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 coherent framework for uncertainty
- Review of basic concepts in probability.
  - Emphasis on conditional probability and conditional independence
- Full joint distributions are intractable to work with.
  - Conditional independence assumptions allow much simpler models
- Bayesian networks are a systematic way to construct:
  parsimonious and structured probability distributions
- **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.
- Using Bayes' rule.
  - Naïve Bayes model (naïve Bayes classifier)

Complete architectures for intelligence?

- Search?
  - Solve the problem of what to do.
- Learning?
  - Learn what to do.
- Logic and inference?
  - Reason about what to do.
  - Encoded knowledge/”expert” systems?
    - Know what to do.
- Modern view: It’s complex & multi-faceted.

A (very brief) History of Probability in AI

- Early AI (1950's and 1960's)
  - Attempts to solve AI problems using probability met with mixed success
- Logical AI (1970's, 80's)
  - Recognized that working with full probability models is intractable
  - Abandoned probabilistic approaches
  - Focused on logic-based representations
  - Problem: Pure logic is "brittle" when applied to real-world problems.
- Probabilistic AI (1990's-present)
  - Judea Pearl invents Bayesian networks in 1988
  - Realization that approximate probability models are tractable and useful
  - Development of machine learning techniques to learn such models from data
  - Probabilistic techniques now widely used in vision, speech recognition, robotics, language modeling, game-playing, etc

Uncertainty

Let action *At* = leave for airport *t* minutes before flight
Will *At* get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood: "*A25*will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"*A25* 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."

"*A1440* should get me there on time but I'd have to stay overnight in the airport."

Methods for handling uncertainty

- Default or nonmonotonic logic:
  - –Assume my car does not have a flat tire
  - –Assume A25 works unless contradicted by evidence
- Issues: What assumptions are reasonable?
  How to handle contradictions?

- Rules with fudge factors:
  - –A25 => 0.3 get there on time
  - –Sprinkler => 0.99 WetGrass
  - –WetGrass => 0.7 Rain
- Issues: Problems with combination, e.g., Sprinkler causes Rain??

- Probability
  - – Model agent's degree of belief
  - – Given the available evidence,
    A25 will get me there on time with probability 0.04

Probability

- Probabilistic assertions summarize effects of
  –laziness: failure to enumerate exceptions, qualifications, etc.
  –ignorance: lack of relevant facts, initial conditions, etc.

- Subjective probability:
  - Probabilities relate propositions to agent's own state of knowledge
    - e.g., P(A25 | no reported accidents) = 0.06

- These are not assertions about the world
  - They indicate degrees of belief in assertions about the world

- Probabilities of propositions change with new evidence:
  - e.g., P(A25 | no reported accidents, 5 a.m.) = 0.15

Making decisions under uncertainty

- Suppose I believe the following:
  - P(A25 gets me there on time | …) = 0.04
  - P(A90 gets me there on time | …) = 0.70
  - P(A120 gets me there on time | …) = 0.95
  - P(A1440 gets me there on time | …) = 0.9999

- Which action to choose?

  Depends on my preferences for missing flight vs. time spent waiting, etc.

  – Utility theory is used to represent and infer preferences

  – Decision theory = probability theory + utility theory

- Expected utility of action a in state s

  $= \sum_{\text{outcome in Results(s,a)}} P(\text{outcome}) * \text{Utility(outcome)}$
- A rational agent acts to maximize expected utility

Making decisions under uncertainty (Example)

- Suppose I believe the following:
  - P(A25 gets me there on time | …) = 0.04
  - P(A90 gets me there on time | …) = 0.70
  - P(A120 gets me there on time | …) = 0.95
  - P(A1440 gets me there on time | …) = 0.9999
  - Utility(on time) = $1,000
  - Utility(not on time) = −$10,000

- Expected utility of action a in state s
  = $\sum_{outcome \in Results(s,a)} P(outcome) * Utility(outcome)$
  - E(Utility(A25)) = 0.04*$1,000 + 0.96*(−$10,000) = −$9,560
  - E(Utility(A90)) = 0.7*$1,000 + 0.3*(−$10,000) = −$2,300
  - E(Utility(A120)) = 0.95*$1,000 + 0.05*(−$10,000) = $450
  - E(Utility(A1440)) = 0.9999*$1,000 + 0.0001*(−$10,000) = $998.90

- Have not yet accounted for disutility of staying overnight at airport, etc.

Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.

- Boolean random variables

  e.g., *Cavity (= do I have a cavity?)*
- Discrete random variables

  e.g., *Weather is one of <sunny,rainy,cloudy,snow>*
- Domain values must be exhaustive and mutually exclusive

- Elementary proposition is an assignment of a value to a random variable:

  e.g., *Weather = sunny; Cavity = false(abbreviated as ¬cavity)*

- Complex propositions formed from elementary propositions and standard logical connectives :

  e.g., *Weather = sunny ∨ Cavity = false*

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
  - P(a) = "prior" or marginal or unconditional probability
  - Assumes no other information is available

- Axioms:
  - 0 <= P(a) <= 1
  - P(NOT(a)) = 1 – P(a)
  - 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 sub-optimally in some cases
  - e.g., de Finetti proved that there will be some combination of bets that forces such an unhappy agent to lose money every time.

- **Rational agents <u>cannot</u> violate probability theory.**

Probability and Logic

- Probability can be viewed as a generalization of propositional logic

- P(*a*):
  - a is any sentence in propositional logic
  - Belief of agent in *a* is no longer restricted to *true, false, unknown*
  - P(a) can range from 0 to 1
    - P(a) = 0, and P(a) = 1 are special cases
    - So logic can be viewed as a special case of probability

Conditional Probability

- P(a|b) is the conditional probability of proposition a, conditioned on knowing that b is true,
  - E.g., P(rain in London tomorrow | raining in London today)
  - P(a|b) is a "posterior" or conditional probability
  - The updated probability that a is true, now that we know b
  - P(a|b) = P(a AND b) / P(b)
  - Syntax: P(a | b) is the probability of a given that b is true
    - a and b can be any propositional sentences
    - e.g., p(John wins OR Mary wins | Bob wins AND Jack loses)

- P(a|b) obeys the same rules as probabilities,
  - E.g., P(a | b) + P(NOT(a) | b) = 1
  - All probabilities in effect are conditional probabilities
    - E.g., P(a) = P(a | our background knowledge)

Random Variables

- A is a random variable taking values a_1, a_2, … a_m
  - Events are A= a_1, A= a_2, ….
  - We will focus on discrete random variables

- Mutual exclusion

  $P(A = a_i \text{ AND } A = a_j) = 0$

- Exhaustive

  $\Sigma\ P(a_i) = 1$

MEE (Mutually Exclusive and Exhaustive) assumption is often useful
(but not always appropriate, e.g., disease-state for a patient)

For finite m, can represent P(A) as a table of m probabilities

For infinite m (e.g., number of tosses before “heads”) we can represent P(A) by a function (e.g., geometric)

Joint Distributions

- Consider 2 random variables: A, B
  - P(a, b) is shorthand for P(A = a AND B=b)
  - $\Sigma_a \Sigma_b P(a, b) = 1$
  - Can represent P(A, B) as a table of m^2 numbers

- Generalize to more than 2 random variables
  - E.g., A, B, C, … Z
  - $\Sigma_a \Sigma_b \ldots \Sigma_z P(a, b, \ldots, z) = 1$
  - P(A, B, …. Z) is a table of m^K numbers, K = # variables
    - This is a potential problem in practice, e.g., m=2, K = 20

Linking Joint and Conditional Probabilities

- Basic fact:

  $P(a, b) = P(a \mid b)\, P(b)$

  - Why? Probability of a and b occurring is the same as probability of a occurring given b is true, times the probability of b occurring

- Bayes rule:

  $P(a, b) = P(a \mid b)\, P(b)$

  $= P(b \mid a)\, P(a)$ by definition

  $\Rightarrow P(b \mid a) = P(a \mid b)\, P(b) \,/\, P(a)$ [Bayes rule]

  Why is this useful?

