CS 331: Artificial Intelligence
Fundamentals of Probability II

Thanks to Andrew Moore for some course material

**Full Joint Probability Distributions**

<table>
<thead>
<tr>
<th>Toothache</th>
<th>Cavity</th>
<th>Catch</th>
<th>P(Toothache, Cavity, Catch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>0.576</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
<td>0.144</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>0.008</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>true</td>
<td>0.072</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>0.064</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>true</td>
<td>0.016</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>0.012</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>0.108</td>
</tr>
</tbody>
</table>

This cell means $P(\text{Toothache} = \text{true}, \text{Cavity} = \text{true}, \text{Catch} = \text{true}) = 0.108$

"Catch" means the dentist’s probe catches in my teeth

The probabilities in the last column sum to 1

**Joint Probability Distribution**

From the full joint probability distribution, we can calculate any probability involving the three random variables in this world e.g.

$$P(\text{Toothache} = \text{true} \text{ OR } \text{Cavity} = \text{true}) =$$

$$P(\text{Toothache}=\text{true}, \text{Cavity}=\text{false}, \text{Catch}=\text{false}) +$$

$$P(\text{Toothache}=\text{true}, \text{Cavity}=\text{false}, \text{Catch}=\text{true}) +$$

$$P(\text{Toothache}=\text{false}, \text{Cavity}=\text{true}, \text{Catch}=\text{false}) +$$

$$P(\text{Toothache}=\text{false}, \text{Cavity}=\text{true}, \text{Catch}=\text{true}) +$$

$$= 0.064 + 0.016 + 0.008 + 0.072 + 0.012 + 0.108 = 0.28$$

**Marginalization**

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

$$P(\text{Toothache} = \text{true}, \text{Cavity} = \text{true}) =$$

$$P(\text{Toothache} = \text{true}, \text{Cavity} = \text{true}, \text{Catch} = \text{true}) +$$

$$P(\text{Toothache} = \text{true}, \text{Cavity} = \text{true}, \text{Catch} = \text{false}) +$$

$$= 0.108 + 0.012 = 0.12$$

**Marginalization**

Or even:

$$P(\text{Cavity} = \text{true}) =$$

$$P(\text{Toothache} = \text{true}, \text{Cavity} = \text{true}, \text{Catch} = \text{true}) +$$

$$P(\text{Toothache} = \text{true}, \text{Cavity} = \text{true}, \text{Catch} = \text{false}) +$$

$$P(\text{Toothache} = \text{false}, \text{Cavity} = \text{true}, \text{Catch} = \text{true}) +$$

$$P(\text{Toothache} = \text{false}, \text{Cavity} = \text{true}, \text{Catch} = \text{false}) +$$

$$= 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$
Marginalization

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

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

or

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

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

Marginalization

For continuous variables, marginalization involves taking the integral:

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

Normalization

$$P(\text{Cavity} = \text{true} \mid \text{Toothache} = \text{true}) = \frac{P(\text{Cavity} = \text{true}, \text{Toothache} = \text{true})}{P(\text{Toothache} = \text{true})} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

$$P(\text{Cavity} = \text{false} \mid \text{Toothache} = \text{true}) = \frac{P(\text{Cavity} = \text{false}, \text{Toothache} = \text{true})}{P(\text{Toothache} = \text{true})} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Note that $1/P(\text{Toothache}=\text{true})$ remains constant in the two equations.

Normalization

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

$$P(\text{Cavity} \mid \text{toothache}) = \alpha P(\text{Cavity}, \text{toothache})$$

Inference

• Suppose you get a query such as $P(\text{Cavity} = \text{true} \mid \text{Toothache} = \text{true})$

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

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

There are also unobserved (aka hidden) variables like Catch

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 $P(X \mid e)$

$$P(X \mid e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)$$

$X = \text{Query variable (a single variable for now)}$

$E = \text{Set of evidence variables}$

$e = \text{the set of observed values for the evidence variables}$

$Y = \text{Unobserved variables}$

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

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

Independence

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

Independence

Suppose the full joint distribution now consists of four variables:

$\text{Toothache} = \{\text{true, false}\}$

$\text{Catch} = \{\text{true, false}\}$

$\text{Cavity} = \{\text{true, false}\}$

$\text{Weather} = \{\text{sunny, rain, cloudy, snow}\}$

There are now 32 entries in the full joint distribution table

Independence

Does the weather influence one’s dental problems?

Is $P(\text{Weather}=\text{cloudy} \mid \text{Toothache} = \text{toothache}, \text{Catch} = \text{catch}, \text{Cavity} = \text{cavity}) = P(\text{Weather}=\text{cloudy})$?

