Basic Probability and Statistics

Yingyu Liang
yliang@cs.wisc.edu

Computer Sciences Department
University of Wisconsin, Madison

[based on slides from Jerry Zhu, Mark Craven]
Reasoning with Uncertainty

• There are two identical-looking envelopes
  ▪ one has a red ball (worth $100) and a black ball
  ▪ one has two black balls. Black balls worth nothing

• You randomly grabbed an envelope, randomly took out one ball – it’s black.

• At this point you’re given the option to switch the envelope. **To switch or not to switch?**
Outline

• Probability
  ▪ random variable
  ▪ 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 head?
    • Will bird flu strike tomorrow?

• Probability is the language of uncertainty
  ▪ Central pillar of modern day artificial intelligence
Sample space

• A space of outcomes that we assign probabilities to
• Outcomes can be binary, multi-valued, or continuous
• Outcomes are mutually exclusive
• Examples
  ▪ Coin flip: \{\text{head, tail}\}
  ▪ Die roll: \{1,2,3,4,5,6\}
  ▪ English words: a dictionary
  ▪ Temperature tomorrow: \mathbb{R}_+ (\text{kelvin})
Random variable

• A variable, \( x \), whose domain is the sample space, and whose value is somewhat uncertain

• Examples:
  ▪ \( x = \) coin flip outcome
  ▪ \( x = \) first word in tomorrow’s headline news
  ▪ \( x = \) tomorrow’s temperature

• Kind of like \( x = \) rand()
Probability for discrete events

- Probability $P(x=a)$ is the fraction of times $x$ takes value $a$
- Often we write it as $P(a)$
- There are other definitions of probability, and philosophical debates… but we’ll not go there
- Examples
  - $P(\text{head})=P(\text{tail})=0.5$ fair coin
  - $P(\text{head})=0.51$, $P(\text{tail})=0.49$ slightly biased coin
  - $P(\text{head})=1$, $P(\text{tail})=0$ Jerry’s coin
  - $P(\text{first word }= \text{“the” } \text{when flipping to a random page in NYT})=?$
- Demo: Search “The Book of Odds”
## Probability table

- **Weather**

<table>
<thead>
<tr>
<th></th>
<th>Sunny</th>
<th>Cloudy</th>
<th>Rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P</strong></td>
<td>200/365</td>
<td>100/365</td>
<td>65/365</td>
</tr>
</tbody>
</table>

- $P(\text{Weather} = \text{sunny}) = P(\text{sunny}) = \frac{200}{365}$

- $P(\text{Weather}) = \{\frac{200}{365}, \frac{100}{365}, \frac{65}{365}\}$

- For now we’ll be satisfied with obtaining the probabilities by counting frequency from data...
Probability for discrete events

• Probability for more complex events A

  - $P(A= \text{“head or tail”})=?$ fair coin
  - $P(A= \text{“even number”})=?$ fair 6-sided die
  - $P(A= \text{“two dice rolls sum to 2”})=?$
Probability for discrete events

• Probability for more complex events $A$

  - $P(A=\text{“head or tail”}) = 0.5 + 0.5 = 1$ fair coin

  - $P(A=\text{“even number”}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5$ fair 6-sided die

  - $P(A=\text{“two dice rolls sum to 2”}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$
The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

The fraction of A can’t be smaller than 0
The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

Sample space

The fraction of $A$ can’t be bigger than 1
The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

Valid sentence: e.g. “$x=$head or $x=$tail”
The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

Invalid sentence: e.g. “$x=$head AND $x=$tail”
The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Some theorems derived from the axioms

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

- If \( A \) can take \( k \) different values \( a_1, \ldots, a_k \):
  \[
  P(A=a_1) + \cdots + P(A=a_k) = 1
  \]

- \( P(B) = P(B \land \neg A) + P(B \land A) \), if \( A \) is a binary event

- \( P(B) = \sum_{i=1}^{k} P(B \land A=a_i) \), if \( A \) can take \( k \) values
Joint probability

- The joint probability $P(A=a, B=b)$ is a shorthand for $P(A=a \land B=b)$, the probability of both $A=a$ and $B=b$ happen

$P(A=a)$, e.g. $P(1^{\text{st}} \text{ word on a random page } = \text{“San”}) = 0.001$

(possibly: San Francisco, San Diego, …)

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

(possibly: San Francisco, Don Francisco, Pablo Francisco …)

$P(A=a, B=b)$, e.g. $P(1^{\text{st}} = \text{“San”}, 2^{\text{nd}} = \text{“Francisco”}) = 0.0007$
### Joint probability table

- P(temp=hot, weather=rainy) = P(hot, rainy) = \( \frac{5}{365} \)

- The full joint probability table between N variables, each taking k values, has \( k^N \) entries (that’s a lot!)