  Often much more natural to express knowledge in a particular "direction", e.g., in the causal direction

  e.g., b = disease, a = symptoms

  More natural to encode knowledge as $P(a|b)$ than as $P(b|a)$

Using Bayes Rule

- Example:
  - P(stiff neck | meningitis) = 0.5 (prior knowledge from doctor)
  - P(meningitis) = 1/50,000 and P(stiff neck) = 1/20
    (e.g., obtained from large medical data sets)

P(m | s) = P(s | m) P(m) / P(s)
= [0.5 * 1/50,000] / [1/20] = 1/5000

So given a stiff neck, and no other information,
p(meningitis|stiff neck) is pretty small

But note that its 10 times more likely that it was before
- so it might be worth measuring more variables for this patient

More Complex Examples with Bayes Rule

- $P(a \mid b, c) = ??$
  $= P(b, c \mid a)\, P(a) \;/\; P(b,c)$

- $P(a, b \mid c, d) = ??$
  $= P(c, d \mid a, b)\, P(a, b) \;/\; P(c, d)$

Both are examples of basic pattern $p(x|y) = p(y|x)p(x)/p(y)$

(it helps to group variables together, e.g., $y = (a,b)$, $x = (c, d)$)

Note also that we can write $P(x \mid y)$ is proportional to $P(y \mid x)\, P(x)$
(the $P(y)$ term on the bottom is just a normalization constant)

Sequential Bayesian Reasoning

- h = hypothesis, e1, e2, .. en = evidence
- P(h) = prior
- P(h | e1) proportional to P(e1 | h) P(h)
  = likelihood of e1 x prior(h)
- P(h | e1, e2) proportional to P(e1, e2 | h) P(h)
  in turn can be written as P(e2| h, e1) P(e1|h) P(h)
  ~ likelihood of e2 x "prior"(h given e1)
- Bayes rule supports sequential reasoning
  - Start with prior P(h)
  - New belief (posterior) = P(h | e1)
  - This becomes the "new prior"
  - Can use this to update to P(h | e1, e2), and so on…..

Computing with Probabilities: Law of Total Probability

Law of Total Probability (aka "summing out" or marginalization)

$$P(a) = \Sigma_b P(a, b)$$
$$= \Sigma_b P(a \mid b) P(b)$$ where B is any random variable

Why is this useful?

Given a joint distribution (e.g., P(a,b,c,d)) we can obtain any "marginal" probability (e.g., P(b)) by summing out the other variables, e.g.,

$$P(b) = \Sigma_a \Sigma_c \Sigma_d P(a, b, c, d)$$

We can compute <u>any conditional probability</u> given a joint distribution, e.g.,

$$P(c \mid b) = \Sigma_a \Sigma_d P(a, c, d \mid b)$$
$$= \Sigma_a \Sigma_d P(a, c, d, b) / P(b)$$

where P(b) can be computed as above

Computing with Probabilities: The Chain Rule or Factoring

We can always write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

(by definition of joint probability)

Repeatedly applying this idea, we can write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c, .. z)\, P(c \mid .. z) .. P(z)$$

This factorization holds for any ordering of the variables

This is the chain rule for probabilities

What does all this have to do with AI?

- Logic-based knowledge representation
  - Set of sentences in KB
  - Agent's belief in any sentence is: true, false, or unknown
- In real-world problems there is uncertainty
  - P(snow in New York on January 1) is not 0 or 1 or unknown
  - P(vehicle speed > 50 | sensor reading)
  - P(Dow Jones will go down tomorrow | data so far)
  - P(pit in square 2,2 | evidence so far)
  - Not acknowledging this uncertainty can lead to brittle systems and inefficient use of information
- Uncertainty is due to:
  - Things we did not measure (which is always the case)
    - E.g., in economic forecasting
  - Imperfect knowledge
    - P(symptom | disease) -> we are not 100% sure
  - Noisy measurements
    - P(speed > 50 | sensor reading > 50) is not 1

Agents, Probabilities, and Degrees of Belief

- What we were taught in school
  - P(*a*) represents the frequency that event *a* will happen in repeated trials
  - -> "relative frequency" interpretation

- Degree of belief
  - P(*a*) represents an agent's degree of belief that event *a* is true
  - This is a more general view of probability
    - Agent's probability is based on what information they have
    - E.g., based on data or based on a theory

- 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)?

