# Probability

## Probability

- Motivation ................................................. 2
- Probability .............................................. 3
- Random Variables ........................................ 4
- Semantics ................................................. 5
- Dice Example ............................................. 6
- Joint Distribution Example ......................... 7
- Axioms of Probability ................................... 8
- Conditional Probabilities .............................. 9
- Using Bayes’ Theorem .................................. 10

## Conditional Independence

- Problem: Lots of Small Numbers .................. 11
- Conditional Independence .......................... 12
- Naive Bayes .............................................. 13

## Bayesian Networks

- Bayesian Networks ...................................... 14
- Example 1 ............................................... 15
- Example 2 ............................................... 16
- Joint Probability Distribution .................. 17
- Examples ................................................. 18
- Model Size .............................................. 19
- Construction ............................................ 20

## Algorithms

- Brute Force ............................................. 21
- Example Calculation, Part 1 ...................... 22
- Example Calculation, Part 2 ...................... 23
- Pruning Irrelevant Variables ..................... 24

## Markov Models

- Markov Chains .......................................... 32
- Hidden Markov Models ................................ 33
- HMM Probabilities .................................... 34

## Joint Probability Distribution

- Examples ................................................. 18

## Model Size

- Construction ............................................ 20

## Algorithms

- Brute Force ............................................. 21
- Example Calculation, Part 1 ...................... 22
- Example Calculation, Part 2 ...................... 23
- Pruning Irrelevant Variables ..................... 24

---

Example ..................................................... 25
Variable Elimination Algorithm ..................... 26
Convert to Factors ..................................... 27
Set Operation ........................................... 28
Set Operation ........................................... 29
Multiply and Sum Out ................................... 30
Multiply and Normalize .............................. 31

---

Markov Models .......................................... 32
Hidden Markov Models ................................ 33
HMM Probabilities .................................... 34
Probability

Motivation
- Agents don’t have complete knowledge about the world.
- Agents need to make decisions in an uncertain world.
- It isn’t enough to assume what the world is like. Example: wearing a seat belt.
- An agent needs to reason about its uncertainty.
- When an agent acts under uncertainty, it is gambling.
- Agents who don’t use probabilities will do worse than those who do.

Probability
- Belief in a proposition \( a \) can be measured by a number between 0 and 1 — this is the **probability of \( a \) = \( P(a) \).**
  - \( P(a) = 0 \) means that \( a \) is believed false.
  - \( P(a) = 1 \) means that \( a \) is believed true.
- \( 0 < P(a) < 1 \) means the agent is unsure.
- Probability is a measure of ignorance.
- Probability is **not** a measure of degree of truth.

Random Variables
- The variables in probability are called **random variables.**
- Each variable \( X \) has a set of possible values.
- A tuple of random variables is written as \( X_1, \ldots, X_n \).
- \( X = x \) means variable \( X \) has value \( x \).
- A **proposition** is a Boolean expression made from assignments of values to variables.
- Example: \( X_1, X_2 \) are the values of two dice. \( X_1 + X_2 = 7 \) is equivalent to \( (X_1 = 1 \land X_2 = 6) \lor (X_1 = 2 \land X_2 = 5) \lor \ldots \)

Semantics
- A **possible world** \( \omega \) is a variable assignment (all the variables).
- Let \( \Omega \) be the set of all possible worlds.
- Assuming each variable has a finite set of possible values.
  - Define \( P(\omega) \) for each world \( \omega \) so that \( 0 \leq P(\omega) \) and they sum to 1. This is the **joint probability distribution.**
  - The probability of proposition \( a \) is defined by:
    \[
    P(a) = \sum_{\omega | a} P(\omega)
    \]

Dice Example
<table>
<thead>
<tr>
<th>( X_1, X_2 )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1</td>
<td>1</td>
<td>3</td>
<td>1/36</td>
</tr>
<tr>
<td>1 2</td>
<td>3</td>
<td>1</td>
<td>1/36</td>
</tr>
<tr>
<td>1 3</td>
<td>2</td>
<td>3</td>
<td>1/36</td>
</tr>
<tr>
<td>1 4</td>
<td>5</td>
<td>3</td>
<td>1/36</td>
</tr>
<tr>
<td>1 5</td>
<td>6</td>
<td>3</td>
<td>1/36</td>
</tr>
<tr>
<td>1 6</td>
<td>4</td>
<td>3</td>
<td>1/36</td>
</tr>
<tr>
<td>2 1</td>
<td>2</td>
<td>4</td>
<td>1/36</td>
</tr>
<tr>
<td>2 2</td>
<td>4</td>
<td>4</td>
<td>1/36</td>
</tr>
<tr>
<td>2 3</td>
<td>6</td>
<td>4</td>
<td>1/36</td>
</tr>
<tr>
<td>2 4</td>
<td>5</td>
<td>4</td>
<td>1/36</td>
</tr>
<tr>
<td>2 5</td>
<td>6</td>
<td>5</td>
<td>1/36</td>
</tr>
<tr>
<td>2 6</td>
<td>6</td>
<td>6</td>
<td>1/36</td>
</tr>
</tbody>
</table>
Joint Distribution Example

