Probability and Uncertainty
Warm-up and Review for Bayesian Networks and Machine Learning

This lecture: Read Chapter 13
Next Lecture: Read Chapter 14.1-14.2

Please do all readings both before and again after lecture.
Outline

• Representing uncertainty is useful in knowledge bases.
  – Probability provides a framework for managing uncertainty

• Review of basic concepts in probability.
  – Emphasis on conditional probability and conditional independence

• Using a full joint distribution and probability rules, we can derive any probability relationship in a probability space.
  – Number of required probabilities can be reduced through independence and conditional independence relationships

• Probabilities allow us to make better decisions.
  – Decision theory and expected utility.

• **Rational agents cannot violate probability theory.**
You will be expected to know

• Basic probability notation/definitions:
  – Probability model, unconditional/prior and conditional/posterior probabilities, factored representation (= variable/value pairs), random variable, (joint) probability distribution, probability density function (pdf), marginal probability, (conditional) independence, normalization, etc.

• Basic probability formulae:
  – Probability axioms, product rule, Bayes’ rule.

• How to use Bayes’ rule:
  – Naïve Bayes model (naïve Bayes classifier)
The Problem: Uncertainty

• We cannot always know everything relevant to the problem before we select an action:
  – Environments that are non-deterministic, partially observable
  – Noisy sensors
  – Some features may be too complex model

• For Example: Trying to decide when to leave for the airport to make a flight
  – Will I get me there on time?
  – Uncertainties:
    ▪ Car failures (flat tire, engine failure) (non-deterministic)
    ▪ Road state, accidents, natural disasters (partially observable)
    ▪ Unreliable weather reports, traffic updates (noisy sensors)
    ▪ Predicting traffic along route (complex modeling)

• A purely logical agent does not allow for strong decision making in the face of such uncertainty.
  – Purely logical agents are based on binary True/False statements, no maybe
  – Forces us to make assumptions to find a solution --> weak solutions
Handling Uncertainty

• **Default** or **non-monotonic** logic:
  – Based on assuming things are a certain way, unless evidence to the contrary.
    ▪ Assume my car does not have a flat tire
    ▪ Assume road ahead is clear, no accidents
  – **Issues:** What assumptions are reasonable?
   How to retract inferences when assumptions found false?

• Rules with **fudge factors**:
  – Based on guesses or rules of thumb for relationships between events.
    ▪ A25 => 0.3 get there on time
    ▪ Rain => 0.99 grass wet
  – **Issues:** No theoretical framework for combination

• **Probability**:
  – Based on degrees of belief, given the available evidence
  – Solidly rooted in statistics
Propositional Logic and Probability

• Their ontological commitments are the same
  – The world is a set of facts that do or do not hold

• Their epistemological commitments differ
  – Logic agent believes true, false, or no opinion
  – Probabilistic agent has a numerical degree of belief between 0 (false) and 1 (true)
Probabilistic Agents
May Make Better Decisions

The Logic Agent has no basis for choosing among [1,3], [2,2], & [3,1].
The Probabilistic Agent can calculate \( P(P_{1,3}) = P(P_{3,1}) \approx 0.31 \),
while \( P(P_{2,2}) \approx 0.86 \) — so it will avoid [2,2] and prefer [1,3] or [3,1].
(R&N, section 13.6)
Probability

• P(a) is the probability of proposition “a”
  – e.g., P(it will rain in London tomorrow)
  – The proposition a is actually true or false in the real-world

• **Probability Axioms:**
  – 0 ≤ P(a) ≤ 1
  – P(NOT(a)) = 1 − P(a) => Σₐ P(A) = 1
  – P(true) = 1
  – P(false) = 0
  – P(A OR B) = P(A) + P(B) − P(A AND B)

• Any agent that holds degrees of beliefs that contradict these axioms will act irrationally in some cases

• **Rational agents cannot violate probability theory.**
  – Acting otherwise results in irrational behavior.
Probability

• Probabilities can be subjective:
  – Agents develop probabilities based on their experiences:
    ▪ Two agents may have different internal probabilities of the same event occurring.