- Probabilities can vary from agent to agent depending on their models of the world and how much data they have

More on Degrees of Belief

- Our interpretation of P(a | e) is that it is an agent's degree of belief in the proposition a, given evidence e
  - Note that proposition a is true or false in the real-world
  - P(a|e) reflects the agent's uncertainty or ignorance

- The degree of belief interpretation does not mean that we need new or different rules for working with probabilities
  - The same rules (Bayes rule, law of total probability, probabilities sum to 1) still apply – our interpretation is different

- If Agent 1 has inconsistent sets of probabilities (violate axioms of probability theory) then there exists a betting strategy that allows Agent 2 to always win in bets against Agent 1
  - See Section 13.2 in text, de Finetti's argument

Maximizing expected utility (or minimizing expected cost)

- What action should the agent take?

- A rational agent should maximize expected utility, or equivalently minimize expected cost

- Expected cost of actions:

  $E[\text{ cost(a) }] = 30\ p(c) - 50\ [1 - p(c)\ ]$

  $E[\text{ cost(b) }] = -100\ p(c)$

Break even point? $30p - 50 + 50p = -100p$

$100p + 30p + 50p = 50$

$\Rightarrow p(c) = 50/180 \sim 0.28$

If $p(c) > 0.28$, the optimal decision is to operate

- Original theory from economics, cognitive science (1950's)
  - But widely used in modern AI, e.g., in robotics, vision, game-playing

- Note that we can only make optimal decisions if we know the probabilities

Constructing a Propositional Probabilistic Knowledge Base

- Define all variables of interest: A, B, C, … Z
- Define a joint probability table for P(A, B, C, … Z)
  - We have seen earlier how this will allow us to compute the answer to any query, p(query | evidence),
    where query and evidence = any propositional sentence
- 2 major problems:
  - Computation time:
    - P(a|b) requires summing out over all other variables in the model, e.g., $O(m^{K-1})$ with K variables
  - Model specification
    - Joint table has $O(m^K)$ entries – where will all the numbers come from?
  - These 2 problems effectively halted the use of probability in AI research from the 1960's up until about 1990

Independence

- 2 random variables A and B are independent iff

  $P(a, b) = P(a)\, P(b)$ for all values a, b

- More intuitive (equivalent) conditional formulation
  - A and B are independent iff

    $P(a \mid b) = P(a)$ OR $P(b \mid a) = P(b)$, for all values a, b

  - Intuitive interpretation:

    $P(a \mid b) = P(a)$ tells us that knowing b provides no change in our probability for a, i.e., b contains no information about a

- Can generalize to more than 2 random variables

- In practice true independence is very rare
  - "butterfly in China" effect
  - Weather and dental example in the text
  - Conditional independence is much more common and useful

- Note: independence is an assumption we impose on our model of the world - it does not follow from basic axioms

Conditional Independence

- 2 random variables A and B are conditionally independent given C iff

$$P(a, b \mid c) = P(a \mid c)\, P(b \mid c) \quad \text{for all values a, b, c}$$

- More intuitive (equivalent) conditional formulation
  - A and B are conditionally independent given C iff

    $P(a \mid b, c) = P(a \mid c)$ OR $P(b \mid a, c) = P(b \mid c)$, for all values a, b, c

  - Intuitive interpretation:

    $P(a \mid b, c) = P(a \mid c)$ tells us that learning about b, given that we already know c, provides no change in our probability for a,

    i.e., b contains no information about a beyond what c provides

- Can generalize to more than 2 random variables
  - E.g., K different symptom variables X1, X2, … XK, and C = disease
  - $P(X1, X2, \ldots XK \mid C) = \prod P(Xi \mid C)$
  - Also known as the naïve Bayes assumption

