Joint Distributions

Set of random variables: \( X_1, \ldots, X_n \)

Joint Distribution:

\[
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]

or \( P(X_1, X_2, \ldots, X_n) \).

Same table:

<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>

SELECT the joint probabilities matching the evidence

\[
P(X_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1|x_2)P(x_2)}{P(x_2)}
\]

Why does this work? Sum of selection is \( P(\text{evidence})! \) (\( P(T=c) \), here)

\[
P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1|x_2)P(x_2)}{P(x_2)}
\]
**Quiz: Normalization Trick**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td>+y</td>
<td>0.2</td>
</tr>
<tr>
<td>+x</td>
<td>-y</td>
<td>0.3</td>
</tr>
<tr>
<td>-x</td>
<td>+y</td>
<td>0.4</td>
</tr>
<tr>
<td>-x</td>
<td>-y</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**SELECT the joint probabilities matching the evidence**  
**NORMALIZE the selection (make it sum to one)**

**To Normalize**

(Dictionary) To bring or restore to a normal condition (sum to one).

Procedure:
- Step 1: Compute $Z = \sum_{q} P(q, e_1, \ldots, e_k)$
- Step 2: Divide every entry by $Z$
- Step 3: Normalize

Example 1:

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Example 2:

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.2</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>

**Probabilistic Inference**

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

We generally compute conditional probabilities:
- $P(\text{on time}|\text{no reported accidents}) = 0.90$
- $P(\text{on time no accidents, 5 a.m.}) = 0.95$
- $P(\text{on time no accidents, 5 a.m. raining}) = 0.80$
- Observing new evidence causes beliefs to be updated

**Obvious problems:**
- Worst-case time complexity $O(d^n)$
- Space complexity $O(d^n)$ to store the joint distribution

**Inference by Enumeration**

**General case:**
- Evidence variables: $E_1, \ldots, E_k = e_1, \ldots, e_k$
- Query* variable: $Q$
- Hidden variables: $H_1, \ldots, H_r$

**Procedure:**
- Step 1: Compute $Z = \sum_{q} P(q, e_1, \ldots, e_k)$
- Step 2: Sum out $H$
- Step 3: Normalize

**Example 1:**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td>+y</td>
<td>0.2</td>
</tr>
<tr>
<td>+x</td>
<td>-y</td>
<td>0.3</td>
</tr>
<tr>
<td>-x</td>
<td>+y</td>
<td>0.4</td>
</tr>
<tr>
<td>-x</td>
<td>-y</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Elements consistent with evidence**

$P(Q(e_1, \ldots, e_k)) = \sum_{q} P(q, e_1, \ldots, e_k)$

**Step 3: Normalize**

$Z = \sum_{q} P(q, e_1, \ldots, e_k)$

$P(Q(e_1, \ldots, e_k)) = \frac{1}{Z} P(Q(e_1, \ldots, e_k))$

**Example:**

<table>
<thead>
<tr>
<th>S</th>
<th>T</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>summer</td>
<td>hot</td>
<td>Sun</td>
</tr>
<tr>
<td>summer</td>
<td>hot</td>
<td>rain</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>sun</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>rain</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>sun</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>rain</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>sun</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>rain</td>
</tr>
</tbody>
</table>

$P(W = \text{sun})$?
Sum out $S, T$.
Get $0.65$

$P(W = \text{sun}|\text{winter})$?
Sum out $S$ over $S=\text{winter}$.
Get $0.25$ for $W=\text{sun}$.
Get $0.25$ for $W=\text{rain}$.
Normalize: $1/2$.

$P(W = \text{sun}|\text{winter, hot})$?
Get $0.10$ for $W=\text{sun}$.
Get $0.05$ for $W=\text{rain}$.
Normalize: $2/3$. 

**Inference by Enumeration**
### The Product Rule

Sometimes have conditional distributions but want the joint

\[ P(x|y) = P(x,y) \]

\[ P(y) ↔ P(x,y) = P(x|y)P(y) \]

### The Chain Rule

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

\[ P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \]

\[ P(x_1, \ldots, x_n) = \prod_i P(x_i | x_1, \ldots, x_{i-1}) \]

Why is this always true?

Induction.

### Bayes Rule

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:

\[ P(x|y) = \frac{P(y|x)P(x)}{P(y)} \]

Why is this all helpful?

- Lets us build one conditional from its reverse
- Often one conditional is tricky but the other one is simple
- Foundation of many systems we’ll see later (e.g. ASR, MT)

In the running for most important AI equation!

