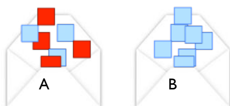
Probabilistic Reasoning

AI Class 9 (Ch. 13)



Based on slides by Dr. Marie desJardin. Some material also adapted from slides by Dr. Matuszek @ Villanova University, which are based in part on www.csc.calpoly.edu/~fkurfess/Courses/CSC-481/W02/Slides/Uncertainty.ppt and www.cs.umbc.edu/courses/graduate/671/fall05/slides/c18_prob.ppt

Bookkeeping

- HW 2 Due 10/3, 11:59pm
  - Blackboard assignment open Friday
- **Important**: understand the math in Chapter 13 **thoroughly**
  - Underpins future work
  - Also basically all of modern AI
- Grading
  - A = 92-100, A- = 90-92, B is 82-87, B- is 80-82, B+ is 88-89, etc
  - These may be revised downward.

Today's Class

- Probability theory
- Bayesian inference
  - From the joint distribution
  - Using independence/factoring
  - From sources of evidence

Probabilistic inference: finding *posterior probability* for a proposition, given observed evidence.

– R&N 490

Bayesian Reasoning

- Posteriors and priors
- What is inference?
- What is uncertainty?
- When/why use probabilistic reasoning?
- What is induction?
- What is the probability of two independent events?
- Frequentist/objectivist/subjectivist assumptions

Sources of Uncertainty

- Uncertain **inputs**
  - Missing data
  - Noisy data
- Uncertain **knowledge**
  - >1 cause → >1 effect
  - Incomplete knowledge of conditions or effects
  - Incomplete knowledge of causality
  - Probabilistic effects

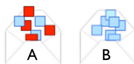
- Uncertain **outputs**
  - Default reasoning (even deduction) is uncertain
  - Abduction & induction inherently uncertain
  - Incomplete deductive inference can be uncertain

Probabilistic reasoning only gives **probabilistic results** (summarizes uncertainty from various sources)

Decision Making with Uncertainty

- **Rational** behavior: for each possible action,
  - Identify possible outcomes
  - Compute **probability** of each outcome
  - Compute **utility** of each outcome
  - Compute probability-weighted **(expected) utility** of possible outcomes for each action
  - Select the action with the highest expected utility (principle of **Maximum Expected Utility**)

Also the definition of "rational" for deterministic decision-making.



Probability

- **World:** The complete set of states
- **Event:** Something that happens
- **Sample Space:** All the things (outcomes) that *could* happen in some set of circumstances
  - Pull 2 squares from envelope A: what is the sample space?
  - How about envelope B?
- **Probability $P(x)$:** likelihood of event x occurring
  - Pull a few more squares.
  - How many of each did you get from A? From B?

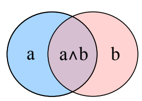
Basic Probability

- Each P is a non-negative value in [0,1]
- Total probability of the sample space is 1
- For **mutually exclusive events**, the probability for at least one of them is the *sum* of their individual probabilities
- Experimental probability
  - Based on frequency of past events
- Subjective probability
  - Based on expert assessment

Why Probabilities Anyway?

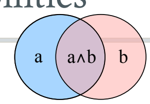
3 simple axioms → all rules of probability theory*

1. All probabilities are between 0 and 1.
   - $0 \leq P(a) \leq 1$
2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0.
   - $P(true) = 1$
   - $P(false) = 0$
3. The probability of a disjunction is:
   - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



*Kolmogorov – https://en.wikipedia.org/wiki/Andrey_Kolmogorov
De Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

Compound Probabilities



- Describe **independent** events
  - Do not affect each other in any way
- **Joint** probability of two independent events A and B
  $P(A \cap B) = P(A) * P(B)$ ← What do these say?
- **Union** probability of two independent events A and B
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  $= P(A) + P(B) - (P(A) * P(B))$

Pull two squares from envelope A. What is the probability that they are BOTH red?

