Probability
Uncertainty

• Let action $A_t = \text{leave for airport } t \text{ minutes before flight}$
  – Will $A_t$ get me there on time?

• Problems:
  • Partial observability (road state, other drivers' plans, etc.)
  • Noisy sensors (traffic reports)
  • Uncertainty in action outcomes (flat tire, etc.)
  • Complexity of modeling and predicting traffic

• Hence a purely logical approach either
  • Risks falsehood: “$A_{25}$ will get me there on time,” or
  • Leads to conclusions that are too weak for decision making:
    • $A_{25}$ will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact, etc., etc.
    • $A_{1440}$ might reasonably be said to get me there on time but I'd have to stay overnight in the airport
Probability

Probabilistic assertions summarize effects of

- **Laziness**: failure to enumerate exceptions, qualifications, etc.
- **Ignorance**: lack of explicit theories, relevant facts, initial conditions, etc.
- Intrinsically random behavior
Making decisions under uncertainty

- Suppose the agent believes the following:
  
  \[ P(A_{25} \text{ gets me there on time}) = 0.04 \]
  
  \[ P(A_{90} \text{ gets me there on time}) = 0.70 \]
  
  \[ P(A_{120} \text{ gets me there on time}) = 0.95 \]
  
  \[ P(A_{1440} \text{ gets me there on time}) = 0.9999 \]

- Which action should the agent choose?
  - Depends on preferences for missing flight vs. time spent waiting
  - Encapsulated by a utility function

- The agent should choose the action that maximizes the expected utility:
  
  \[ P(A_t \text{ succeeds}) \times U(A_t \text{ succeeds}) + P(A_t \text{ fails}) \times U(A_t \text{ fails}) \]

- Utility theory is used to represent and infer preferences
- Decision theory = probability theory + utility theory
Where do probabilities come from?

• **Frequentism**
  – Probabilities are relative frequencies
  – For example, if we toss a coin many times, $P(\text{heads})$ is the proportion of the time the coin will come up heads
  – But what if we’re dealing with events that only happen once?
    • E.g., what is the probability that Republicans will take over Congress in 2010?
  – “Reference class” problem

• **Subjectivism**
  – Probabilities are degrees of belief
  – But then, how do we assign belief values to statements?
  – What would constrain agents to hold consistent beliefs?
Random variables

- We describe the (uncertain) state of the world using \textit{random variables}
  - Denoted by capital letters
    - $\mathbf{R}$: Is it raining?
    - $\mathbf{W}$: What’s the weather?
    - $\mathbf{D}$: What is the outcome of rolling two dice?
    - $\mathbf{S}$: What is the speed of my car (in MPH)?

- Just like variables in CSP’s, random variables take on values in a \textit{domain}
  - Domain values must be mutually exclusive and exhaustive
    - $\mathbf{R}$ in \{True, False\}
    - $\mathbf{W}$ in \{Sunny, Cloudy, Rainy, Snow\}
    - $\mathbf{D}$ in \{(1,1), (1,2), … (6,6)\}
    - $\mathbf{S}$ in [0, 200]
Events

- Probabilistic statements are defined over events, or sets of world states
  - “It is raining”
  - “The weather is either cloudy or snowy”
  - “The sum of the two dice rolls is 11”
  - “My car is going between 30 and 50 miles per hour”

- Events are described using propositions:
  - \( R = \text{True} \)
  - \( W = \text{“Cloudy”} \lor W = \text{“Snowy”} \)
  - \( D \in \{(5,6), (6,5)\} \)
  - \( 30 \leq S \leq 50 \)

- Notation: \( P(A) \) is the probability of the set of world states in which proposition A holds
  - \( P(X = x) \), or \( P(x) \) for short, is the probability that random variable \( X \) has taken on the value \( x \)
Kolmogorov’s axioms of probability

- For any propositions (events) $A, B$
  - $0 \leq P(A) \leq 1$
  - $P(\text{True}) = 1$ and $P(\text{False}) = 0$
  - $P(A \lor B) = P(A) + P(B) - P(A \land B)$
    - Subtraction accounts for double-counting

- Based on these axioms, what is $P(\neg A)$?

- These axioms are sufficient to completely specify probability theory for discrete random variables
  - For continuous variables, need *density functions*
Probabilities and rationality

- Why should a rational agent hold beliefs that are consistent with axioms of probability?

