Probability and Uncertainty

Bayesian Networks

First Lecture Today (Tue 28 Jun)
Review Chapters 8.3-8.5,
Read 9.1-9.2 (optional: 9.5)

Second Lecture Today (Tue 28 Jun)
Read Chapters 13, 14.1-14.5

Next Lecture (Thu 30 Jun)
Read Chapters 3.1-3.4, 18.6.1-2, 20.3.1

(Please read lecture topic material before and after each lecture on that topic)
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
Propositional Logic and Probability

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

    Ontology is the philosophical study of the nature of being, becoming, existence, or reality

• 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)

    Epistemology: the branch of philosophy concerned with the nature and scope of knowledge
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)  =>  Σ_A 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.
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**
  – $P(a)$, the probability of “a” being true, or $P(a=True)$
  – Does not depend on anything else to be true (**unconditional**)
  – Represents the probability prior to further information that may adjust it (**prior**)

• **Conditional Probability**
  – $P(a|b)$, the probability of “a” being true, given that “b” is true (**conditional**)
  – 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 $(10^\circ, 200^\circ)$

• **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.
    - \(P(a) \equiv P(A = a)\)
    - \(P(a | b) \equiv P(A = a | B = b)\)
    - \(P(a, b) \equiv P(A = a, B = b)\)

- **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:
Prob = \( P(A) \)

Prob = 1 - \( P(A) \)

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

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

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

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

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

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

Area = Probability of Event
Using the Product Rule

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

- **Factoring:** (AKA Chain Rule for probabilities)
  - By the product rule, we can always write:
    \( P(a, b, c, \ldots z) = P(a | b, c, \ldots z) \ P(b, c, \ldots z) \)
  - Repeatedly applying this idea, we can write:
    \( P(a, b, c, \ldots z) = P(a | b, c, \ldots z) \ P(b | c, \ldots z) \ P(c | \ldots z) \ldots P(z) \)
  - 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 \( P(\text{CatchFish}, \text{Day}, \text{Lake}) \)
  – **Where:**
    ▪ \( \text{CatchFish} = \{\text{true, false}\} \)
    ▪ \( \text{Day} = \{\text{mon, tues, wed, thurs, fri, sat, sun}\} \)
    ▪ \( \text{Lake} = \{\text{buel lake, ralph lake, crystal lake}\} \)

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

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

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

Area = Probability of Event

\[ P(A \land B) = P(A) + P(B) - P(A \lor B) \]
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):**
  
  \[- P(a|b) = P(b|a) P(a) / P(b) \quad <-- \text{Bayes’ Rule} \]
Who’s Bayes?

• **English theologian** and mathematician Thomas Bayes has greatly contributed to the field of probability and statistics. **His ideas have created much controversy and debate among statisticians over the years.**

• Bayes wrote a number of papers that discussed his work. However, **the only ones known to have been published while he was still living are: Divine Providence and Government Is the Happiness of His Creatures** (1731) and **An Introduction to the Doctrine of Fluxions, and a Defense of the Analyst** (1736)...

  http://mnstats.morris.umn.edu/introstat/history/w98/Bayes.html
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 \( \text{Windspeed} = \{0-10\text{mph}, 10-20\text{mph}, 20-30\text{mph}, \text{etc.}\} \)

• **Bayes’ Rule:**
  - \( P(b|a) = \frac{P(a|b) P(b)}{P(a)} \)
  - 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.
Independence

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

• **Informal Definition:**
  - 2 random variables A and B are independent iff:
    \[ P(a \mid b) = P(a) \quad \text{OR} \quad P(b \mid a) = P(b), \quad \text{for all values } a, b \]
  - \( P(a \mid b) = P(a) \) 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:
    \[ P(a, b | c) = P(a | c) P(b | c), \quad \text{for all values } a, b, c \]

- **Informal Definition:**
  - 2 random variables $A$ and $B$ are conditionally independent given $C$ iff:
    \[ P(a | b, c) = P(a | c) \quad \text{OR} \quad P(b | a, c) = P(b | c), \quad \text{for all values } a, b, c \]
  - $P(a | b, c) = P(a | c)$ 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 $C = \text{disease}$, reducing to:
    \[ P(X_1, X_2, \ldots, X_k \mid C) = \prod P(X_i \mid C) \]
Full Joint vs Conditional Independence

• Example: 4 Binary Random Variable (A, B, C, D)
  – Full Joint Probability Table
    • 1 Table with 16 rows
  – Conditional Independence
    • \( P(A, B, C, D) = P(A) \cdot P(B \mid A) \cdot P(C \mid A, B) \cdot P(D \mid A, B, C) \) (no saving yet..)
    • if... \( P(D \mid A, B) = P(C \mid A), P(D \mid A, B, C) = P(D \mid A) \) [Naïve Bayes Model]
      – \( P(A, B, C, D) = P(A) \cdot P(B \mid A) \cdot P(C \mid A) \cdot P(D \mid A) \)
      – 4 Tables. With at most 4 rows