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>P</th>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.04</td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.04</td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.32</td>
<td></td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.00</td>
<td></td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.00</td>
<td></td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.08</td>
<td></td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.16</td>
<td></td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.16</td>
<td></td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Axioms of Probability

Three axioms for probabilities (finite case)
1. \(0 \leq P(a)\) for any proposition \(p\).
2. \(P(a) = 1\) if \(a\) is a tautology.
3. \(P(a \lor b) = P(a) + P(b)\) if \(a\) and \(b\) contradict.

For all propositions \(a\) and \(b\), these axioms imply:

\[P(\neg a) = 1 - P(a)\]
\[P(a \land b) = P(a) + P(b) - P(a \land b)\]
\[P(a) = P(a \land b) + P(a \land \neg b)\]
\[\text{If variable } V \text{ has possible values } D, \text{ then } P(a) = \sum_{d \in D} P(a \land V = d)\]

Conditional Probabilities

- Conditional probabilities specify how to revise beliefs based on evidence, known values for one or more variables.
- If \(e\) is the evidence, the conditional probability of \(h\) given \(e\) is \(P(h \mid e) = P(h \land e) / P(e)\).
- Chain Rule: \(P(a \land b \land c) = P(a \mid b \land c)P(b \land c)P(c)\)
- Bayes' Theorem: \(P(b \mid e) = P(e \mid h)P(h) / P(e)\)
- Update with additional evidence \(e'\).
  \(P(h \mid e \land e') = P(e' \mid h \land e)P(h \mid e) / P(e' \mid e)\)

Using Bayes' Theorem

- Often you have causal knowledge:
  \(P(\text{symptom} \mid \text{disease})\)
  \(P(\text{light is off} \mid \text{status of switches})\)
  \(P(\text{alarm} \mid \text{fire})\)
  \(P(\text{looks, swims, quacks like a duck} \mid \text{a duck})\)
- and want to do evidential reasoning:
  \(P(\text{disease} \mid \text{symptom})\)
  \(P(\text{status of switches} \mid \text{light is off})\)
  \(P(\text{fire} \mid \text{alarm})\).
  \(P(\text{a duck} \mid \text{looks, swims, quacks like a duck})\)
- Bayes' theorem tells you how.
Conditional Independence

**Problem: Lots of Small Numbers**

- If there are \( n \) binary variables, there are \( 2^n \) numbers to be assigned for a joint probability distribution.
- They need to add up to one, so if \( 2^n - 1 \) numbers are assigned, the last one can be solved for. [Doesn’t help much.]
- In addition, each number is likely to be very, very small, so it is unrealistic to use frequencies (even with Big Data).
- Reduce work by using knowledge of when one variable is independent of another variable.

\[ P(X = x \wedge Y = y) = P(X = x)P(Y = y). \]

- One die is independent of the other die.
- Rain is independent of the day of the week.

\[ P(X = x | Y = y \wedge Z = z) = P(X = x | Z = z) \] for any \( x, y, z \). Knowing \( Z \), ignore \( Y \) to infer \( X \).
- A nice day (\( Y \)) makes me more likely to exercise (\( Z \)), and so more likely to be tired (\( X \)).
- Suppose \( X, Y, Z \) randomly chosen, but \( X \neq Z \) and \( Z \neq Y \).

Naive Bayes

- Suppose we want to determine the probability of a hypothesis given the evidence \( P(H | E_1, \ldots, E_n) \)
- A naive, but often effective, assumption is that the evidence is conditionally independent of the hypothesis.
  \[ P(H | E_1, \ldots, E_n) = P(H)P(E_1, \ldots, E_n | H) / P(E_1, \ldots, E_n) \approx P(H) \prod_{i=1}^{n} P(E_i | H) / P(E_1, \ldots, E_n) \]
- Different values for \( H \) have same denominator, so only need to compare numerators.

