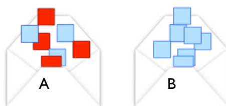
Probabilistic Reasoning

AI Class 9 (Ch. 13)



Based on slides by Dr. Marie desJardin and Dr. Tim Oates. 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

Today's Class

- Probability theory
- Probability notation
- 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

Today's Class

We don't (can't!) know everything about most problems.

- Most problems are not:
  - Deterministic
  - Fully observable
- Or, we can't calculate everything.
  - Continuous problem spaces

Probability lets us understand, quantify, and work with this uncertainty.

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

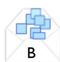
- Uncertain **outputs**
  - Default reasoning (even deduction) is uncertain
  - Abduction & induction inherently uncertain
  - Incomplete deductive inference can be uncertain

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
    - "goodness" or "desirability" per some formally specified definition
  - Compute probability-weighted **(expected) utility** of possible outcomes for each action
  - Select the action with the highest expected utility (principle of **Maximum Expected Utility**)

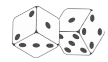
Also the definition of "rational" for deterministic decision-making!



Probability

- **World:** The complete set of possible states
- **Random variables:** Problem aspects that take a value
  - "The number of blue squares we are holding," *B*
  - "The combined value of two dice we rolled," *C*
- **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?
- **World, redux:** A *complete assignment of values to variables*

Basic Probability



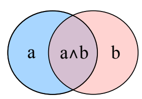
- Each P is a non-negative value in [0,1]
  - $P(\{1,1\}) = 1/36$
- Total probability of the sample space is 1
  - $P(\{1,1\}) + P(\{1,2\}) + P(\{1,3\}) + \ldots + P(\{6,6\}) = 1$
- For mutually exclusive events, the probability for at least one of them is the sum of their individual probabilities
  - $P(\text{sunny}) \lor P(\text{cloudy}) = P(\text{sunny}) + P(\text{cloudy})$
- Experimental probability: Based on frequency of past events
- Subjective probability: Based on expert assessment

commons.wikimedia.org/wiki/File:2-Dice-Icon.svg

Why Probabilities Anyway?

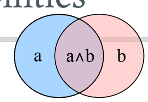
3 simple axioms → all rules of probability theory*

1. All probabilities are between 0 and 1.
   - $0 \le P(a) \le 1$
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:
   - $P(a \lor b) = P(a) + P(b) - P(a \land b)$



*Kolmogorov – 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,** $P(A|B)$

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

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

Probability Distributions

- A distribution is the probabilities of **all possible values** of a random variable
- Ex: weather can be sunny, rainy, cloudy, or snowy
  - $P(\text{Weather} = \text{sun}) = 0.6$
  - $P(\text{Weather} = \text{rain}) = 0.1$
  - $P(\text{Weather} = \text{cloud}) = 0.29$
  - $P(\text{Weather} = \text{snow}) = 0.01$
  - $\mathbf{P}(\text{Weather}) = \langle 0.6, 0.1, 0.29, 0.01 \rangle$ ← shortcut
- $\mathbf{P}(\text{Weather})$: **probability distribution on Weather**

Probability Theory: Definitions

- **Conditional probability:** Probability of some effect given that we know cause(s)
  - Example: $P(alarm|burglary)$
    - (Technically, we only know b is true, not causal, but…)
- Computing it:
  - $P(a \mid b) = \dfrac{P(a \wedge b)}{P(b)}$
- $P(b)$: **normalizing constant**
  - (Later we'll call this alpha)

14

Probability Theory: Definitions

- **Product rule**:
  - $P(a \wedge b) = P(a \mid b)\, P(b)$
- **Marginalizing** (summing out):
  - Finding distribution over *one* or a *subset* of variables
  - Marginal probability of B summed over all alarm states:
  - $P(B) = \Sigma_a P(B,\, a)$
- **Conditioning** over a subset of variables:
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$

15

Try It…

	alarm	¬alarm
burglary	0.09	0.01
¬burglary	0.1	0.8

- **Cond'l probability**
  - P(effect, cause[s])
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **normalizing** constant ($1/\alpha$)
- **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**)

