Basic Probability and Statistics

CS540 Bryan R Gibson University of Wisconsin-Madison

Slides adapted from those used by Prof. Jerry Zhu, CS540-1
Reasoning with Uncertainty

- There are two identical-looking envelopes
  - one has a red coin (worth $100) and a black coin (worth $0)
  - the other has two black coins

- You randomly grab an envelope and randomly pick out one coin - it’s black
- You’re then given the chance to switch envelopes: Should you?
Outline

Probability:
- Sample Space
- Random Variables
- Axioms of Probability
- Conditional Probability
- Probabilistic Inference: Bayes Rule
- Independence
- Conditional Independence
Uncertainty

- Randomness
  - Is our world random?

- Uncertainty
  - Ignorance (practical and theoretical)
    - Will my coin flip end in heads?
    - Will a pandemic flu strike tomorrow?

- Probability is the language of uncertainty
  - A central pillar of modern day A.I.
Sample Space

- A space of **Events** that we assign probabilities to
- Events can be binary, multi-valued or continuous
- Events are mutually exclusive

**Examples:**
- Coin flip: \{head,tail\}
- Die roll: \{1,2,3,4,5,6\}
- English words: a dictionary
- Temperature tomorrow: \(\mathbb{R}^+\) (kelvin)
Random Variable

- A variable $X$, whose domain is the sample space, and whose value is somewhat uncertain.

- Examples:
  - $X = \text{coin flip outcome}$
  - $X = \text{first word in tomorrow’s headline news}$
  - $X = \text{tomorrow’s temperature}$

- Kind of like $x = \text{rand()}$
Probability for Discrete Events

- Probability $P(X = a)$ is the fraction of times $X$ takes value $a$
- Often written as $P(a)$
- There are other definitions of prob. and philosophical debates, but we’ll set those aside for now

Examples:
- $P(\text{head}) = P(\text{tail}) = 0.5$ : a fair coin
- $P(\text{head}) = 0.51$, $P(\text{tail}) = 0.49$ : a slightly biased coin
- $P(\text{head}) = 1$, $P(\text{tail}) = 0$ : Jerry’s coin
- $P(\text{first word} = \text{“the” when flip to random page in R&N}) =$?

Demo: bookofodds
Example: Weather

<table>
<thead>
<tr>
<th></th>
<th>sunny</th>
<th>cloudy</th>
<th>rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>200/365</td>
<td>100/365</td>
<td>65/365</td>
</tr>
</tbody>
</table>

\[ P(\text{Weather} = \text{sunny}) = P(\text{sunny}) = \frac{200}{365} \]

\[ P(\text{Weather}) = \left\{ \frac{200}{365}, \frac{100}{365}, \frac{65}{365} \right\} \]

(For now, we’ll be satisfied with just using counted frequency of data to obtain probabilities . . . )
Prob. for Discrete Events (cont.)

Probability for more complex events: we’ll call it event $A$

- $P(A = \text{“head or tail”}) = ?$ (for a fair coin?)
- $P(A = \text{“even number”}) = ?$ (for a fair 6-sided die?)
- $P(A = \text{“two dice rolls sum to 2”}) = ?$
▶ Probability for more complex events: we’ll call it event $A$

▶ $P(A = “\text{head or tail}”) = \frac{1}{2} + \frac{1}{2} = 1$ (fair coin)

▶ $P(A = “\text{even number}”) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$ (fair 6-sided die)

▶ $P(A = “\text{two dice rolls sum to 2}”) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$
The Axioms of Probability

- \( P(A) \in [0, 1] \)
- \( P(true) = 1, \ P(false) = 0 \)
- \( P(A \lor B) = P(A) + P(B) - P(A \land B) \)
The Axioms of Probability (cont.)

- $P(A) \in [0, 1]$
The Axioms of Probability (cont.)

- \( P(A) \in [0, 1] \)

No fraction of \( A \) can be \textbf{bigger} than 1
The Axioms of Probability (cont.)

- \( P(true) = 1, \ P(false) = 0 \)

Sample Space

Valid sentence: e.g. “\( x = \text{head OR } x = \text{tail} \)”
The Axioms of Probability (cont.)

- \( P(true) = 1, \ P(false) = 0 \)

Sample Space

Invalid sentence: e.g. “\( x = \text{head AND } x = \text{tail} \)”
The Axioms of Probability (cont.)