Probability Theory

- **Random variables:**
  - Domain: possible values
- **Atomic event:**
  - Complete specification of a state
- **Prior probability:**
  - Degree of belief without any new evidence
- **Joint probability:**
  - Matrix of **combined probabilities of a set of variables**

- Alarm (A), Burglary (B), Earthquake (E)
  - Boolean, discrete, continuous
- A=*true* ∧ B=*true* ∧ E=*false* :
  - alarm ∧ burglary ∧ ¬earthquake
- $P(B) = 0.1$
- $P(A, B) =$

	alarm	**¬alarm**
burglary	0.09	0.01
¬burglary	0.1	0.8

Probability Theory: Definitions

- **Conditional probability**
  - Probability of effect given cause(s)
- Computing conditional probability:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
- $P(b)$: **normalizing** constant

- **Product rule**:
  - $P(a \wedge b) = P(a \mid b)\, P(b)$
- **Marginalizing**:
  - Finding distribution over a *subset* of variables
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$ (**conditioning**)

Try It…

- **Cond'l probability**
  - P(effect, cause[s])
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **normalizing** constant
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b)\, P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$ (**conditioning**)

- $P(A \mid B) = ?$
- $P(B \mid A) = ?$
- $P(B \wedge A) = ?$
- $P(A) = ?$

	alarm	¬ alarm
burglary	0.09	0.01
¬ burglary	0.1	0.8

13

Probability Theory (cont.)

- **Cond'l probability**
  - P(effect, cause[s])
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - Here, $P(b)$: **normalizing** constant (α)
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b)\, P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$ (**conditioning**)

- $P(A \mid B) = 0.9$
  $P(B \mid A) = 0.47$
  - $P(B \mid A) = P(B \wedge A) / P(A) = 0.09 / 0.19 = 0.47$
- $P(B \wedge A) = P(B \mid A)\, P(A) = 0.47 \times 0.19 = 0.09$
- $P(A) = P(A \wedge B) + P(A \wedge \neg B) = 0.09 + 0.1 = 0.19$

14

Example: Inference from the Joint

- $P(B \mid A) = \alpha\, P(B, A)$
  $= \alpha\, [P(B, A, E) + P(B, A, \neg E)$
  $= \alpha\, [(.01, .01) + (.08, .09)]$
  $= \alpha\, [(.09, .1)]$

	A		¬A	
	E	¬E	E	¬E
B	0.01	0.08	0.001	0.009
¬B	0.01	0.09	0.01	0.79

- Since
  $P(B \mid A) + P(\neg B \mid A) = 1$, $\alpha = 1 / (0.09 + 0.1) = 5.26$
  (i.e., $P(A) = 1/\alpha = 0.19$)
- $P(B \mid A) = 0.09 * 5.26 = 0.474$
- $P(\neg B \mid A) = 0.1 * 5.26 = 0.526$

quizlet: how can you verify this?

15

Exercise: Inference from the Joint

- Queries: what is…
  - The prior probability of *smart*?
  - The prior probability of *study*?
  - The conditional probability of *prepared*, gi[ven s]*tudy* and *smart*?
- Save these answers for later! ☺

Where do these come from?

P(smart ∧ study ∧ prep)	***smart***		***¬smart***	
	study	***¬study***	***study***	***¬study***
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

16

Independence: ⫫

- **Independent:** Two sets of propositions that do not affect each others' probabilities
- Easy to calculate **joint** and **conditional** probability of independence:
  $(A, B) \Leftrightarrow P(A \wedge B) = P(A)\, P(B)$ or $P(A \mid B) = P(A)$
- Examples:
  - A = alarm
  - B = burglary
  - E = earthquake
  - M = moon phase
  - L = light level

