomorrow
- Rainfall
- Number of votes in election
- Height
PDFs

Continuous probabilities described by probability density function $f(x)$. 

PDF is about density, not probability.

- Non-negative.
- $\int_{x} f(x) = 1$ integrates to 1
- $f(x)$ might be greater than 1.
PDFs

Can’t ask $P(x = 0.0014245)$?

The probability of a single real-valued number is zero.

Instead we can ask for a range:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx$$
Distributions

Distributions usually specified by a PDF type or family.

- Each family is a parametrized function describing the PDF.
- Get a specific distribution by fixing the parameters.
Uniform Distribution

For example, uniform distribution over $[0, 0.5]$.

Parameter: mean.
Gaussian (Normal)

A mean + an exponential drop-off, characterized by variance.

\[ f(x, \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
PDFs

When dealing with a real-valued variable, two steps:

• Specifying the family of distribution.
• Specifying the values of the parameters.

Conditioning on a discrete variable just means picking from a discrete number of parameter settings.

<table>
<thead>
<tr>
<th>$\mu_A$</th>
<th>$\sigma^2_A$</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.02</td>
<td>True</td>
</tr>
<tr>
<td>0.1</td>
<td>0.06</td>
<td>False</td>
</tr>
</tbody>
</table>
PDFs

Conditioning on real-valued RV:
• Parameters function of RV

Linear regression:
\[ f(x) = w \cdot x + \epsilon \]
\[ y \sim N(w \cdot x, \sigma^2) \]
Parametrized Forms

Many machine learning algorithms start with parametrized, generative models.

Find PDFs / CPTs (i.e., parameters) such that probability that they generated the data is maximized.

There are also non-parametric forms: describe the PDF directly from the data itself, not a function.