CS 331: Artificial Intelligence
Fundamentals of Probability II

Thanks to Andrew Moore for some course material

Full Joint Probability Distributions

<table>
<thead>
<tr>
<th>Coin</th>
<th>Card</th>
<th>Candy</th>
<th>P(Coin, Card, Candy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tails</td>
<td>black</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>tails</td>
<td>black</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>tails</td>
<td>black</td>
<td>3</td>
<td>0.09</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>3</td>
<td>0.12</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>1</td>
<td>0.075</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>3</td>
<td>0.045</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>1</td>
<td>0.035</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>2</td>
<td>0.105</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>3</td>
<td>0.21</td>
</tr>
</tbody>
</table>

This cell means \( P(\text{Coin=heads, Card=red, Candy=3}) = 0.21 \)

The probabilities in the last column sum to 1
Joint Probability Distribution

From the full joint probability distribution, we can calculate any probability involving these three random variables.

e.g. \( P(\text{Coin} = \text{heads OR Card} = \text{red}) \)

\[
P(\text{Coin} = \text{heads OR Card} = \text{red}) = P(\text{Coin=heads, Card=black, Candy=1}) + P(\text{Coin=heads, Card=black, Candy=2}) + P(\text{Coin=heads, Card=black, Candy=3}) + P(\text{Coin=tails, Card=red, Candy=1}) + P(\text{Coin=tails, Card=red, Candy=2}) + P(\text{Coin=tails, Card=red, Candy=3}) + P(\text{Coin=heads, Card=red, Candy=1}) + P(\text{Coin=heads, Card=red, Candy=2}) + P(\text{Coin=heads, Card=red, Candy=3})
\]

\[
= 0.075 + 0.03 + 0.045 + 0.02 + 0.06 + 0.12 + 0.035 + 0.105 + 0.21 = 0.7
\]
Marginalization

We can even calculate marginal probabilities (the probability distribution over a subset of the variables) e.g.:

\[
P(\text{Coin}=\text{tails}, \text{Card}=\text{red}) = \\
P(\text{Coin}=\text{tails}, \text{Card}=\text{red}, \text{Candy}=1) + \\
P(\text{Coin}=\text{tails}, \text{Card}=\text{red}, \text{Candy}=2) + \\
P(\text{Coin}=\text{tails}, \text{Card}=\text{red}, \text{Candy}=3) \\
= 0.02 + 0.06 + 0.12 = 0.2
\]

Or even:

\[
P(\text{Card}=\text{black}) = \\
P(\text{Coin}=\text{heads}, \text{Card}=\text{black}, \text{Candy}=1) + \\
P(\text{Coin}=\text{heads}, \text{Card}=\text{black}, \text{Candy}=2) + \\
P(\text{Coin}=\text{heads}, \text{Card}=\text{black}, \text{Candy}=3) + \\
P(\text{Coin}=\text{tails}, \text{Card}=\text{black}, \text{Candy}=1) + \\
P(\text{Coin}=\text{tails}, \text{Card}=\text{black}, \text{Candy}=2) + \\
P(\text{Coin}=\text{tails}, \text{Card}=\text{black}, \text{Candy}=3) \\
= 0.075 + 0.03 + 0.045 + 0.015 + 0.06 + 0.09 = 0.315
\]
Marginalization

The general marginalization rule for any sets of variables $Y$ and $Z$:

$$P(Y) = \sum_z P(Y, z)$$

or

$$P(Y) = \sum_z P(Y \mid z) P(z)$$

$z$ is over all possible combinations of values of $Z$ (remember $Z$ is a set)

Marginalization

For continuous variables, marginalization involves taking the integral:

$$P(Y) = \int P(Y, z) dz$$
Practice

Compute \( P(Candy = 2) \).

<table>
<thead>
<tr>
<th>Coin</th>
<th>Card</th>
<th>Candy</th>
<th>( P(Coin, \text{Card, Candy}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>tails</td>
<td>black</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>tails</td>
<td>black</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>tails</td>
<td>black</td>
<td>3</td>
<td>0.09</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>3</td>
<td>0.12</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>1</td>
<td>0.075</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>3</td>
<td>0.045</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>1</td>
<td>0.035</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>2</td>
<td>0.105</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>3</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Conditional Probabilities

We can also compute conditional probabilities from the joint. Recall:

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

\[ P(Coin = \text{heads}|Card = \text{black}) \]
\[ = \frac{P(Coin=\text{heads},Card=\text{black})}{P(Card=\text{black})} \]
\[ = \frac{0.075+0.03+0.045}{0.15+0.06+0.09+0.075+0.03+0.045} = 0.333 \]

\[ P(Coin = \text{tails}|Card = \text{black}) \]
\[ = \frac{P(Coin=\text{tails},Card=\text{black})}{P(Card=\text{black})} \]
\[ = \frac{0.15+0.06+0.09}{0.15+0.06+0.09+0.075+0.03+0.045} = 0.667 \]