$A ⫫ B ⫫ E = f$
$M ⫫ L = f$
$A ⫫ M = t$

17

Independence Example

- ***{moon-phase, light-level} ⫫ {burglary, alarm, earthquake}***
  - But maybe burglaries increase in low light
  - But, *if* we know the light level, *moon-phase* ⫫ *burglary*
  - *Once* we're burglarized, light level doesn't affect whether the alarm goes off; {*light-level*} ⫫ {*alarm*}
- We need:
  1. A more complex notion of independence
  2. Methods for reasoning about these kinds of (common) relationships

18

Exercise: Independence

P(smart ∧ study ∧ prep)	*smart*		*¬smart*	
	study	*¬study*	*study*	*¬study*
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

- Queries:
  - Is *smart* independent of
  - Is *prepared* independent

Smart	Study		
t	*t*	0.432 + 0.48	0.480
t	*f*	0.16 + 0.16	0.32
f	*t*	0.084 + 0.008	0.092
f	*f*	0.036 + 0.72	0.756

Conditional Probabilities

- Describes **dependent** events
  - Affect each other in some way
- Typical in the real world
- If we know some event has occurred, what does that tell us about the likelihood of another event?

Conditional Independence

- *moon-phase* and *burglary* are **conditionally independent given** *light-level*
  - That is, $M \perp\!\!\!\perp B$ if we already know L
- Conditional independence is:
  - Weaker than absolute independence
  - Useful in decomposing full joint probability distributions

Conditional Independence

- **Absolute** independence: $A \perp\!\!\!\perp B$, if:
  - $P(A \wedge B) = P(A)\,P(B)$
  - Equivalently, $P(A) = P(A \mid B)$ and $P(B) = P(B \mid A)$
- A and B are **conditionally independent** given C if:
  - $P(A \wedge B \mid C) = P(A \mid C)\,P(B \mid C)$
- This lets us decompose the joint distribution:
  - $P(A \wedge B \wedge C) = P(A \mid C)\,P(B \mid C)\,P(C)$
- What does this mean?

Exercise: Conditional Independence

P(smart ∧ study ∧ prep)	*smart*		*¬smart*	
	study	*¬study*	*study*	*¬study*
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

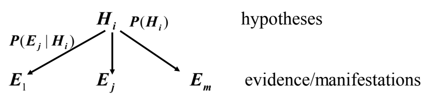
- Queries:
  - Is *smart* conditionally independent of *prepared*, given *study*?
  - Is *study* conditionally independent of *prepared*, given *smart*?

Bayes' Rule

- Derive the probability of an **event** given **another event**
  - Assumption of attribute independency (Naïve assumption): Naïve Bayes assumes that all of the *attributes* are independent.
- Bayes' rule is derived from the product rule: R&N 495
  - $P(Y \mid X) = P(X \mid Y)\,P(Y) / P(X)$
- Often useful for **diagnosis**. If we have:
  - X = (observed) effects, Y = (hidden) causes
  - A model for how causes lead to effects: $P(X \mid Y)$
  - Prior beliefs about frequency of occurrence of effects: $P(Y)$
- We can reason abductively from effects to causes:
  - $P(Y \mid X)$

Bayesian Inference

- In the setting of diagnostic/evidential reasoning



  - Know: prior probability of hypothesis $P(H_i)$
    conditional probability $P(E_j | H_i)$
  - Want to compute the *posterior probability* $P(H_i | E_j)$
- Bayes' theorem (formula 1):

$$P(H_i | E_j) = P(H_i)P(E_j | H_i) / P(E_j)$$

For our Envelopes

- Envelope A has 10 blue and 10 red
- Envelope B as 7 blue and 13 red
- So if we pull a red square it is slightly more likely to be from Envelope B
- A blue square is slightly more likely to be from Envelope A