• If we had N Binary Random Variables
  – Full Joint Probability Table
    • 1 Table with \( 2^N \) Rows; \( N = 100, 2^{100} \approx 10^{30} \)
  – Naïve Bayes Model (Conditional Independence)
    • \( N \) tables with at most 4 rows!
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., \( P(\text{Toothache, Catch, Cavity}) \)
    \[ = P(\text{Toothache, Catch} | \text{Cavity}) \cdot P(\text{Cavity}) \]
    \[ = P(\text{Toothache} | \text{Cavity}) \cdot P(\text{Catch} | \text{Cavity}) \cdot P(\text{Cavity}) \]

• All three variables are Boolean.
• Before conditional independence, requires \( 2^3 \) probabilities for full specification:
  \[ \text{--> Space Complexity: } O(2^n) \]
• After conditional independence, requires 3 probabilities for full specification:
  \[ \text{--> Space Complexity: } O(n) \]
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.**
• **Snake Robot Climbs a Tree**
  https://www.youtube.com/watch?v=8VLjDjXzTiU

• **Asterisk - Omni-directional Insect Robot Picks Up Prey #DigInfo**
  https://www.youtube.com/watch?v=kMF83m8lNrw

• **Freaky AI robot, taken from Nova science now**
  https://www.youtube.com/watch?v=UIWWLg4wLEY
Bayesian Networks

Read R&N Ch. 14.1-14.2

Next lecture: Read R&N 18.1-18.4
You will be expected to know

• Basic concepts and vocabulary of Bayesian networks.
  – Nodes represent random variables.
  – Directed arcs represent (informally) direct influences.
  – Conditional probability tables, $P( X_i \mid \text{Parents}(X_i) )$.

• Given a Bayesian network:
  – Write down the full joint distribution it represents.

• Given a full joint distribution in factored form:
  – Draw the Bayesian network that represents it.

• Given a variable ordering and some background assertions of conditional independence among the variables:
  – Write down the factored form of the full joint distribution, as simplified by the conditional independence assertions.
Bayesian Networks

JUDEA PEARL
United States – 2011
Lecture Video

THE MECHANIZATION OF CAUSAL INFERENCE:
A "mini" Turing Test And Beyond
Judea Pearl
Department of Computer Science
UCLA
Why Bayesian Networks?

• Knowledge Representation & Reasoning (Inference)
  – Propositional Logic
    • Knowledge Base : Propositional Logic Sentences
    • Reasoning : KB |= Theory
      – Find a model or Count models
  – Probabilistic Reasoning
    • Knowledge Base : Full Joint Probability Distribution over All Random Variables
    • Reasoning: Compute Pr ( KB |= Theory )
      – Find the most probable assignments
      – Compute marginal / conditional probability

• Why Bayesian Net?
  – Manipulating full joint probability distribution is Very Hard!
  – We want to exploit a special property of probability distribution, conditional independence
  – Bayesian Network captures conditional independence
    • Graphical Representation (Probabilistic Graphical Models)
    • Tool for Reasoning, Computation (Probabilistic Reasoning bases on the Graph)
Extended example of 3-way Bayesian Networks

Common Cause
A: Fire
B: Heat
C: Smoke

Conditionally independent effects:
\[ p(A, B, C) = p(B|A)p(C|A)p(A) \]

B and C are conditionally independent
Given A

“Where there’s Smoke, there’s Fire.”

If we see Smoke, we can infer Fire.

If we see Smoke, observing Heat tells us very little additional information.
Extended example of 3-way Bayesian Networks

Suppose I build a fire in my fireplace about once every 10 days...

<table>
<thead>
<tr>
<th>Fire</th>
<th>P(Smoke)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.90</td>
</tr>
<tr>
<td>f</td>
<td>.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fire</th>
<th>P(Heat)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.99</td>
</tr>
<tr>
<td>f</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Conditionally independent effects:

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

Smoke and Heat are conditionally independent given Fire.

If we see B=Smoke, observing C=Heat tells us very little additional information.
Let’s Compute Probability of…
“Where there’s smoke, there’s (probably) fire.”

What is P(Fire=t | Smoke=t)?
P(Fire=t | Smoke=t)
= P(Fire=t & Smoke=t) / P(Smoke=t)

Extended example of 3-way Bayesian Networks

Definition of conditional probability

<table>
<thead>
<tr>
<th>Fire</th>
<th>P(Smoke)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.90</td>
</tr>
<tr>
<td>f</td>
<td>.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fire</th>
<th>P(Heat)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.99</td>
</tr>
<tr>
<td>f</td>
<td>.0001</td>
</tr>
</tbody>
</table>
Extended example of 3-way Bayesian Networks

What is $P(\text{Fire}=t \& \text{Smoke}=t)$?