Note that \(1/P(Card=\text{black})\) remains constant in the two equations.
Normalization

- In fact, $1/P(\text{Card})$ can be viewed as a normalization constant for $P(\text{Coin} | \text{Card})$, ensuring it adds up to 1
- We will refer to normalization constants with the symbol $\alpha$

\[ P(\text{Coin}|\text{black}) = \alpha P(\text{Coin}, \text{black}) \]

Practice

Compute $P(\text{Candy} = 1 | \text{Card} = \text{red})$.

<table>
<thead>
<tr>
<th>Coin</th>
<th>Card</th>
<th>Candy</th>
<th>P(Coin, Card, Candy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tails</td>
<td>black</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>tails</td>
<td>black</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>tails</td>
<td>black</td>
<td>3</td>
<td>0.09</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>tails</td>
<td>red</td>
<td>3</td>
<td>0.12</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>1</td>
<td>0.075</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>heads</td>
<td>black</td>
<td>3</td>
<td>0.045</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>1</td>
<td>0.035</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>2</td>
<td>0.105</td>
</tr>
<tr>
<td>heads</td>
<td>red</td>
<td>3</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Inference

• Suppose you get a query such as
  \[ P(Card = \text{red} \mid Coin = \text{heads}) \]

  *Coin* is called the evidence variable because we observe it. More generally, it’s a set of variables.

  *Card* is called the query variable (we’ll assume it’s a single variable for now).

  There are also unobserved (aka hidden) variables like *Candy*.

Inference

• We will write the query as \( P(X \mid e) \)

  This is a probability distribution hence the boldface.

  \[
  X = \text{Query variable (a single variable for now)} \\
  E = \text{Set of evidence variables} \\
  e = \text{the set of observed values for the evidence variables} \\
  Y = \text{Unobserved variables}
  \]
Inference

We will write the query as \( P(X | e) \)

\[
P(X | e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)
\]

Summation is over all possible combinations of values of the unobserved variables \( Y \)

\( X = \) Query variable (a single variable for now)
\( E = \) Set of evidence variables
\( e = \) the set of observed values for the evidence variables
\( Y = \) Unobserved variables

---

Inference

\[
P(X | e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)
\]

Computing \( P(X | e) \) involves going through all possible entries of the full joint probability distribution and adding up probabilities with \( X=x_i, E=e, \) and \( Y=y \)

Suppose you have a domain with \( n \) Boolean variables. What is the space and time complexity of computing \( P(X | e) \)?
Independence

• How do you avoid the exponential space and time complexity of inference?
• Use independence (aka factoring)

We say that variables $X$ and $Y$ are independent if any of the following hold:
(note that they are all equivalent)

$P(X \mid Y) = P(X)$  or  
$P(Y \mid X) = P(Y)$  or  
$P(X, Y) = P(X)P(Y)$
Independence

Consider the full joint distribution over these variables:

Card = \{red, black\}

Candy = \{1,2,3\}

By the product rule, we know:

\[ P(\text{Card, Candy}) \]
\[ = P(\text{Card} | \text{Candy}) P(\text{Candy}) \]

Independence

Suppose I tell you that these two events are independent (i.e. they do not influence each other).

Then:

\[ P(\text{Card, Candy}) \]
\[ = P(\text{Card} | \text{Candy}) P(\text{Candy}) \]
\[ = P(\text{Card}) P(\text{Candy}) \]
Why is independence useful?

\[ P(\text{Card, Candy}) = P(\text{Card})P(\text{Candy}) \]

- You now need to store 5 values to calculate \( P(\text{Coin, Card, Candy}) \)
- Without independence, we needed 6

Independence

Another example:

- Suppose you have \( n \) coin flips and you want to calculate the joint distribution \( P(C_1, \ldots, C_n) \)
- If the coin flips are not independent, you need \( 2^n \) values in the table
- If the coin flips are independent, then

\[ P(C_1, \ldots, C_n) = \prod_{i=1}^{n} P(C_i) \]

Each \( P(C_i) \) table has 2 entries and there are \( n \) of them for a total of \( 2n \) values
Independence

- Independence is powerful!
- It required extra domain knowledge. A different kind of knowledge than numerical probabilities. It needed an understanding of relationships among the random variables.

Practice

Are Coin and Card independent in this distribution?

Recall:

\[ P(X \mid Y) = P(X) \]
\[ P(Y \mid X) = P(Y) \]

\[ P(X, Y) = P(X)P(Y) \]

for independent \( X \) and \( Y \)