Probability

Lirong Xia

Rensselaer

Spring 2017
Today's schedule

• Probability
  – probability space
  – events
  – independence
  – conditional probability
  – Bayes’ rule

• Random variables

• Probabilistic inference
A sample space $\Omega$ – states of the world,
  - or equivalently, outcomes
  - only need to model the states that are relevant to the problem

A probability mass function $p$ over the sample space
  - for all $a \in \Omega$, $p(a) \geq 0$
  - $\sum_{a \in \Omega} p(a) = 1$

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Weather</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Events

• An event is a subset of the sample space
  – can be seen as a “property” of the states of the world
  – note: this does not need to consider the probability mass function

• An atomic event is a single outcome in $\Omega$

<table>
<thead>
<tr>
<th>$\Omega$</th>
<th>hot</th>
<th>sun</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hot</td>
<td>rain</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>sun</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>rain</td>
</tr>
</tbody>
</table>

• Atomic events:
  – $\{(\text{hot, sun}), (\text{hot, rain}), (\text{cold, sun}), (\text{cold, rain})\}$
  – note: no need to consider the probability mass function

• Event 1 (hot days): $\{(\text{hot, sun}), (\text{hot, rain})\}$

• Event 2 (sunny days): $\{(\text{hot, sun}), (\text{cold, sun})\}$
Marginal probability of an event

• Given
  – a probability space \((\Omega, p)\)
  – an event \(A\)

• The marginal probability of \(A\), denoted by \(p(A)\) is \(\sum_{a \in A} p(a)\)

• Example
  – \(p(\text{sun}) = 0.6\)
  – \(p(\text{hot}) = 0.5\)

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Weather</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Joint probability

- Given
  - a probability space
  - two events A and B
- The joint probability of A and B is $p(A \text{ and } B)$
Conditional probability

- \( p(A \mid B) = \frac{p(A \text{ and } B)}{p(B)} \)

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

- \( p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \)
  
  – Bayes’ rule
Example of conditional probability

- $A = \text{hot days}$
- $B = \text{sun days}$
- $p(A \text{ and } B) = 0.4$
- $p(A | B) = 0.4 / 0.6$
- $p(B | A) = 0.4 / 0.5$

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Weather</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Independent events

• Given
  – a probability space
  – two events A and B

• A and B are independent, if and only if
  – \( p(A \text{ and } B) = p(A)p(B) \)
  – equivalently, \( p(A|B) = p(A) \)

• Interpretation
  – Knowing that the state of the world is in B does not affect the probability that the state of the world is in A (vs. not in A)

• Different from [A and B are disjoint]
The Monty Hall problem

- Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

- After the problem appeared in Parade, approximately 10,000 readers, including nearly 1,000 with PhDs, wrote to the magazine, most of them claiming vos Savant was wrong.

- Paul Erdos remained unconvinced until he was shown a computer simulation confirming the predicted result.
As a probability problem

- Sample space:
  - (door with the car, door opened by the host)
  - \{(1,2), (1,3), (2,2), (2,3), (3,2), (3,3)\}
- Equal prob behind each door: \(p(1)=p(2)=p(3)=1/3\)
- Objective: compare \(p(1|\text{host } = 2)\) and \(p(3|\text{host } = 2)\), and \(p(1|\text{host } = 3)\) and \(p(2|\text{host } = 3)\)
- Probability mass function?
  - option 1: all 1/6
  - option 2: \(p(2,2)=p(3,3)=0, p(2,3)=p(3,2)=1/3\)
- Depends on the behavior of the host!
  - assuming he picks 2 and 3 uniformly at random, then no need to switch
  - assuming he always picks a door with a goat, then should switch
Random variables

All we need to care in this course
Random Variables

• A random variable is some aspect of the world about which we (may) have uncertainty
  – \( W \) = Is it raining?
  – \( D \) = How long will it take to drive to work?
  – \( L \) = Where am I?
  – We denote random variables with capital letters

• Like variables in a CSP, random variables have domains
  – \( W \) in \{rain, sun\}
  – \( D \) in \([0, \infty)\)
  – \( L \) in possible locations, maybe \{(0,0),{0,1},…\}