<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>
Marginal probability

• Sum over other variables

<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>

\[ \sum = \frac{200}{365} \quad \frac{100}{365} \quad \frac{65}{365} \]

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

• The name comes from the old days when the sums are written on the margin of a page
Marginal probability

- Sum over other variables

\[
P(B) = \sum_{i=1}^{k} P(B \land A=a_i), \text{ if } A \text{ can take } k \text{ values}
\]

<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>

\[
P(\text{temp}) = \{195/365, 170/365\}
\]
Conditional probability

- The conditional probability $P(A=a | B=b)$ is the fraction of times $A=a$, within the region that $B=b$

$P(A=a)$, e.g. $P(1^{st} \text{ word on a random page } = \text{“San”}) = 0.001$

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

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

(possibly: San, Don, Pablo …)

Although “San” is rare and “Francisco” is rare, given “Francisco” then “San” is quite likely!
Conditional probability

- $P(\text{San} \mid \text{Francisco})$
  
  $= \frac{\#(1^{\text{st}}=\text{S} \text{ and } 2^{\text{nd}}=\text{F})}{\#(2^{\text{nd}}=\text{F})}$
  
  $= \frac{P(\text{San} \wedge \text{Francisco})}{P(\text{Francisco})}$
  
  $= 0.0007 / 0.0008$
  
  $= 0.875$

$P(\text{S}) = 0.001$

$P(\text{F}) = 0.0008$

$P(\text{S,F}) = 0.0007$

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

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

(possibly: San, Don, Pablo …)
Conditional probability

• In general, the conditional probability is

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

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

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

‘|’ has low precedence. This should read \( P(A \mid (B,C)) \)
The chain rule

• From the definition of conditional probability we have the chain rule

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

• It works the other way around

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

• It works with more than 2 events too

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

How do we use probabilities in AI?

• You wake up with a headache (D’oh!).
• Do you have the flu?
• \( H = \) headache, \( F = \) flu

**Logical** Inference: if \( (H) \) then \( F \). (but the world is often not this clear cut)

**Statistical** Inference: compute the probability of a query given (conditioned on) evidence, \( i.e. \ P(F|H) \)

[Example from Andrew Moore]
Inference with Bayes’ rule: Example 1

Inference: compute the probability of a query given evidence
(H = headache, F = flu)

You know that

- \( P(H) = 0.1 \) “one in ten people has headache”
- \( P(F) = 0.01 \) “one in 100 people has flu”
- \( P(H|F) = 0.9 \) “90% of people who have flu have headache”

- How likely do you have the flu?
  - 0.9?
  - 0.01?
  - …?
Inference with Bayes’ rule

Bayes’ rule

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

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

- \( P(H) = 0.1 \)  “one in ten people has headache”
- \( P(F) = 0.01 \)  “one in 100 people has flu”
- \( P(H|F) = 0.9 \)  “90% of people who have flu have headache”

- \( P(F|H) = 0.9 \times 0.01 / 0.1 = 0.09 \)
- So there’s a 9% chance you have flu – much less than 90%
- But it’s higher than \( P(F)=1\% \), since you have the headache
Inference with Bayes’ rule

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$  
  Bayes’ rule
- Why do we make things this complicated?
  - Often $P(B|A)$, $P(A)$, $P(B)$ are easier to get
  - Some names:
    - **Prior $P(A)$**: probability before any evidence
    - **Likelihood $P(B|A)$**: assuming $A$, how likely is the evidence
    - **Posterior $P(A|B)$**: conditional prob. after knowing evidence
    - **Inference**: deriving unknown probability from known ones
- In general, if we have the full joint probability table, we can simply do $P(A|B)=P(A, B) / P(B)$ – more on this later…
Inference with Bayes’ rule: Example 2

• In a bag there are two envelopes
  ▪ one has a red ball (worth $100) and a black ball
  ▪ one has two black balls. Black balls worth nothing

• You randomly grabbed an envelope, randomly took out one ball – it’s black.
• At this point you’re given the option to switch the envelope. To switch or not to switch?
Inference with Bayes’ rule: Example 2

• E: envelope, 1=(R,B), 2=(B,B)
• B: the event of drawing a black ball
• \( P(E|B) = \frac{P(B|E)*P(E)}{P(B)} \)
• We want to compare \( P(E=1|B) \) vs. \( P(E=2|B) \)
Inference with Bayes’ rule: Example 2