• Probabilities of propositions change with new evidence:
  – $P(\text{party tonight}) = 0.15$
  – $P(\text{party tonight} \mid \text{Friday}) = 0.60$
Interpretations of Probability

• **Relative Frequency:** *What we were taught in school*
  – $P(a)$ represents the frequency that event $a$ will happen in repeated trials.
  – Requires event $a$ to have happened enough times for data to be collected.

• **Degree of Belief:** *A more general view of probability*
  – $P(a)$ represents an agent’s degree of belief that event $a$ is true.
  – Can predict probabilities of events that occur rarely or have not yet occurred.
  – Does not require new or different rules, just a different interpretation.

• **Examples:**
  – $a = “life exists on another planet”$
    • What is $P(a)$? We will all assign different probabilities
  – $a = “Hilary Clinton will be the next US president”$
    • What is $P(a)$?
  – $a = “over 50% of the students in this class will get A’s”$
    • What is $P(a)$?
Concepts of Probability

• **Unconditional Probability** (AKA marginal or prior probability):
  – $P(a)$, the probability of “a” being true
  – Does not depend on anything else to be true (**unconditional**)
  – Represents the probability prior to further information that may adjust it (**prior**)

• **Conditional Probability** (AKA posterior probability):
  – $P(a|b)$, the probability of “a” being true, given that “b” is true
  – Relies on “b” = true (**conditional**)
  – Represents the prior probability adjusted based upon new information “b” (**posterior**)
  – Can be generalized to more than 2 random variables:
    ▪ e.g. $P(a|b, c, d)$

• **Joint Probability** :
  – $P(a, b) = P(a \land b)$, the probability of “a” and “b” both being true
  – Can be generalized to more than 2 random variables:
    ▪ e.g. $P(a, b, c, d)$
Random Variables

• **Random Variable**: 
  – Basic element of probability assertions
  – Similar to CSP variable, but values reflect probabilities not constraints.
    ▪ Variable: A
    ▪ Domain: \{a_1, a_2, a_3\} <-- events / outcomes

• **Types of Random Variables**: 
  – **Boolean** random variables = \{true, false\}
    ▪ e.g., Cavity (= do I have a cavity?)
  – **Discrete** random variables = One value from a set of values
    ▪ e.g., Weather is one of <sunny, rainy, cloudy, snow>
  – **Continuous** random variables = A value from within constraints
    ▪ e.g., Current temperature is bounded by (10°, 200°)

• Domain values must be **exhaustive and mutually exclusive**: 
  – One of the values must always be the case **(Exhaustive)**
  – Two of the values cannot both be the case **(Mutually Exclusive)**
Random Variables

• **For Example:** Flipping a coin
  – Variable = R, the result of the coin flip
  – Domain = \{heads, tails, edge\} \text{ <-- must be exhaustive}
  – \( P(R = \text{heads}) = 0.4999 \)
  – \( P(R = \text{tails}) = 0.4999 \) \text{ -- must be exclusive}
  – \( P(R = \text{edge}) = 0.0002 \)

• **Shorthand is often used for simplicity:**
  – Upper-case letters for variables, lower-case letters for values.
  – e.g.
    \[
    P(a) \equiv P(A = a) \\
    P(a \mid b) \equiv P(A = a \mid B = b) \\
    P(a, b) \equiv P(A = a, B = b)
    \]

• **Two kinds of probability propositions:**
  – **Elementary propositions** are an assignment of a value to a random variable:
    – e.g., \( \text{Weather} = \text{sunny}; \text{Cavity} = \text{false} \) (abbreviated as \( \neg \text{cavity} \))
  – **Complex propositions** are formed from elementary propositions and standard logical connectives:
    – e.g., \( \text{Cavity} = \text{false} \lor \text{Weather} = \text{sunny} \)
Probability Space

\[ P(A) + P(\neg A) = 1 \]

Entire Sample Space: \( P(S) = 1 \)

Event A: \( \text{Prob} = P(A) \)

Area = Probability of Event
AND Probability

\[ P(A, B) = P(A \land B) = P(A) + P(B) - P(A \lor B) \]

Entire Sample Space: \( P(S) = 1 \)