### Inference with Bayes’ Rule

Example: Diagnostic probability from causal probability:

\[ P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})} \]

Example:

- M: meningitis, S: stiff neck
- Given.
- \[ P(+m) = 0.001, P(+s|+m) = 0.8, P(+s|-m) = 0.01 \]
- Bayes Rule
- \[ P(+m|+s) = \frac{P(+s|m)P(+m)}{P(+s)} = \frac{0.8(0.001)}{0.9999} \approx 0.0008 \]

Note: posterior probability of meningitis still very small

Note: you should still get stiff necks checked out! Why?
Quiz: Bayes’ Rule

Given:

<table>
<thead>
<tr>
<th>H</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.8</td>
</tr>
<tr>
<td>rain</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Pr(D—W)  

<table>
<thead>
<tr>
<th>D</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>wet</td>
<td>sun</td>
<td>0.1</td>
</tr>
<tr>
<td>dry</td>
<td>sun</td>
<td>0.9</td>
</tr>
<tr>
<td>wet</td>
<td>rain</td>
<td>0.7</td>
</tr>
<tr>
<td>dry</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

What is $P(\text{sun} | \text{dry})$?

$P(\text{dry} | \text{sun}) = \frac{P(\text{dry}, \text{sun})}{P(\text{sun})} = \frac{0.1}{0.8} = 0.125$

$P(\text{dry}) = P(\text{dry, sun}) + P(\text{dry, rain})$

$= P(\text{dry} | \text{sun}) \times P(\text{sun}) + P(\text{dry} | \text{rain}) \times P(\text{rain}) = 0.72 + 0.06 = 0.78.$

Ghostbusters, Revisited

Let’s say we have two distributions:

- Prior distribution over ghost location: $P(G)$
- Let's say this is uniform
- Sensor model: $P(R | G)$
- Given: we know what our sensors do
  - $R = \text{color measured at (1,1)}$
  - E.g. $P(R = \text{yellow} | G = (1,1)) = 0.1$

We can calculate the posterior distribution $P(G | r)$ over ghost locations given a reading using Bayes’ rule:

$P(g | r) \propto P(r | g)P(g)$

[Demo: Ghostbuster – with probability (L12D2)]

Next up: bayes nets.

CS 188: Artificial Intelligence


Probabilistic Models


Models describe how (a portion of) the world works

Models are always simplifications

- May not account for every variable
- May not account for all interactions between variables
- “All models are wrong; but some are useful.” – George E. P. Box

What do we do with probabilistic models?

- We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information
### Probability Recap


Conditional probability:
\[ P(x|y) = \frac{P(x,y)}{P(y)} \]

Product rule:
\[ P(x,y) = P(x|y)P(y) \]

Bayes Rule:
\[ P(y|x) = \frac{P(x|y)P(y)}{P(x)} \]

Chain rule:
\[ P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\ldots = \prod_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1}) \]

\( X, Y \) independent if and only if:
\[ \forall x, y : P(x,y) = P(x)P(y) \]

\( X \) and \( Y \) are conditionally independent given \( Z \) if and only if:
\[ \forall x, y, z : P(x,y|z) = P(x|z)P(y|z) \]

### Ghostbusters Chain Rule


Each sensor depends only on where the ghost is. That means, the two sensors are conditionally independent, given the ghost position.

\( T \): Top square is red.
\( B \): Bottom square is red.
\( G \): Ghost is in the top.

\[ P(T,B,G) = P(G)P(T|G)P(B|G) \]

<table>
<thead>
<tr>
<th>T</th>
<th>B</th>
<th>G</th>
<th>P(T,B,G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>0.16</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>0.24</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>0.24</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>0.06</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.06</td>
</tr>
</tbody>
</table>

**Givens:**
\[ P(+g) = 0.5 \]
\[ P(-g) = 0.5 \]
\[ P(+t|+g) = 0.8 \]
\[ P(+t|-g) = 0.4 \]
\[ P(+b|+g) = 0.4 \]
\[ P(+b|-g) = 0.8 \]

### Bayes’ Nets: Big Picture

Two problems with using full joint distribution tables as our probabilistic models:
- Unless there are only a few variables, the joint is WAY too big to represent explicitly
- Hard to learn (estimate) anything empirically about more than a few variables at a time

Bayes’ nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
- More properly called graphical models
- We describe how variables locally interact
- Local interactions chain together to give global, indirect interactions
- For now, we’ll be vague about how these interactions are specified

**Example Bayes’ Net: Insurance**

**Example Bayes’ Net: Car**
**Graphical Model Notation**

Nodes: variables (with domains)
- Can be assigned (observed) or unassigned (unobserved)

Arcs: interactions
- Similar to CSP constraints
- Indicate "direct influence" between variables
- Formally: encode conditional independence (more later)

For now: imagine that arrows mean direct causation (in general, they don't!)