- De Finetti (1931): If an agent has some degree of belief in proposition A, he/she should be able to decide whether or not to accept a bet for/against A
  - E.g., if the agent believes that $P(A) = 0.4$, should he/she agree to bet $6 that A will occur against $4 that A will not occur?

- **Theorem**: An agent who holds beliefs inconsistent with axioms of probability can be tricked into accepting a combination of bets that are guaranteed to lose them money.
Atomic events

- **Atomic event**: a complete specification of the state of the world, or a complete assignment of domain values to all random variables
  - Atomic events are mutually exclusive and exhaustive

- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:
  
  \[
  \begin{align*}
  &Cavity = false \land Toothache = false \\
  &Cavity = false \land Toothache = true \\
  &Cavity = true \land Toothache = false \\
  &Cavity = true \land Toothache = true
  \end{align*}
  \]
Joint probability distributions

- A **joint distribution** is an assignment of probabilities to every possible atomic event.

<table>
<thead>
<tr>
<th>Atomic event</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Cavity = false \land Toothache = false$</td>
<td>0.8</td>
</tr>
<tr>
<td>$Cavity = false \land Toothache = true$</td>
<td>0.1</td>
</tr>
<tr>
<td>$Cavity = true \land Toothache = false$</td>
<td>0.05</td>
</tr>
<tr>
<td>$Cavity = true \land Toothache = true$</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- Why does it follow from the axioms of probability that the probabilities of all possible atomic events must sum to 1?
Joint probability distributions

• Suppose we have a joint distribution $P(X_1, X_2, \ldots, X_n)$ of $n$ random variables with domain sizes $d$
  – What is the size of the probability table?
  – Impossible to write out completely for all but the smallest distributions

• Notation:
  – $P(X = x)$ is the probability that random variable $X$ takes on value $x$
  – $P(X)$ is the distribution of probabilities for all possible values of $X$
Marginal probability distributions

- Suppose we have the joint distribution \( P(X,Y) \) and we want to find the marginal distribution \( P(Y) \).

\[
\begin{array}{|c|c|}
\hline
P(\text{Cavity, Toothache}) & \text{Probability} \\
\hline
\text{Cavity} = \text{false} \land \text{Toothache} = \text{false} & 0.8 \\
\text{Cavity} = \text{false} \land \text{Toothache} = \text{true} & 0.1 \\
\text{Cavity} = \text{true} \land \text{Toothache} = \text{false} & 0.05 \\
\text{Cavity} = \text{true} \land \text{Toothache} = \text{true} & 0.05 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
P(\text{Cavity}) & \text{Probability} \\
\hline
\text{Cavity} = \text{false} & ? \\
\text{Cavity} = \text{true} & ? \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
P(\text{Toothache}) & \text{Probability} \\
\hline
\text{Toothache} = \text{false} & ? \\
\text{Toothache} = \text{true} & ? \\
\hline
\end{array}
\]
Marginal probability distributions

• Suppose we have the joint distribution $P(X,Y)$ and we want to find the marginal distribution $P(Y)$

\[ P(X = x) = P((X = x \land Y = y_1) \lor \ldots \lor (X = x \land Y = y_n)) \]
\[ = P((x, y_1) \lor \ldots \lor (x, y_n)) = \sum_{i=1}^{n} P(x, y_i) \]

• General rule: to find $P(X = x)$, sum the probabilities of all atomic events where $X = x$. 
Conditional probability

- Probability of cavity given toothache:
  \[ P(\text{Cavity} = \text{true} \mid \text{Toothache} = \text{true}) \]

- For any two events A and B, \[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)} \]
Conditional probability

<table>
<thead>
<tr>
<th>P(Cavity, Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false \∧ Toothache = false</td>
<td>0.8</td>
</tr>
<tr>
<td>Cavity = false \∧ Toothache = true</td>
<td>0.1</td>
</tr>
<tr>
<td>Cavity = true \∧ Toothache = false</td>
<td>0.05</td>
</tr>
<tr>
<td>Cavity = true \∧ Toothache = true</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(Cavity)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false</td>
<td>0.9</td>
</tr>
<tr>
<td>Cavity = true</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Toothache = false</td>
<td>0.85</td>
</tr>
<tr>
<td>Toothache = true</td>
<td>0.15</td>
</tr>
</tbody>
</table>

• What is $P(Cavity = true \mid Toothache = false)$?
  $0.05 / 0.85 = 0.059$

• What is $P(Cavity = false \mid Toothache = true)$?
  $0.1 / 0.15 = 0.667$
Conditional distributions

- A conditional distribution is a distribution over the values of one variable given fixed values of other variables.