Area = Probability of Event
OR Probability

\[ P(A \lor B) = P(A) + P(B) - P(A, B) \]

Entire Sample Space: \( P(S) = 1 \)

Area = Probability of Event
Conditional Probability

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

Entire Sample Space: \( P(S) = 1 \)

Area = Probability of Event
Product Rule

\[ P(A,B) = P(A|B) \cdot P(B) \]

Entire Sample Space: \( P(S) = 1 \)

Area = Probability of Event

\[ P(A \land B) = P(A) + P(B) - P(A \lor B) \]
Using the Product Rule

- **Applies to any number of variables:**
  - \( P(a, b, c) = P(a, b|c) P(c) = P(a|b, c) P(b, c) \)
  - \( P(a, b, c|d, e) = P(a|b, c, d, e) P(b, c) \)

- **Factoring:** (AKA Chain Rule for probabilities)
  - By the product rule, we can always write:
    \( P(a, b, c, ... z) = P(a \mid b, c, ..., z) P(b, c, ... z) \)
  - Repeatedly applying this idea, we can write:
    \( P(a, b, c, ... z) = P(a \mid b, c, ..., z) P(b \mid c, ... z) P(c \mid .. z) P(z) \)
  - This holds for any ordering of the variables
Sum Rule

\[ P(A) = \sum_{B,C} P(A, B, C) \]

Entire Sample Space: \( P(S) = 1 \)

Area = Probability of Event
Using the Sum Rule

• We can marginalize variables out of any joint distribution by simply summing over that variable:
  – \( P(b) = \sum_a \sum_c \sum_d P(a, b, c, d) \)
  – \( P(a, d) = \sum_b \sum_c P(a, b, c, d) \)

• For Example: Determine probability of catching a fish
  – Given a set of probabilities \( P(CatchFish, Day, Lake) \)
  – Where:
    ▪ \( \text{CatchFish} = \{\text{true}, \text{false}\} \)
    ▪ \( Day = \{\text{mon}, \text{tues}, \text{wed}, \text{thurs}, \text{fri}, \text{sat}, \text{sun}\} \)
    ▪ \( Lake = \{\text{buel lake}, \text{ralph lake}, \text{crystal lake}\} \)

  – Need to find \( P(CatchFish = \text{true}) \):
    ▪ \( P(CatchFish = \text{true}) = \sum_{\text{day}} \sum_{\text{lake}} P(CatchFish = \text{true}, \text{day}, \text{lake}) \)
Bayes’ Rule

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

Entire Sample Space: \( P(S) = 1 \)

\[ P(A \land B) = P(A) + P(B) - P(A \lor B) \]

Area = Probability of Event
Derivation of Bayes’ Rule

• **Start from Product Rule:**
  
  \[ P(a, b) = P(a|b) P(b) = P(b|a) P(a) \]

• **Isolate Equality on Right Side:**
  
  \[ P(a|b) P(b) = P(b|a) P(a) \]

• **Divide through by P(b):**
  
  \[ P(a|b) = P(b|a) P(a) / P(b) \]  
  \[ \text{ <-- Bayes’ Rule} \]
Using Bayes’ Rule

• **For Example**: Determine probability of meningitis given a stiff neck
  
  – **Given**:
    - \( P(\text{stiff neck} \mid \text{meningitis}) = 0.5 \) 
    - \( P(\text{meningitis}) = \frac{1}{50,000} \) -- From medical databases
    - \( P(\text{stiff neck}) = \frac{1}{20} \)
  
  – **Need to find** \( P(\text{meningitis} \mid \text{stiff neck}) \):
    - \[ P(m \mid s) = \frac{P(s \mid m) \ P(m)}{P(s)} \] \hspace{3.0cm} \text{[Bayes’ Rule]}
    - \[ = \frac{0.5 \times 1/50,000}{1/20} = 1/5,000 \]
  
  – 10 times more likely to have meningitis given a stiff neck

• **Applies to any number of variables**:
  – Any probability \( P(X \mid Y) \) can be rewritten as \( P(Y \mid X) \ P(X) / P(Y) \), even if \( X \) and \( Y \) are lists of variables.
  – \( P(a \mid b, c) = P(b, c \mid a) \ P(a) / P(b, c) \)
  – \( P(a, b \mid c, d) = P(c, d \mid a, b) \ P(a, b) / P(c, d) \)
Summary of Probability Rules