- \( P(A \lor B) = P(A) + P(B) - P(A \land B) \)
Some Theorems Derived from Axioms

- \( P(\neg A) = 1 - P(A) \)

- If \( A \) can take \( k \) different values \( a_1, \ldots, a_k \):

\[
P(A = a_1) + \ldots + P(A = a_k) = 1
\]

- If \( A \) is a binary event:

\[
P(B) = P(B \land \neg A) + P(B \land A)
\]

- If \( A \) can take \( k \) values:

\[
P(B) = \sum_{i=1..k} P(B \land A = a_i)
\]
Joint Probability

- **Joint Probability:**
  \[ P(A = a, B = b), \text{ shorthand for } P(A = a \land B = b), \]
  is the probability of both \( A = a \) and \( B = b \) happening

\[
\begin{align*}
P(A = a) & : \text{ e.g. } P(1\text{st word} = \text{"San"}) = 0.001 \\
P(B = b) & : \text{ e.g. } P(2\text{nd word} = \text{"Francisco"}) = 0.0008 \\
P(A = a, B = b) & : \text{ e.g. } P(1\text{st} = \text{"San"}, \ 2\text{nd} = \text{"Francisco"}) = 0.0007
\end{align*}
\]
### Joint Probability Table

<table>
<thead>
<tr>
<th>Temp</th>
<th>Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sunny</td>
</tr>
<tr>
<td>hot</td>
<td>150/365</td>
</tr>
<tr>
<td>cold</td>
<td>50/365</td>
</tr>
</tbody>
</table>

- $P(\text{Temp} = \text{hot}, \text{Weather} = \text{rainy}) = P(\text{hot, rainy}) = \frac{5}{365}$
- The full joint probability table between $N$ variables, each taking $k$ values, has $k^N$ entries!
Marginal Probability

- Marginalize = Sum over “other” variables
- For example, marginalize over/out Temp:

<table>
<thead>
<tr>
<th>Temp</th>
<th>sunny</th>
<th>cloudy</th>
<th>rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>150/365</td>
<td>40/365</td>
<td>5/365</td>
</tr>
<tr>
<td>cold</td>
<td>50/365</td>
<td>60/365</td>
<td>60/365</td>
</tr>
</tbody>
</table>

\[
\sum 200/365 \quad 100/365 \quad 65/365
\]

\[
P(\text{Weather}) = \left\{ \frac{200}{365}, \frac{100}{365}, \frac{65}{365} \right\}
\]

- “Marginalize” comes from old practice of writing sums in margin
Marginal Probability (cont.)

- Marginalize = Sum over “other” variables
- Now marginalize over Weather:

<table>
<thead>
<tr>
<th></th>
<th>sunny</th>
<th>cloudy</th>
<th>rainy</th>
<th>(\sum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>150/365</td>
<td>40/365</td>
<td>5/365</td>
<td>195/365</td>
</tr>
<tr>
<td>cold</td>
<td>50/365</td>
<td>60/365</td>
<td>60/365</td>
<td>170/365</td>
</tr>
</tbody>
</table>

\[ P(T_{\text{Temp}}) = \left\{ \frac{195}{365}, \frac{170}{365} \right\} \]

- This is nothing but \( P(B) = \sum_{i=1..k} P(B \land A = a_i) \) if \( A \) can take \( k \) values
Conditional Probability

- $P(A = a \mid B = b)$: fraction of times $A = a$ within the region that $B = b$, or given that $B = b$

$P(A = a)$: e.g. $P(1st \text{ word} = \text{“San”}) = 0.001$

$P(B = b)$: e.g. $P(2nd \text{ word} = \text{“Francisco”}) = 0.0008$

$P(A = a \mid B = b)$: e.g. $P(1st = \text{“San”} \mid 2nd = \text{“Francisco”}) = 0.875$

Although both “San” and “Francisco” are rare, given “Francisco”, “San” quite likely!
In general, conditional probability is defined as:

\[ P(A = a \mid B) = \frac{P(A = a, B)}{P(B)} = \frac{P(A = a, B)}{\sum_{a_i} P(A = a_i, B)} \]

We can have everything conditioned on some other events C, to get a conditional version of conditional probability:

\[ P(A \mid B, C) = \frac{P(A, B \mid C')}{P(B \mid C)} \]

This should be read as \( P(A \mid (B, C')) \)
The Chain Rule

- From the definition of conditional probability we get the chain rule:

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

- It works for more than two items too:

\[
P(A_1, A_2, \ldots, A_n) = \\
P(A_1) \, P(A_2 \mid A_1) \, P(A_3 \mid A_1, A_2) \, \ldots \, P(A_n \mid A_1, A_2, \ldots, A_{n-1})
\]
Reasoning

- How do we use probabilities in A.I.?
- Example:
  - You wake up with a headache
  - Do you have the flue?
  - $H =$ headache, $F =$ flu

- **Logical Inference:** if $H$ then $F$. (world often not this clear)

- **Statistical Inference:** compute probability of a query given (or conditioned on) evidence, i.e. $P(F \mid H)$
Inference with Bayes’ Rule: Example 1

- Inference: compute the probability of a query given evidence
- $H =$ have headache, $F =$ have flu
- You know that:
  
  $P(H) = 0.1$  
  “1 in 10 people has a headache”

  $P(F) = 0.01$  
  “1 in 100 people has the flu”

  $P(H \mid F) = 0.9$  
  “90% of people who have flu have headache”

- How likely is it that you have the flu?
  