In other words, is $\text{Weather}$ independent of $\text{Toothache}, \text{Catch}$ and $\text{Cavity}$?

Independence

We say that variables $X$ and $Y$ are independent if any of the following hold:

(note that they are all equivalent)

$$P(X \mid Y) = P(X) \quad \text{or}$$

$$P(Y \mid X) = P(Y) \quad \text{or}$$

$$P(X, Y) = P(X)P(Y)$$
Why is independence useful?

Assume that Weather is independent of toothache, catch, cavity i.e.
\[ P(\text{Weather=cloudy} \mid \text{Toothache = toothache, Catch = catch, Cavity = cavity}) = P(\text{Weather=cloudy}) \]

Now we can calculate:
\[ P(\text{Weather=cloudy, Toothache = toothache, Catch = catch, Cavity = cavity}) = P(\text{Weather=cloudy} \mid \text{Toothache = toothache, Catch = catch, Cavity = cavity}) \times P(\text{toothache, catch, cavity}) \]
\[ = P(\text{Weather=cloudy}) \times P(\text{Toothache = toothache, Catch = catch, Cavity = cavity}) \]

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

Bayes’ Rule

The product rule can be written in two ways:
\[ P(A, B) = P(A \mid B)P(B) \]
\[ P(A, B) = P(B \mid A)P(A) \]

You can combine the equations above to get:
\[ P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \]

Bayes’ Rule

More generally, the following is known as Bayes’ Rule:
\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]

Note that these are distributions

Sometimes, you can treat \( P(B) \) as a normalization constant \( \alpha \)
\[ P(A \mid B) = \alpha P(B \mid A)P(A) \]
More General Forms of Bayes Rule

If A takes 2 values:

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

If A takes \( n_A \) values:

\[ P(A = v_i | B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{k=1}^{n_A} P(B|A = v_k)P(A = v_k)} \]

Bayes Rule Example

Meningitis causes stiff necks with probability 0.5. The prior probability of having meningitis is 0.00002. The prior probability of having a stiff neck is 0.05. What is the probability of having meningitis given that you have a stiff neck?

Let \( m = \) patient has meningitis
Let \( s = \) patient has stiff neck

\[ P(s|m) = 0.5 \]
\[ P(m) = 0.00002 \]
\[ P(s) = 0.05 \]

\[ P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{(0.5)(0.00002)}{0.05} = 0.0002 \]

Bayes Rule With More Than One Piece of Evidence

Suppose you now have 2 evidence variables

\( Toothache = \text{true} \) and \( Catch = \text{catch} \) (note that Cavity is uninstantiated below)

\[ P(\text{Cavity} | Toothache = \text{true}, Catch = \text{catch}) = \frac{P(D|h)P(h)}{P(D)} \]

In order to calculate \( P(\text{Toothache} = \text{true}, Catch = \text{true} | \text{Cavity})P(\text{Cavity}) \), you need a table of 4 probability values. With \( N \) evidence variables, you need \( 2^N \) probability values.

When is Bayes Rule Useful?

Sometimes it’s easier to get \( P(X|Y) \) than \( P(Y|X) \).

Information is typically available in the form \( P(\text{effect} | \text{cause}) \) rather than \( P(\text{cause} | \text{effect}) \)

For example, \( P(\text{symptom} | \text{disease}) \) is easy to measure empirically but obtaining \( P(\text{disease} | \text{symptom}) \) is harder.
Conditional Independence

Are Toothache and Catch independent?
No – if probe catches in the tooth, it likely has a cavity which causes the toothache.

But given the presence or absence of the cavity, they are independent (since they are directly caused by the cavity but don’t have a direct effect on each other)

Conditional independence:
\[ P(\text{Toothache} = \text{true}, \text{Catch} = \text{true} \mid \text{Cavity} ) = \]
\[ P(\text{Toothache} = \text{true} \mid \text{Cavity} ) \times P(\text{Catch} = \text{true} \mid \text{Cavity} ) \]

Conditional Independence

General form:
\[ P(A, B \mid C) = P(A \mid C)P(B \mid C) \]
Or equivalently:
\[ P(A \mid B, C) = P(A \mid C) \quad \text{and} \quad P(B \mid A, C) = P(B \mid C) \]

How to think about conditional independence:
In \( P(A \mid B, C) = P(A \mid C) \): if knowing \( C \) tells me everything about \( A \), I don’t gain anything by knowing \( B \)

What You Should Know

• How to do inference in joint probability distributions
• How to use Bayes Rule
• Why independence and conditional independence is useful

Conditional independence permits probabilistic systems to scale up!