$P(\text{Fire}=t \& \text{Smoke}=t)$

$= \sum_{\text{heat}} P(\text{Fire}=t \& \text{Smoke}=t \& \text{heat})$

$= \sum_{\text{heat}} P(\text{Smoke}=t \& \text{heat}|\text{Fire}=t)P(\text{Fire}=t)$

$= \sum_{\text{heat}} P(\text{Smoke}=t|\text{Fire}=t)P(\text{heat}|\text{Fire}=t)P(\text{Fire}=t)$

$= P(\text{Smoke}=t|\text{Fire}=t) P(\text{heat}=t|\text{Fire}=t)P(\text{Fire}=t)$

$+ P(\text{Smoke}=t|\text{Fire}=t)P(\text{heat}=f|\text{Fire}=t)P(\text{Fire}=t)$

$= (.90 \times .99 \times .1) + (.90 \times .01 \times .1)$

$= 0.09$

<table>
<thead>
<tr>
<th>Fire</th>
<th>$P(\text{Smoke})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.90</td>
</tr>
<tr>
<td>f</td>
<td>.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fire</th>
<th>$P(\text{Heat})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.99</td>
</tr>
<tr>
<td>f</td>
<td>.0001</td>
</tr>
</tbody>
</table>
What is $P(\text{Smoke}=t)$?

$P(\text{Smoke}=t)$

$= \sum_{\text{fire}} \sum_{\text{heat}} P(\text{Smoke}=t \& \text{fire} \& \text{heat})$

$= \sum_{\text{fire}} \sum_{\text{heat}} P(\text{Smoke}=t \& \text{heat}|\text{fire}) P(\text{fire})$

$= \sum_{\text{fire}} \sum_{\text{heat}} P(\text{Smoke}=t|\text{fire}) P(\text{heat}|\text{fire}) P(\text{fire})$

$= P(\text{Smoke}=t|\text{fire}=t) P(\text{heat}=t|\text{fire}=t) P(\text{fire}=t)$

$+ P(\text{Smoke}=t|\text{fire}=t) P(\text{heat}=f|\text{fire}=t) P(\text{fire}=t)$

$+ P(\text{Smoke}=t|\text{fire}=f) P(\text{heat}=t|\text{fire}=f) P(\text{fire}=f)$

$+ P(\text{Smoke}=t|\text{fire}=f) P(\text{heat}=f|\text{fire}=f) P(\text{fire}=f)$

$\approx 0.0909$
What is $P(\text{Fire}=t \mid \text{Smoke}=t)$?

$P(\text{Fire}=t \mid \text{Smoke}=t)$

$= P(\text{Fire}=t \& \text{Smoke}=t) / P(\text{Smoke}=t)$

$\approx 0.09 / 0.0909$

$\approx 0.99$

So we’ve just proven that

“Where there’s smoke, there’s (probably) fire.”
Bayesian Networks

- Structure of the graph ⇔ Conditional independence relations

In general,

\[ p(X_1, X_2, \ldots, X_N) = \prod p(X_i \mid \text{parents}(X_i)) \]

- Requires that graph is acyclic (no directed cycles)

- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

- Also known as belief networks, graphical models
Bayesian Network

- A Bayesian network specifies a joint distribution in a structured form:

  \[ p(A,B,C) = p(C|A,B)p(A)p(B) \]

- Dependence/independence represented via a directed graph:
  - Node = random variable
  - Directed Edge = conditional dependence
  - Absence of Edge = conditional independence

- Allows concise view of joint distribution relationships:
  - Graph nodes and edges show conditional relationships between variables.
  - Tables provide probability data.
Examples of 3-way Bayesian Networks

Marginal Independence:
\[ p(A,B,C) = p(A) \cdot p(B) \cdot p(C) \]

Nodes: Random Variables
A, B, C

Edges: \( P(X_i \mid \text{Parents}) \rightarrow \) Directed edge from parent nodes to \( X_i \)
No Edge!
Examples of 3-way Bayesian Networks

Independent Causes:
A Earthquake
B Burglary
C Alarm

p(A, B, C) = p(C | A, B)p(A)p(B)

"Explaining away" effect:
Given C, observing A makes B less likely
e.g., earthquake/burglary/alarm example

A and B are (marginally) independent
but become dependent once C is known

You heard alarm, and observe Earthquake
…. It explains away burglary

Nodes: Random Variables
A, B, C

Edges: P(Xi | Parents) → Directed edge from parent nodes to Xi
A → C
B → C
Examples of 3-way Bayesian Networks