• For now let us assume that a random variable is a variable with a domain
  – Random variables: capital letters, e.g. \( W, D, L \)
  – values: small letters, e.g. \( w, d, l \)
Joint Distributions

• 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)
  - $p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
  - $p(x_1, x_2, \ldots, x_n)$
  - Size of distribution if $n$ variables with domain sizes $d$?
  - Must obey: $p(x_1, x_2, \ldots, x_n) \geq 0$
    $$\sum_{(x_1, x_2, \ldots, x_n)} p(x_1, x_2, \ldots, x_n) = 1$$

• For all but the smallest distributions, impractical to write out

• This is a probability space
  - Sample space $\Omega$: all combinations of values
  - probability mass function is $p$

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables.

- Probabilistic models:
  - (Random) variables with domains.
  - Assignments are called outcomes (atomic events).
  - Joint distributions: say whether assignments (outcomes) are likely.
  - Normalized: sum to 1.0

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Events in probabilistic models

• An event in a probabilistic model is a set $E$ of outcomes

$$ p(E) = \sum_{(x_1 \cdots x_n) \in E} p(x_1 \cdots x_n) $$

• From a joint distribution, we can calculate the probability of any event
  – Probability that it’s hot AND sunny?
  – Probability that it’s hot?
  – Probability that it’s hot OR sunny?

• Typically, the events we care about are partial assignments, like $p(T=\text{hot})$

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables.
- Marginalization (summing out): combine collapsed rows by adding:

\[ p(X_1 = x_1) = \sum_{x_2} p(X_1 = x_1, X_2 = x_2) \]

### Example

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

W: sun 0.6, rain 0.4
T: hot 0.5
   | 0.4   | 0.1   |
   | 0.2   | 0.3   |
   cold 0.5
Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun (hot)</td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>rain (hot)</td>
<td></td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun (cold)</td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>rain (cold)</td>
<td></td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot sun</td>
<td></td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>hot rain</td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>cold sun</td>
<td></td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>cold rain</td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>
Normalization Trick

• A trick to get a whole conditional distribution at once:
  – Select the joint probabilities matching the assignments (evidence)
  – Normalize the selection (make it sum to one)

\[
p(T,W)
\]

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>hot</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>sun</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>hot</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>rain</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>cold</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>sun</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>cold</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>rain</td>
<td></td>
</tr>
</tbody>
</table>

– Why does this work? Sum of selection is \( p(\text{evidence}) \)! (\( p(r) \), here)

\[
p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{p(x_1, x_2)}{\sum_{x_1} p(x_1, x_2)}
\]
Probabilistic Inference

• Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

• We generally compute conditional probabilities
  – $p(\text{on time} \mid \text{no reported accidents}) = 0.9$
  – 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
Inference by Enumeration

- \( p(\text{sun})? \)
- \( p(\text{sun} \mid \text{winter})? \)
- \( p(\text{sun} \mid \text{winter}, \text{warm})? \)

<table>
<thead>
<tr>
<th>S</th>
<th>T</th>
<th>W</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>summer</td>
<td>hot</td>
<td>sun</td>
<td>0.30</td>
</tr>
<tr>
<td>summer</td>
<td>hot</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>sun</td>
<td>0.15</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>rain</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Inference by Enumeration

• General case:
  – Evidence variables: \( E_1 \cdots E_k = e_1 \cdots e_k \)
  – Query variable: \( Q \)
  – Hidden variable: \( H_1 \cdots H_r \)

• We want: \( p(Q|e_1 \cdots e_k) \)

• First, select the entries consistent with the evidence
• Second, sum out H to get joint of query and evidence:
  \[
p(Q,e_1 \cdots e_k) = \sum_{h_1 \cdots h_r} p(Q,h_1 \cdots h_r,e_1 \cdots e_k)
\]

• Finally, normalize the remaining entries
• Obvious problems:
  – Worst-case time complexity \( O(d^n) \)
  – Space complexity \( O(d^n) \) to store the joint distribution
The Product Rule

- Sometimes have conditional distributions but want the joint

\[ p(x|y) = \frac{p(x, y)}{p(y)} \iff p(x, y) = p(x|y)p(y) \]