  - 0.9?
  
  - 0.01?
  
  - ...?
Inference with Bayes’ Rule: Example 1 (cont.)

Bayes Rule

in Essay Towards Solving a Problem in the Doctrine of Chances (1764)

\[
P(F \mid H) = \frac{P(F, H)}{P(H)} = \frac{P(H \mid F)P(F)}{P(H)}
\]

Using:

\[
P(H) = 0.1 \quad \text{“1 in 10 people has a headache”}
\]
\[
P(F) = 0.01 \quad \text{“1 in 100 people has the flu”}
\]
\[
P(H \mid F) = 0.9 \quad \text{“90% of people who have flu have headache”}
\]

We find:

\[
P(F \mid H) = \frac{0.9 \times 0.01}{0.1} = 0.09
\]

- So there’s a 9% chance you have the flu – much less than 90%
- But it’s higher than \( P(F) = 1\% \), since you have a headache
Inference with Bayes’ Rule (cont.)

- **Bayes Rule**

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

- **Why make things so complicated?**
  - Often \( P(B \mid A), P(A) \) and \( P(B) \) are easier to get

- **Some terms:**
  - **prior** \( P(A) \): probability before any evidence
  - **likelihood** \( P(B \mid A) \): assuming A, how likely is evidence
  - **posterior** \( P(A \mid B) \): conditional prob. after knowing evidence
  - **inference**: deriving unknown probs. from known ones

- **In general, if we have full joint prob. table, we can simply do:**

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

more on this later . . .
Inference with Bayes’ Rule: Example 2

- There are two identical-looking envelopes
  - one has a red coin (worth $100) and a black coin (worth $0)
  - the other has two black coins

- You randomly grab an envelope and randomly pick out one coin - it’s black
- You’re then given the chance to switch envelopes:
  Should you?
Inference with Bayes’ Rule: Example 2 (cont.)

- \( E \): envelope, 1=(R,B), 2=(B,B)
- \( B \): event of drawing a black coin

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

- We want to compare \( P(E = 1 \mid B) \) vs. \( P(E = 2 \mid B) \)
- \( P(B \mid E = 1) = 0.5, P(B \mid E = 2) = 1 \)
- \( P(E = 1) = P(E = 2) = 0.5 \)
- \( P(B) = \frac{3}{4} \) (and in fact we don’t need this for the comparison)
- \( P(E = 1 \mid B) = \frac{1}{3}, P(E = 2 \mid B) = \frac{2}{3} \)
- After seeing a black coin, the posterior probability of the this envelope being 1 (worth $100) is smaller than it being 2
- You should switch!
Two events $A$, $B$ are independent if: (the following are equivalent)

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

For a fair 4-sided die, let

- $A = \text{outcome is small } \{1,2\}$
- $B = \text{outcome is even } \{2,4\}$

Are $A$ and $B$ independent?

How about for a fair 6-sided die?
Independence (cont.)

- Independence can be domain knowledge

- If $A$, $B$ are independent, the joint probability table is simple:
  - it has $k^2$ cells, but only $2k - 2$ parameters
    - This is good news – more on this later . . .

- Example: $P(\text{burglary}) = 0.001$, $P(\text{earthquake}) = 0.002$.
  - Let’s say they are independent.
  - The full joint probability table $= ?$
Independence Misused

A famous statistician would never travel by airplane, because he had studied air travel and estimated that the probability of there being a bomb on any given flight was one in a million, and he was not prepared to accept these odds.

One day, a colleague met him at a conference far from home.
"How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:one million, then the odds of two bombs are \((1/1,000,000) \times (1/1,000,000)\). This is a very, very small probability, which I can accept. So now I bring my own bomb along!"

An old math joke
Conditional Independence

- Random variables can be dependent, but still **conditionally independent**

- Example: Your house has an alarm
  - Neighbor John will call when he hears the alarm
  - Neighbor Mary will call when she hears the alarm
  - Assume John and Mary don’t talk to each other

- Is JohnCall independent of MaryCall?
  - No – if John calls, it’s likely that the alarm went off, which increases the likelihood that Mary will call
  - \( P(\text{MaryCall} \mid \text{JohnCall}) \neq P(\text{MaryCall}) \)
But, if we know status of the alarm, JohnCall won’t affect MaryCall

\[ P(\text{MaryCall} \mid \text{JohnCall}, \text{Alarm}) = P(\text{MaryCall} \mid \text{Alarm}) \]

We say JohnCall and MaryCall are conditionally independent, given Alarm

In general \( A, B \) are conditionally independent given \( C \) if:

\[ P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C), \text{ or} \]
\[ P(A \mid B, C) = P(A \mid C), \text{ or} \]
\[ P(B \mid A, C) = P(B \mid C) \]