Bayesian Networks

A Bayesian network\(^*\) consists of:

- a directed acyclic graph, where nodes correspond to variables,
- a set of possible values for each variable,
- a prior probability table for each variable with no parents, and
- a conditional probability table for each variable with parents, specifying the probability of the variable given its parents.

\(^*\)I prefer “Bayesian network” over “belief network”.

---

CS 3793/5233 Artificial Intelligence

Probability – 11

Probability – 13

Probability – 14
Joint Probability Distribution

- Let $X_1, X_2, \ldots, X_n$ be the variables in the Bayesian network.
- Let $\omega$ be an assignment of values to variables.
- Let $\omega(X)$ be the value assigned to $X$.
- Let $\text{parents}(X)$ be the parents of $X$ in the Bayesian network. $\text{parents}(X) = \emptyset$ if $X$ has no parents.
- The joint probability distribution of a Bayesian network is specified by:

$$P(\omega) = \prod_{i=1}^n P(X_i = \omega(X_i) | \text{parents}(X_i) = \omega(\text{parents}(X_i)))$$

Examples

- Suppose $\neg B, E, A, J, \neg M$ (no burglary, an earthquake, an alarm, John calls, Mary doesn’t call)

$$P(\neg B, E, A, J, \neg M) = P(\neg B)P(E)P(A | \neg B, E)P(J | A)P(\neg M | A) = (0.999)(0.002)(0.29)(0.9)(0.3)$$

- Suppose $\neg C, S, \neg R, W$ (not cloudy, sprinkler was on, no rain, wet grass)

$$P(\neg C, S, \neg R, W) = P(\neg C)P(S | \neg C)P(\neg R | C)P(W | S, \neg R) = (0.5)(0.5)(0.8)(0.9)$$

Model Size

- Suppose $n$ variables, each with $k$ possible values.
- Defining a joint probability table requires $k^n$ probabilities ($k^n - 1$ to be exact).
- In a Bayesian network, a variable with $j$ parents requires a table of $k^{j+1}$ probabilities ($k^{j+1} - 1$ to be exact).
- If no variable has more than $j$ parents, then less than $n \cdot k^{j+1}$ probabilities are required.
- Number of probabilities reduced from exponential in $n$ to linear in $n$ and exponential in $j$. 

---

Example 1

| B | E | P(A|B,E) |
|---|---|---------|
| T | T | .95     |
| T | F | .94     |
| F | T | .29     |
| F | F | .01     |

Example 2

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |
Construction

To represent a problem in a Bayesian network:

□ What are the relevant variables?
  - What will you observe?
  - What would you like to infer?
  - Are there hidden variables that would make the model simpler?
□ What are the possible values of the variables?
□ What is the relationship between them? A cause should be a parent of what it directly affects.
□ How does the value of each variable depend on its parents, if any? This is expressed by prior and conditional probabilities.

Example Calculation, Part 1

Calculate \( P(W \mid C, \neg R) \) in the cloudy example.

1. Apply the conditional probability rule.
   \[
P(W \mid C, \neg R) = \frac{P(W, C, \neg R)}{P(C, \neg R)}
   \]
2. Determine which values in the joint probability table are needed.
   \[
   P(C, S, \neg R, W) \\
   P(C, \neg S, \neg R, W) \\
   P(C, S, \neg R, \neg W) \\
   P(C, \neg S, \neg R, \neg W)
   \]

Example Calculation, Part 2

3. Apply the joint probability distribution for Bayesian networks.
   \[
   P(C, S, \neg R, W) = (0.5)(0.1)(0.2)(0.9) = 0.009 \\
   P(C, \neg S, \neg R, W) = (0.5)(0.9)(0.2)(0.01) = 0.0009 \\
   P(C, S, \neg R, \neg W) = (0.5)(0.1)(0.2)(0.1) = 0.001 \\
   P(C, \neg S, \neg R, \neg W) = (0.5)(0.9)(0.2)(0.99) = 0.0891
   \]
   \[
   P(W \mid C, \neg R) = \frac{P(W, C, \neg R)}{P(C, \neg R)} \\
   = \frac{0.009 + 0.0009}{0.009 + 0.0009 + 0.001 + 0.0891} = 0.099
   \]
Pruning Irrelevant Variables

Some variables might not be relevant to $P(h \mid e)$.