- Example:

\[
\begin{array}{c|c|c}
W & p & p(D|W) \\
\hline
\text{Wet} & \text{Sun} & 0.1 \\
\text{Dry} & \text{Sun} & 0.9 \\
\text{Wet} & \text{Rain} & 0.7 \\
\text{Dry} & \text{Rain} & 0.3 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c}
D & W & p & p(D, W) \\
\hline
\text{Wet} & \text{Sun} & 0.08 \\
\text{Dry} & \text{Sun} & 0.72 \\
\text{Wet} & \text{Rain} & 0.14 \\
\text{Dry} & \text{Rain} & 0.06 \\
\end{array}
\]
The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

\[
p(x_1, x_2, x_3) = p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1, x_2)
\]

\[
p(x_1, x_2, \ldots, x_n) = \prod_{i} p(x_i \mid x_1 \ldots x_{i-1})
\]

- Why is this always true?
Bayes’ Rule revisited

- Two ways to factor a joint distribution over two variables:
  \[ p(x, y) = p(x|y)p(y) = p(y|x)p(x) \]

- Dividing, we get:
  \[ p(x|y) = \frac{p(y|x)}{p(y)} p(x) \]

- Why is this helpful?
  - Update belief (X) based on evidence y
  - when \( p(y|x) \) is much easier to compute
Inference with Bayes’ Rule

- Example: diagnostic probability from causal probability:

\[ p(Cause|Effect) = \frac{p(Effect|Cause)p(Cause)}{p(Effect)} \]

- Example:
  - F is fire, \{f, \neg f\}  
  - A is the alarm, \{a, \neg a\}  
  - \( p(f) = 0.01 \)  
  - \( p(a) = 0.1 \)  
  - \( p(a|f) = 0.9 \)  

\[ p(f|a) = \frac{p(a|f)p(f)}{p(a)} = \frac{0.9 \times 0.001}{0.1} = 0.009 \]

- Note: posterior probability of fire still very small
- Note: you should still run when hearing an alarm! Why?
Independence

• Two variables are *independent* in a joint distribution if for all $x, y$, the events $X=x$ and $Y=y$ are independent:

$$p(X, Y) = p(X) p(Y)$$

$$\forall x, y \ p(x, y) = p(x) p(y)$$

– The joint distribution factors into a product of two simple ones
– Usually variables aren’t independent!

• Can use independence as a modeling assumption
  – Independence can be a simplifying assumption
  – What could we assume for {Weather, Traffic, stock price}?
Mathematical definition of Random variables

- Just for your curiosity
- Mathematically, given a sample space $\Omega$, a random variable is a function $X: \Omega \rightarrow S$
  - $S$ is the domain of $X$
  - for any $s \in S$, $X^{-1}(s)$ is an event

Example 1
- $W$: weather
  - $S_w = \{\text{rain, sun}\}$
  - $W^{-1}(\text{rain}) = \{(\text{hot, rain}), (\text{cold, rain})\}$
  - $W^{-1}(\text{sun}) = \{(\text{hot, sun}), (\text{cold, sun})\}$

Example 2
- $T$: temperature
  - $S_T = \{\text{hot, cold}\}$
  - $T^{-1}(\text{hot}) = \{(\text{hot, sun}), (\text{hot, rain})\}$
  - $T^{-1}(\text{cold}) = \{(\text{cold, sun}), (\text{cold, rain})\}$
What is probability, anyway?

Different philosophical positions:

- **Frequentism**: numbers only come from repeated experiments
  - As we flip a coin lots of times, we see experimentally that it comes out heads $\frac{1}{2}$ the time
  - Problem: “No man ever steps in the same river twice”
    - Probability that the Democratic candidate wins the next election?

- **Objectivism**: probabilities are a real part of the universe
  - Maybe true at level of quantum mechanics
  - Most of us agree that the result of a coin flip is (usually) determined by initial conditions + classical mechanics

- **Subjectivism**: probabilities merely reflect agents’ beliefs
Recap

• Probability
  – which we still do not know

• Random variable
  – which will be our main focus

• Probabilistic inference
  – more in the next class (next Friday)
  – Reassigned project 1 due