Conditional Independence vs. Independence

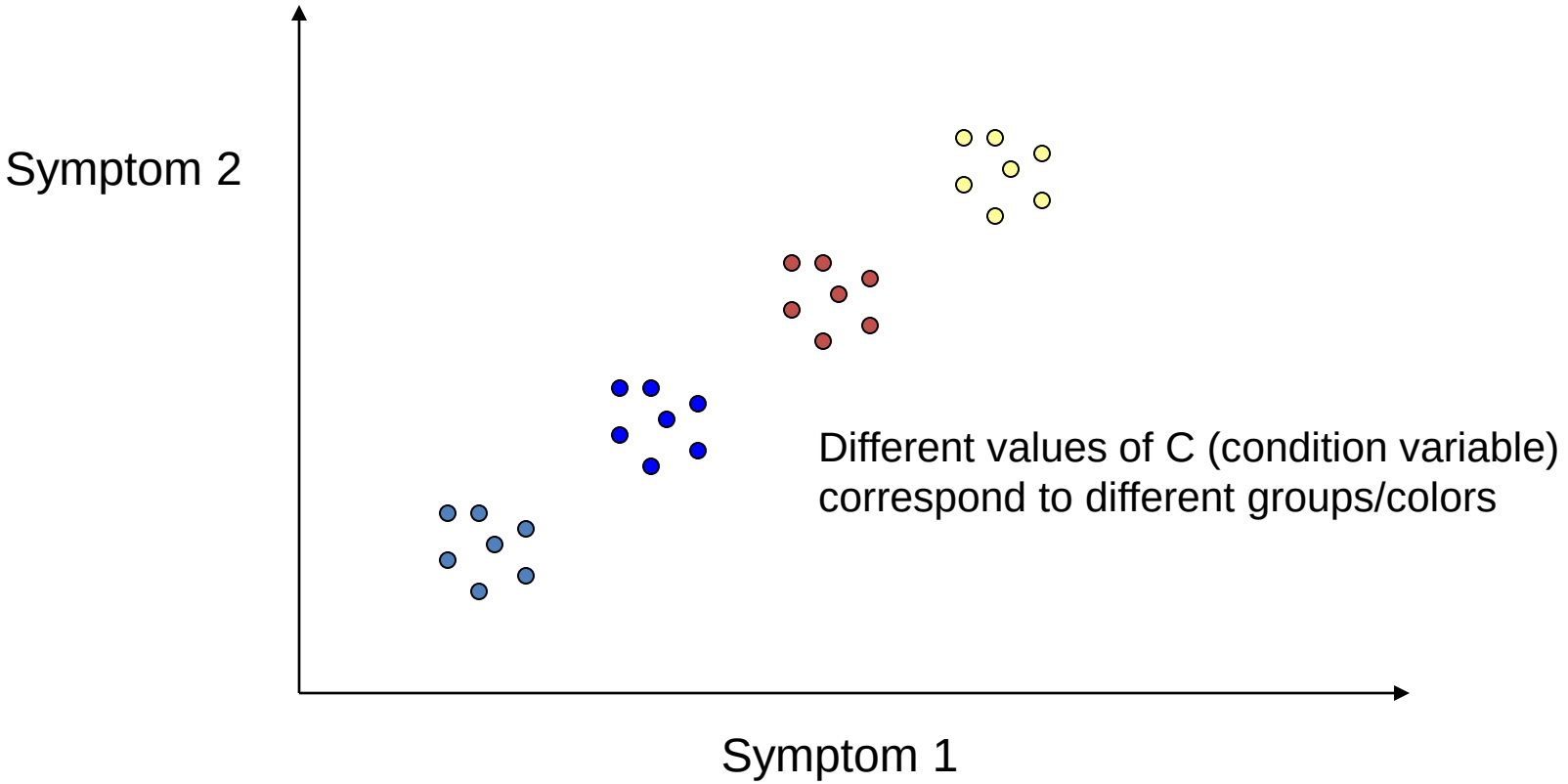
- Conditional independence does not imply independence

- Example:
  - A = height
  - B = reading ability
  - C = age

  - P(reading ability | age, height) = P(reading ability | age)
  - P(height | reading ability, age) = P(height | age)

- Note:
  - Height and reading ability are dependent (not independent) but are conditionally independent given age

Another Example



In each group, symptom 1 and symptom 2 are conditionally independent.

But clearly, symptom 1 and 2 are marginally dependent (unconditionally).

"…probability theory is more fundamentally concerned with the structure of reasoning and causation than with numbers."

Glenn Shafer and Judea Pearl
***Introduction to Readings in Uncertain Reasoning*,**
Morgan Kaufmann, 1990

Conclusions…

- Representing uncertainty is useful in knowledge bases
  - Probability provides a coherent framework for uncertainty
- Full joint distributions are intractable to work with
- Conditional independence assumptions allow much simpler models of real-world phenomena
- Bayesian networks are a systematic way to construct parsimonious structured distributions

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