**Example: Coin Flips**

N independent coin flips
No interactions between variables: absolute independence

**Example: Traffic**

Variables:
- R: It rains
- T: There is traffic

Model 1: independence

Model 2: rain causes traffic

Why is an agent using model 2 better?

**Example: Traffic II**

Let's build a causal graphical model!

- T: Traffic
- R: It rains
- L: Low pressure
- D: Roof drips
- B: Ballgame
- C: Cavity

**Example: Alarm Network**

Variables
- B: Burglary
- A: Alarm goes off
- M: Mary calls
- J: John calls
- E: Earthquake!

**Bayes’ Net Semantics**
Bayes’ Net Semantics

- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
- A collection of distributions over X, one for each combination of parents’ values
- CPT: conditional probability table
- Description of a noisy “causal” process

A Bayes net = Topology (graph) + Local Conditional Probabilities

Probabilities in Bayesnets

Bayes’ nets implicitly encode joint distributions
- As a product of local conditional distributions
  \[ P(x_1, \ldots, x_n) = \prod_i P(x_i | \text{Parents}(X_i)) \]
- To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:
  \[ P(x_1, \ldots, x_n) = \prod_i P(x_i | \text{Parents}(X_i)) \]

Example:
\[ P(G = b | T = r, B = r) = P(G = r | T = r) P(S = s | B = r) \]

Probabilities in BNs

Why are we guaranteed that setting results is joint distribution?
Chain rule (valid for all distributions):
\[ P(x_1, \ldots, x_n) = \prod_i P(x_i | x_1, \ldots, x_i) \]
Assume conditional independences:
\[ P(x_1, \ldots, x_n) = \prod_i P(x_i | \text{parents}(X_i)) \]

Consequence: Not every BN can represent every joint distribution.
- The topology enforces certain conditional independencies!

Example: Coin Flips

Only distributions whose variables are absolutely independent can be represented by a Bayes’ net with no arcs.

Example: Traffic

\[ P(s, t) = P(s) \times P(t | s) = (1/4) \times (1/4) = 1/16 \]

Example: Alarm Network
**Example: Traffic**

Causal direction

<table>
<thead>
<tr>
<th>R</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>1/4</td>
</tr>
<tr>
<td>-1</td>
<td>3/4</td>
</tr>
</tbody>
</table>

P(T|R)

| +1 | +1 | 3/4 |
| +1 | -1 | 1/4 |
| -1 | +1 | 1/2 |
| -1 | -1 | 1/2 |

P(R,T)

| +1 | +1 | 3/16 |
| +1 | -1 | 1/16 |
| -1 | +1 | 6/16 |
| -1 | -1 | 6/16 |

**Example: Reverse Traffic**

Reverse causality?

<table>
<thead>
<tr>
<th>R</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>9/16</td>
</tr>
<tr>
<td>-1</td>
<td>7/16</td>
</tr>
</tbody>
</table>

P(R|T)

| +1 | +1 | 1/3 |
| +1 | -1 | 2/3 |
| -1 | +1 | 1/7 |
| -1 | -1 | 6/7 |

P(R,T)

| +1 | +1 | 3/16 |
| +1 | -1 | 1/16 |
| -1 | +1 | 6/16 |
| -1 | -1 | 6/16 |

**Causality?**

When Bayes’ nets reflect the true causal patterns:

- Often simpler (nodes have fewer parents)
- Often easier to think about
- Often easier to elicit from experts

BNs need not actually be causal

- Sometimes no causal net exists over the domain (especially if variables are missing)
- E.g. consider the variables Traffic and Drips
- End up with arrows that reflect correlation, not causation

\[
P(X_i | X_1, \ldots, X_{i-1}) = P(X_i; \text{parents}(X_i))
\]

What do the arrows really mean?

- Topology may happen to encode causal structure
- Topology really encodes conditional independence

**Bayes’ Nets**

So far: how a Bayes’ net encodes a joint distribution

Next: how to answer queries about that distribution

- So far:
  - First assembled BNs using an intuitive notion of conditional independence as causality.
  - Then saw that key property is conditional independence
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)