<table>
<thead>
<tr>
<th>P(Cavity, Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false ∧ Toothache = false</td>
<td>0.8</td>
</tr>
<tr>
<td>Cavity = false ∧ Toothache = true</td>
<td>0.1</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = false</td>
<td>0.05</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = true</td>
<td>0.05</td>
</tr>
</tbody>
</table>

| P(Cavity | Toothache = true) |        |
|-----------|-------------------|
| Cavity = false | 0.667           |
| Cavity = true   | 0.333           |

| P(Cavity | Toothache = false) | P(Toothache | Cavity = true) | P(Toothache | Cavity = false) |
|-----------|---------------------|----------------|----------------|
| Cavity = false | 0.5               | 0.5            | 0.889          |
| Cavity = true   | 0.5               | 0.5            | 0.111          |
Normalization trick

- To get the whole conditional distribution $P(X \mid y)$ at once, select all entries in the joint distribution matching $Y = y$ and renormalize them to sum to one.

<table>
<thead>
<tr>
<th>$P(Cavity, Toothache)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Cavity = false \land Toothache = false$</td>
</tr>
<tr>
<td>$Cavity = false \land Toothache = true$</td>
</tr>
<tr>
<td>$Cavity = true \land Toothache = false$</td>
</tr>
<tr>
<td>$Cavity = true \land Toothache = true$</td>
</tr>
</tbody>
</table>

Select

<table>
<thead>
<tr>
<th>Toothache, Cavity = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Toothache = false$</td>
</tr>
<tr>
<td>$Toothache = true$</td>
</tr>
</tbody>
</table>

Renormalize

<table>
<thead>
<tr>
<th>$P(Toothache \mid Cavity = false)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Toothache = false$</td>
</tr>
<tr>
<td>$Toothache = true$</td>
</tr>
</tbody>
</table>
Normalization trick

• To get the whole conditional distribution \( P(X \mid y) \) at once, select all entries in the joint distribution matching \( Y = y \) and renormalize them to sum to one.

• Why does it work?

\[
\frac{P(x, y)}{\sum_{a'} P(x', y)} = \frac{P(x, y)}{P(y)} \quad \text{by marginalization}
\]
Product rule

- Definition of conditional probability: \( P(A \mid B) = \frac{P(A, B)}{P(B)} \)

- Sometimes we have the conditional probability and want to obtain the joint:

\[ P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]
Product rule

• Definition of conditional probability: 
\[ P(A | B) = \frac{P(A, B)}{P(B)} \]

• Sometimes we have the conditional probability and want to obtain the joint:
\[ P(A, B) = P(A | B)P(B) = P(B | A)P(A) \]

• The chain rule:
\[
P(A_1, \ldots, A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2) \ldots P(A_n | A_1, \ldots, A_{n-1}) \\
= \prod_{i=1}^{n} P(A_i | A_1, \ldots, A_{i-1})
\]
Bayes Rule

• The product rule gives us two ways to factor a joint distribution:

\[ P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]

• Therefore,

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]

• Why is this useful?
  – Can get diagnostic probability \( P(\text{cavity} \mid \text{toothache}) \) from causal probability \( P(\text{toothache} \mid \text{cavity}) \)
  – Can update our beliefs based on evidence
  – Important tool for probabilistic inference
Independence

• Two events A and B are independent if and only if
  \[ P(A \land B) = P(A) \cdot P(B) \]
  – In other words, \[ P(A | B) = P(A) \] and \[ P(B | A) = P(B) \]
  – This is an important simplifying assumption for modeling, e.g., Toothache and Weather can be assumed to be independent

• Are two mutually exclusive events independent?
  – No, but for mutually exclusive events we have
    \[ P(A \lor B) = P(A) + P(B) \]

• Conditional independence: A and B are conditionally independent given C iff
  \[ P(A \land B | C) = P(A | C) \cdot P(B | C) \]
Conditional independence: Example

- **Toothache**: boolean variable indicating whether the patient has a toothache
- **Cavity**: boolean variable indicating whether the patient has a cavity
- **Catch**: whether the dentist’s probe catches in the cavity