• **Product Rule**:  
  - \( P(a, b) = P(a|b) \ P(b) = P(b|a) \ P(a) \)  
  - Probability of “a” and “b” occurring is the same as probability of “a” occurring given “b” is true, times the probability of “b” occurring.  
    ▪ e.g., \( P(\text{rain, cloudy}) = P(\text{rain} | \text{cloudy}) \times P(\text{cloudy}) \)  

• **Sum Rule**: (AKA Law of Total Probability)  
  - \( P(a) = \sum_b P(a, b) = \sum_b P(a|b) \ P(b), \) where \( B \) is any random variable  
  - Probability of “a” occurring is the same as the sum of all joint probabilities including the event, provided the joint probabilities represent all possible events.  
  - Can be used to “marginalize” out other variables from probabilities, resulting in prior probabilities also being called marginal probabilities.  
    ▪ e.g., \( P(\text{rain}) = \sum_{\text{Windspeed}} P(\text{rain, Windspeed}) \)  
    where Windspeed = \{0-10mph, 10-20mph, 20-30mph, etc.\}  

• **Bayes’ Rule**:  
  - \( P(b|a) = P(a|b) \ P(b) / P(a) \)  
  - Acquired from rearranging the product rule.  
  - Allows conversion between conditionals, from \( P(a|b) \) to \( P(b|a) \).  
    ▪ e.g., \( b = \text{disease}, a = \text{symptoms} \)  
      More natural to encode knowledge as \( P(a|b) \) than as \( P(b|a) \).
Full Joint Distribution

• We can fully specify a probability space by constructing a full joint distribution:
  – A full joint distribution contains a probability for every possible combination of variable values. This requires:
    \[ \prod_{\text{vars}} (n_{\text{var}}) \text{ probabilities} \]
    where \( n_{\text{var}} \) is the number of values in the domain of variable \( \text{var} \)
  – e.g. \( P(A, B, C) \), where \( A, B, C \) have 4 values each
    Full joint distribution specified by \( 4^3 \) values = 64 values

• Using a full joint distribution, we can use the product rule, sum rule, and Bayes’ rule to create any combination of joint and conditional probabilities.
Decision Theory: Why Probabilities are Useful

• We can use probabilities to make better decisions!

• For Example: Deciding whether to operate on a patient
  – Given:
    ▪ $\textit{Operate} = \{\text{true, false}\}$
    ▪ $\textit{Cancer} = \{\text{true, false}\}$
    ▪ A set of evidence $e$

  – So far, agent’s degree of belief is $p(\textit{Cancer} = \text{true} \mid e)$.

  – Which action to choose?
    ▪ Depends on the agent’s preferences:
      o How willing is the agent to operate if there is no cancer?
      o How willing is the agent to not operate when there is cancer?
    ▪ Preferences can be quantified by a Utility Function, or a Cost Function.
Utility Function / Cost Function

• **Utility Function**:
  – Quantifies an agent’s utility from (happiness with) a given outcome.
  – Rational agents act to maximize expected utility.
  – **Expected Utility** of action $A = a$, resulting in outcomes $B = b$:
    ▪ $\text{Expected Utility} = \sum_b P(b|a) \times \text{Utility}(b)$

• **Cost Function**:
  – Quantifies an agent’s cost from (unhappiness with) a given outcome.
  – Rational agents act to minimize expected cost.
  – **Expected Cost** of action $a$, resulting in outcomes $o$:
    ▪ $\text{Expected Cost} = \sum_b P(b|a) \times \text{Cost}(b)$
Decision Theory: Why Probabilities are Useful

• Utility associated with various outcomes:
  – Operate = true, Cancer = true: utility = 30
  – Operate = true, Cancer = false: utility = -50
  – Operate = false, Cancer = true: utility = -100
  – Operate = false, Cancer = false: utility = 0