Simple Bayesian Diagnostic Reasoning

- Knowledge base:
  - Evidence / manifestations: $E_1, \dots E_m$
  - Hypotheses / disorders: $H_1, \dots H_n$
    - E_j and H_i are **binary**; hypotheses are **mutually exclusive** (non-overlapping) and **exhaustive** (cover all possible cases)
  - Conditional probabilities: $P(E_j \mid H_i)$, $i = 1, \dots n$; $j = 1, \dots m$
- Cases (evidence for a particular instance): $E_1, \dots, E_m$
- Goal: Find the hypothesis H_i with the highest posterior
  - $\text{Max}_i\ P(H_i \mid E_1, \dots, E_m)$

Priors

- Four values total here:
  - P(H|E) = (P(E|H) * P(H)) / P(E)
- P(H|E) — what we want to compute
- Three we already know, called the *priors*
  - P(E|H)
  - P(H)
  - P(E)

(In ML we use the training set to estimate the priors)

Bayesian Diagnostic Reasoning II

- Bayes' rule says that
  - $P(H_i \mid E_1, \dots, E_m) = P(E_1, \dots, E_m \mid H_i)\, P(H_i) / P(E_1, \dots, E_m)$
- Assume each piece of evidence E_i is conditionally independent of the others, **given** a hypothesis H_i, then:
  - $P(E_1, \dots, E_m \mid H_i) = \prod_{j=1}^{l} P(E_j \mid H_i)$
- If we only care about relative probabilities for the H_i, then we have:
  - $P(H_i \mid E_1, \dots, E_m) = \alpha\, P(H_i) \prod_{j=1}^{l} P(E_j \mid H_i)$

Bayes Example: Diagnosing Meningitis

$$P(H_i | E_j) = P(H_i)P(E_j | H_i) / P(E_j)$$

- Suppose we know that
  - Stiff neck is a symptom in 50% of meningitis cases
  - Meningitis (m) occurs in 1/50,000 patients
  - Stiff neck (s) occurs in 1/20 patients
- Then
  - P(s|m) = 0.5, P(m) = 1/50000, P(s) = 1/20
  - P(m|s) = (P(s|m) P(m))/P(s)
    = (0.5 x 1/50000) / 1/20 = .0002
- So we expect that one in 5000 patients with a stiff neck to have meningitis.

Analysis of Naïve Bayes Algorithm

- Advantages:
  - Sound theoretical basis
  - Works well on numeric and textual data
  - Easy implementation and computation
  - Has been effective in practice (e.g., typical spam filter)

32

Limitations of Simple Bayesian Inference

- Cannot easily handle multi-fault situations, nor cases where intermediate (hidden) causes exist:
  - Disease D causes syndrome S, which causes correlated manifestations M_1 and M_2
- Consider a composite hypothesis $H_1 \wedge H_2$, where H_1 and H_2 are independent. What is the relative posterior?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) = \alpha\ P(E_1, \ldots, E_l \mid H_1 \wedge H_2)\ P(H_1 \wedge H_2)$
    $= \alpha\ P(E_1, \ldots, E_l \mid H_1 \wedge H_2)\ P(H_1)\ P(H_2)$
    $= \alpha \prod^{l}_{j=1} P(E_j \mid H_1 \wedge H_2)\ P(H_1)\ P(H_2)$
- How do we compute $P(E_j \mid H_1 \wedge H_2)$??

33

Limitations of Simple Bayesian Inference II

- Assume H1 and H2 are independent, given E1, …, El?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) = P(H_1 \mid E_1, \ldots, E_l)\ P(H_2 \mid E_1, \ldots, E_l)$
- This is a very unreasonable assumption
  - Earthquake and Burglar are independent, but *not* given Alarm:
    - P(burglar | alarm, earthquake) << P(burglar | alarm)
- Simple application of Bayes' rule doesn't handle causal chaining:
  - A: this year's weather; B: cotton production; C: next year' s cotton price
  - A influences C indirectly: A→ B → C
  - P(C | B, A) = P(C | B)
- Need a richer representation to model interacting hypotheses, conditional independence, and causal chaining
- Next time: conditional independence and Bayesian networks!

34