If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether he/she has a toothache

\[ P(Catch \mid Toothache, Cavity) = P(Catch \mid Cavity) \]

Therefore, **Catch** is conditionally independent of **Toothache** given **Cavity**

Likewise, **Toothache** is conditionally independent of **Catch** given **Cavity**

\[ P(Toothache \mid Catch, Cavity) = P(Toothache \mid Cavity) \]

Equivalent statement:

\[ P(Toothache, Catch \mid Cavity) = P(Toothache \mid Cavity) P(Catch \mid Cavity) \]
Conditional independence: Example

• How many numbers do we need to represent the joint probability table $P(\text{Toothache, Cavity, Catch})$?
  \[2^3 - 1 = 7 \text{ independent entries}\]

• Write out the joint distribution using chain rule:
  \[
P(\text{Toothache, Catch, Cavity})
  = P(\text{Cavity}) P(\text{Catch | Cavity}) P(\text{Toothache | Catch, Cavity})
  = P(\text{Cavity}) P(\text{Catch | Cavity}) P(\text{Toothache | Cavity})
  \]

• How many numbers do we need to represent these distributions?
  \[1 + 2 + 2 = 5 \text{ independent numbers}\]

• In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in $n$ to linear in $n$
Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) that we want to use to diagnose the underlying cause
- It is usually impractical to directly estimate or store the joint distribution $P(Cause, Effect_1, \ldots, Effect_n)$. 
- To simplify things, we can assume that the different effects are conditionally independent given the underlying cause
Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) that we want to use to diagnose the underlying cause.
- It is usually impractical to directly estimate or store the joint distribution \( P(Cause, Effect_1, \ldots, Effect_n) \).
- To simplify things, we can assume that the different effects are conditionally independent *given the underlying cause*.
- Then we can estimate the joint distribution as

\[
P(Cause, Effect_1, \ldots, Effect_n) = P(Cause) \prod_i P(Effect_i \mid Cause)
\]

- This is usually not accurate, but very useful in practice.
Example: Naïve Bayes Spam Filter

- **Bayesian decision theory**: to minimize the probability of error, we should classify a message as spam if
  \[ P(\text{spam} | \text{message}) > P(\neg\text{spam} | \text{message}) \]
  - *Maximum a posteriori (MAP)* decision

- We have
  \[
P(\text{spam} | \text{message}) = \frac{P(\text{message} | \text{spam}) P(\text{spam})}{P(\text{message})}
  \]
  and
  \[
P(\neg\text{spam} | \text{message}) = \frac{P(\text{message} | \neg\text{spam}) P(\neg\text{spam})}{P(\text{message})}
  \]

- Notice that \( P(\text{message}) \) is just a constant normalizing factor and doesn’t affect the decision

- Therefore, all we need is to find \( P(\text{message} | \text{spam}) P(\text{spam}) \) and \( P(\text{message} | \neg\text{spam}) P(\neg\text{spam}) \)
Example: Naïve Bayes Spam Filter

- We need to find $P(\text{message} \mid \text{spam}) P(\text{spam})$ and $P(\text{message} \mid \neg\text{spam}) P(\neg\text{spam})$
- The message is a sequence of words $(w_1, \ldots, w_n)$
- **Bag of words** representation
  - The order of the words in the message is not important
  - Each word is conditionally independent of the others given message class (spam or not spam)

\[
P(\text{message} \mid \text{spam}) = P(w_1, \ldots, w_n \mid \text{spam}) = \prod_{i=1}^{n} P(w_i \mid \text{spam})
\]

- Our filter will classify the message as spam if

\[
P(\text{spam}) \prod_{i=1}^{n} P(w_i \mid \text{spam}) > P(\neg\text{spam}) \prod_{i=1}^{n} P(w_i \mid \neg\text{spam})
\]
Example: Naïve Bayes Spam Filter

\[
P(\text{spam} \mid w_1, \ldots, w_n) = P(\text{spam}) \prod_{i=1}^{n} P(w_i \mid \text{spam})
\]

- posterior
- prior
- likelihood
Probabilistic inference

• In general, the agent observes the values of some random variables $X_1, X_2, \ldots, X_n$ and needs to reason about the values of some other unobserved random variables $Y_1, Y_2, \ldots, Y_m$
  – Figuring out a diagnosis based on symptoms and test results
  – Classifying the content type of an image or a document based on some features
• This will be the subject of the next few lectures