- E: envelope, 1=(R,B), 2=(B,B)
- B: the event of drawing a black ball
- \( P(E|B) = \frac{P(B|E)\cdot P(E)}{P(B)} \)
- We want to compare \( P(E=1|B) \) vs. \( P(E=2|B) \)
- \( P(B|E=1) = 0.5, P(B|E=2) = 1 \)
- \( P(E=1)=P(E=2)=0.5 \)
- \( P(B)=3/4 \) (it in fact doesn’t matter for the comparison)
Inference with Bayes’ rule: Example 2

- E: envelope, 1 = (R,B), 2 = (B,B)
- B: the event of drawing a black ball
- \( P(E|B) = \frac{P(B|E) \cdot P(E)}{P(B)} \)
- We want to compare \( P(E=1|B) \) vs. \( P(E=2|B) \)
- \( P(B|E=1) = 0.5, P(B|E=2) = 1 \)
- \( P(E=1)=P(E=2)=0.5 \)
- \( P(B)=\frac{3}{4} \) (it in fact doesn’t matter for the comparison)
- \( P(E=1|B)=\frac{1}{3}, P(E=2|B)=\frac{2}{3} \)
- After seeing a black ball, the posterior probability of this envelope being 1 (thus worth $100) is smaller than it being 2
- Thus you should switch
Independence

• Two events A, B are independent, if (the following are equivalent)
  ▪ $P(A, B) = P(A) \times P(B)$
  ▪ $P(A \mid B) = P(A)$
  ▪ $P(B \mid A) = P(B)$

• For a 4-sided die, let
  ▪ A=outcome is small
  ▪ B=outcome is even
  ▪ Are A and B independent?

• How about a 6-sided die?
Independence

• Independence is a domain knowledge

• If A, B are independent, the joint probability table between A, B 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=?
Conditional independence

• Random variables can be dependent, but conditionally independent

• 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

• JohnCall independent of MaryCall?
  ▪ No – If John called, likely the alarm went off, which increases the probability of Mary calling
  ▪ \( P(\text{MaryCall} \mid \text{JohnCall}) \neq P(\text{MaryCall}) \)
Conditional independence

• If we know the status of the alarm, JohnCall won’t affect Mary at all
  \[ P(\text{MaryCall} | \text{Alarm, JohnCall}) = P(\text{MaryCall} | \text{Alarm}) \]

• We say JohnCall and MaryCall are conditionally independent, given Alarm

• In general A, B are conditionally independent given C
  - if \( P(A | B, C) = P(A | C) \), or
  - \( P(B | A, C) = P(B | C) \), or
  - \( P(A, B | C) = P(A | C) \cdot P(B | C) \)
Independence example #1

Are $X$ and $Y$ independent here?
Independence example #2

<table>
<thead>
<tr>
<th>$x, y$</th>
<th>$P(X = x, Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, fly-United</td>
<td>0.27</td>
</tr>
<tr>
<td>rain, fly-United</td>
<td>0.45</td>
</tr>
<tr>
<td>snow, fly-United</td>
<td>0.18</td>
</tr>
<tr>
<td>sun, fly-Delta</td>
<td>0.03</td>
</tr>
<tr>
<td>rain, fly-Delta</td>
<td>0.05</td>
</tr>
<tr>
<td>snow, fly-Delta</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>rain</td>
<td>0.5</td>
</tr>
<tr>
<td>snow</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y$</th>
<th>$P(Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>fly-United</td>
<td>0.9</td>
</tr>
<tr>
<td>fly-Delta</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Are $X$ and $Y$ independent here?
Expected values

- The *expected value* of a random variable that takes on numerical values is defined as:

  \[
  \mathbb{E}[X] = \sum_x xP(x)
  \]

  This is the same thing as the *mean*

- We can also talk about the expected value of a function of a random variable

  \[
  \mathbb{E}[g(X)] = \sum_x g(x)P(x)
  \]
Expected value examples

• Shoesize

\[ \mathbb{E}[\text{Shoesize}] \]
\[ = 5 \times P(\text{Shoesize} = 5) + \cdots + 14 \times P(\text{Shoesize} = 14) \]

• Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

What is the expectation of the gain?
Expected value examples

• Shoesize

\[ E[\text{Shoesize}] = 5 \times P(\text{Shoesize} = 5) + \cdots + 14 \times P(\text{Shoesize} = 14) \]

• Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

\[ E[\text{gain(Lottery)}] = \text{gain(winning)}P(\text{winning}) + \text{gain(losing)}P(\text{losing}) \]
\[ = (\$100 - \$1) \times 0.001 - \$1 \times 0.999 \]
\[ = -\$0.9 \]
Summary

- Axioms of probability and related properties
- Joint/marginal/conditional probabilities
- Bayes’ rule for reasoning
- Independence and conditional independence
- Expectation