- Prune any variables that have no observed or queried descendents, that is, not part of $e$ or $h$.
- Connect the parents of any observed variable.
- Remove arc directions and observed variables.
- Prune any variables not connected to $h$ in the (undirected) graph.
- Calculate $P(h \mid e)$ in original network minus pruned variables.

In example on the next slide, compare $P(H \mid Q, P)$ and $P(H \mid K, Q, P)$.

Example

Variable Elimination Algorithm

Variable elimination is an exact algorithm for computing $P(h \mid e)$. Assume $h$ is one variable.

- Convert each probability table into a factor. Let $F$ be all the factors.
- Eliminate each non-query variable $X$ in turn:
  - Identify the factors $F'$ in which $X$ appears.
  - If $X$ is observed, set $X$ to the observed value in each factor in $F'$.
  - Otherwise, multiply the factors in $F'$ together, sum out $X$, add the result to $F$, and remove $F'$ from $F$.
- Multiply the remaining factors and normalize.

Convert to Factors

These are the factors of the cloudy network.

\[
\begin{array}{|c|c|c|c|}
\hline
C & S & R & W \\
\hline
T & 0.5 & T & T & 0.99 \\
F & 0.5 & T & F & 0.01 \\
F & 0.5 & F & T & 0.90 \\
F & 0.5 & F & F & 0.10 \\
\hline
\end{array}
\]

Want $P(W \mid C, \neg R)$.

\[
\begin{array}{|c|c|c|c|c|}
\hline
C & S & R & W & \text{val} \\
\hline
T & T & T & T & 0.8 \\
T & F & F & F & 0.90 \\
F & T & F & F & 0.10 \\
F & F & T & F & 0.01 \\
F & F & F & F & 0.99 \\
\hline
\end{array}
\]
Set Operation

Setting $C$ to $T$ and $R$ to $F$.

<table>
<thead>
<tr>
<th>C</th>
<th>S val</th>
<th>R</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$0.5$</td>
<td>$T$</td>
<td>$0.8$</td>
</tr>
<tr>
<td>$E$</td>
<td>$0.5$</td>
<td>$T$</td>
<td>$4.2$</td>
</tr>
<tr>
<td>$F$</td>
<td>$0.5$</td>
<td>$T$</td>
<td>$4.2$</td>
</tr>
<tr>
<td>$F$</td>
<td>$0.5$</td>
<td>$F$</td>
<td>$0.8$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$0.1$</td>
</tr>
<tr>
<td>$F$</td>
<td>$0.9$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$0.5$</td>
</tr>
<tr>
<td>$F$</td>
<td>$0.9$</td>
</tr>
</tbody>
</table>

Multiply and Normalize

Multiply remaining tables (only $h$ is left).

$$\begin{align*}
\text{val} & \times \text{val} \times W \frac{\text{val}}{T} = W \frac{\text{val}}{T} \\
0.5 & \times 0.2 \times W \frac{0.099}{T} = W \frac{0.099}{T} \\
\text{Normalize} = W \frac{0.099}{T} \\
\end{align*}$$

Markov Models

Markov Chains

A Markov chain is a Bayesian network for representing a sequence of values.

- This represents the Markov assumption:
  $$P(S_{t+1} \mid S_0, \ldots, S_t) = P(S_{t+1} \mid S_t)$$
- A stationary Markov chain has
  $$P(S_{t+1} \mid S_t) = P(S_1 \mid S_0)$$
- Sequence of states over time, e.g., queueing theory.
- Probability of next item (word, letter) given previous item.
Hidden Markov Models

A hidden Markov model adds observations to a Markov chain.

- States are not directly observable.
- Observations provide evidence for states.
- $P(S_0)$ specifies initial conditions.
- $P(S_{t+1} \mid S_t)$ specifies the dynamics.
- $P(O_k \mid S_i)$ specifies the sensor model.

HMM Probabilities

- To compute $P(S_i \mid O_0, \ldots, O_i, \ldots, O_k)$
  - Compute variable elimination forward. This computes $P(S_i \mid O_0, \ldots, O_{i-1})$.
  - Compute variable elimination backward. This computes $P(O_i, \ldots, O_k \mid S_i)$.
  - $P(S_i \mid O_0, \ldots, O_k) \propto P(S_i \mid O_0, \ldots, O_{i-1}) \cdot P(O_i, \ldots, O_k \mid S_i)$

- To compute most probable sequence of states:
  - Viterbi algorithm.
  - Idea: Find most probable sequence to $S_i$.
  - Use that information to extend sequence by one state, and repeat.