• Expected utility of actions:
  – \( P(c) = P(\text{Cancer} = \text{true}) \) \( <--- \) for simplicity
  – \( E[\text{utility}(\text{Operate} = \text{true})] = 30 \, P(c) - 50 \, [1-P(c)] \)
  – \( E[\text{utility}(\text{Operate} = \text{false})] = -100 \, P(c) \)

• Break even point?
  – \( 30 \, P(c) - 50 + 50 \, P(c) = -100 \, P(c) \)
  – \( P(c) = 50/180 \approx 0.28 \)
  – If \( P(c) > 0.28 \), the optimal decision (highest expected utility) is to operate!
Independence

• **Formal Definition:**
  – 2 random variables A and B are **independent** iff:
    \[ P(a, b) = P(a) \ P(b), \text{ for all values } a, b \]

• **Informal Definition:**
  – 2 random variables A and B are **independent** iff:
    \[ P(a \mid b) = P(a) \text{ OR } P(b \mid a) = P(b), \text{ for all values } a, b \]
    – \( P(a \mid b) = P(a) \) tells us that knowing b provides no change in our probability for a, and thus b contains no information about a.

• Also known as **marginal independence**, as all other variables have been marginalized out.

• In practice true independence is very rare:
  – “butterfly in China” effect
  – Conditional independence is much more common and useful
Conditional Independence

• **Formal Definition:**
  – 2 random variables A and B are *conditionally independent* given C iff:
    \[ P(a, b | c) = P(a | c) P(b | c), \text{ for all values } a, b, c \]

• **Informal Definition:**
  – 2 random variables A and B are *conditionally independent* given C iff:
    \[ P(a | b, c) = P(a | c) \quad \text{OR} \quad P(b | a, c) = P(b | c), \text{ for all values } a, b, c \]
  – \( P(a | b, c) = P(a | c) \) tells us that learning about b, given that we already know c, provides no change in our probability for a, and thus b contains no information about a beyond what c provides.

• **Naïve Bayes Model:**
  – Often a single variable can directly influence a number of other variables, all of which are conditionally independent, given the single variable.
  – E.g., k different symptom variables \( X_1, X_2, \ldots X_k \), and C = disease, reducing to:
    \[ P(X_1, X_2, \ldots X_k \mid C) = \prod P(X_i \mid C) \]
Conditional Independence vs. Independence

• For Example:
  – $A = \text{height}$
  – $B = \text{reading ability}$
  – $C = \text{age}$

  – $P(\text{reading ability} \mid \text{age, height}) = P(\text{reading ability} \mid \text{age})$
  – $P(\text{height} \mid \text{reading ability, age}) = P(\text{height} \mid \text{age})$

• Note:
  – Height and reading ability are dependent (not independent) but are conditionally independent given age
In each group, symptom 1 and symptom 2 are conditionally independent.

But clearly, symptom 1 and 2 are marginally dependent (unconditionally).
Putting It All Together

• Full joint distributions can be difficult to obtain:
  – Vast quantities of data required, even with relatively few variables
  – Data for some combinations of probabilities may be sparse

• Determining independence and conditional independence allows us to decompose our full joint distribution into much smaller pieces:
  – e.g., \( P(\text{Toothache}, \text{Catch}, \text{Cavity}) \)
    \[ = P(\text{Toothache}, \text{Catch}|\text{Cavity}) \cdot P(\text{Cavity}) \]
    \[ = P(\text{Toothache}|\text{Cavity}) \cdot P(\text{Catch}|\text{Cavity}) \cdot P(\text{Cavity}) \]

• All three variables are Boolean.
• Before conditional independence, requires \( 2^3 \) probabilities for full specification:
  --> Space Complexity: \( O(2^n) \)
• After conditional independence, requires 3 probabilities for full specification:
  --> Space Complexity: \( O(n) \)
Conclusions...

• Representing uncertainty is useful in knowledge bases.

• Probability provides a framework for managing uncertainty.

• Using a full joint distribution and probability rules, we can derive any probability relationship in a probability space.

• Number of required probabilities can be reduced through independence and conditional independence relationships.

• Probabilities allow us to make better decisions by using decision theory and expected utilities.

• **Rational agents cannot violate probability theory.**