Markov Dependence
A Rain on Mon
B Ran on Tue
C Rain on Wed

Markov dependence:
\[ p(A, B, C) = p(C|B) \ p(B|A)p(A) \]

A affects B and B affects C
Given B, A and C are independent

e.g.
If it rains today, it will rain tomorrow with 90%

On Wed morning...
If you know it rained yesterday, it doesn’t matter whether it rained on Mon

Nodes: Random Variables
A, B, C

Edges: \( P(X_i \mid \text{Parents}) \) → Directed edge from parent nodes to \( X_i \)
A → B
B → C
Inference in Bayesian Networks

Simple Example

P(A) = 0.05  
P(B) = 0.02

Disease1  Disease2

A B P(C|A,B)

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>t</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>.90</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>.90</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>.005</td>
<td></td>
</tr>
</tbody>
</table>

C P(D|C)

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.90</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>.002</td>
<td></td>
</tr>
</tbody>
</table>

D Fever

(A=True, B=False | D=True) : Probability of getting Disease1 when we observe Fever

Note: Not an anatomically correct model of how diseases cause fever!

Suppose that two different diseases influence some imaginary internal body temperature regulator, which in turn influences whether fever is present.
Inference in Bayesian Networks

- \( X = \{ X_1, X_2, ..., X_k \} = \text{query variables} \) of interest
- \( E = \{ E_1, ..., E_l \} = \text{evidence variables} \) that are observed
- \( Y = \{ Y_1, ..., Y_m \} = \text{hidden variables} \) (nonevidence, nonquery)

- What is the posterior distribution of \( X \), given \( E \)?
  \[- P( X | e ) = \alpha \sum_y P( X, y, e ) \]
  Normalizing constant \( \alpha = \sum_x \sum_y P( X, y, e ) \)

- What is the most likely assignment of values to \( X \), given \( E \)?
  \[- \arg\max_x P( x | e ) = \arg\max_x \sum_y P( x, y, e ) \]
Given a graph, can we “read off” conditional independencies?

The “Markov Blanket” of X (the gray area in the figure)

X is conditionally independent of everything else, GIVEN the values of:
* X’s parents
* X’s children
* X’s children’s parents

X is conditionally independent of its non-descendants, GIVEN the values of its parents.

Bayesian Network Captures Conditional Independence…
From Graphical Model, what variables are relevant when you query on X?
Pr( X | all other nodes ) = Pr( X | parents, children, children’s parent)
Naïve Bayes Model

\[ P(C \mid X_1, \ldots, X_n) = \alpha \prod P(X_i \mid C) \cdot P(C) \]

Features \( X \) are conditionally independent given the class variable \( C \)

Widely used in machine learning
  e.g., spam email classification: \( X \)'s = counts of words in emails

Probabilities \( P(C) \) and \( P(X_i \mid C) \) can easily be estimated from labeled data
Naïve Bayes Model (2)

\[ P(C | X_1, \ldots X_n) = \alpha \prod P(X_i | C) P(C) \]

<Learning Naïve Bayes Model>

Probabilities \( P(C) \) and \( P(X_i | C) \) can easily be estimated from labeled data

\[ P(C = c_j) \approx \frac{\text{#(Examples with class label } c_j)}{\text{#(Examples)}} \]
\[ P(X_i = x_{ik} | C = c_j) \approx \frac{\text{#(Examples with } X_i \text{ value } x_{ik} \text{ and class label } c_j)}{\text{#(Examples with class label } c_j)} \]

Usually easiest to work with logs

\[ \log \left[ P(C | X_1, \ldots X_n) \right] = \log \alpha + \sum \left[ \log P(X_i | C) + \log P(C) \right] \]

DANGER: Suppose ZERO examples with \( X_i \) value \( x_{ik} \) and class label \( c_j \)?
An unseen example with \( X_i \) value \( x_{ik} \) will NEVER predict class label \( c_j \)!

Practical solutions: Pseudocounts, e.g., add 1 to every \( \#() \), etc.
Theoretical solutions: Bayesian inference, beta distribution, etc.
Hidden Markov Model (HMM)

Two key assumptions:
1. hidden state sequence is Markov
2. observation $Y_t$ is **Conditionally Independent** of all other variables given $S_t$

Widely used in speech recognition, protein sequence models
Since this is a Bayesian network polytree, inference is linear in $n$, Viterbi Algorithm. (Special Graph Structure Allows Efficient Computation!)
Summary

• Bayesian networks represent a joint distribution using a graph

• The graph encodes a set of conditional independence assumptions

• Answering queries (or inference or reasoning) in a Bayesian network amounts to efficient computation of appropriate conditional probabilities

• Probabilistic inference is intractable in the general case
  – But can be carried out in linear time for certain classes of Bayesian networks