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

16

Example: Inference from the Joint

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

- $P(B \mid A) = \alpha\, P(B, A)$
  $= \alpha\, [P(B, A, E) + P(B, A, \neg E)$
  $= \alpha\, [(.01,\ .01) + (.08,\ .09)]$
  $= \alpha\, [(.09,\ .1)]$
- Since
  $P(B \mid A) + P(\neg B \mid A) = 1,\ \alpha = 1 / (0.09 + 0.1) = 5.26$
  (i.e., $P(A) = 1/\alpha = 0.19$)
- $P(B \mid A) = 0.09 * 5.26 = 0.474$
- $P(\neg B \mid A) = 0.1 * 5.26 = 0.526$

17

Exercise: Inference from the Joint

- Queries: what is…
  - The prior probability (knowing noth
  - The prior probability of *study*?
  - The conditional probability of *prepare* ven *study* and *smart*?

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

18

Exercise: Inference from the joint

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

Queries:

- **What is the prior probability of *smart*?**
- What is the prior probability of *study*?
- What is the conditional probability of *prepared*, given *study* and *smart*?

P(*smart*) = .432 + .16 + .048 + .16 = **0.8**

19

Exercise: Inference from the joint

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

Queries:

- What is the prior probability of *smart*?
- **What is the prior probability of *study*?**
- What is the conditional probability of *prepared*, given *study* and *smart*?

P(*study*) = .432 + .048 + .084 + .036 = **0.6**

Exercise: Inference from the joint

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

Queries:

- **What is the conditional probability of *prepared*, given *study* and *smart*?**

P(*prep*|*smart*,*study*) = P(*prep*, *smart*, *study*)/P(*smart*, *study*)
= .432 / (.432 + .048)
= **0.9**

Independence: $\perp\!\!\!\perp$

- **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

$A \perp\!\!\!\perp B \perp\!\!\!\perp E = f$
$M \perp\!\!\!\perp L = f$
$A \perp\!\!\!\perp M = t$

Independence Example

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

Exercise: Independence

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

- Is *smart* independent of *study*?
  - P(*smart*|*study*) = P(*smart*)
- Is *prepared* independent of *study*?
  - P(*prep*|*study*) = P(*prep*)

Exercise: Independence

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

- Is *smart* independent of
  - P(*smart*|*study*) = P(*smart*)
- Is *prepared* independent
  - P(*prep*|*study*) = P(*prep*)

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

Exercise: Independence

P(smart ∧ study ∧ prep)	stu			
prepared	.4			
¬prepared	.048	.16	.036	.072

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

- P(*smart* | *study*) = P(*smart*)
- P(*smart* | *study*) = P(*smart*, *study*) / P(*study*)
- 0.8 = (.432 + .048) / .6
- 0.8 = 0.8 ✔

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 ⫫ 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 some **event,** *given* **another event**
  - Assumption of attribute independency (AKA the Naïve assumption)
  - Naïve Bayes assumes that all *attributes* are independent.
- Bayes' rule is derived from the product rule:
  - $P(Y \mid X) = P(X \mid Y)\,P(Y) / P(X)$ R&N 495

Bayes' Rule

- Bayes' rule is derived from the product rule:
  - $P(Y \mid X) = P(X \mid Y)\, P(Y) \,/\, P(X)$
- 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:
  - $P(Y \mid X)$

32

Naïve Bayes Algorithm

- Estimate the probability of each class:
  - Compute the posterior probability (Bayes rule)

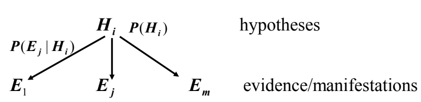
$$\boxed{P(c_i \mid D) = \frac{P(c_i)P(D \mid c_i)}{P(D)}}$$

  - Choose the class with the highest probability
- Assumption of attribute independency (Naïve assumption): Naïve Bayes assumes that all of the attributes are independent.

33

Bayesian Inference

- In the setting of diagnostic/evidential reasoning



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

$$P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$$

34

Simple Bayesian Diagnostic Reasoning

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

36

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)

37

Bayesian Diagnostic Reasoning II

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

38

Bayes Example: Diagnosing Meningitis

$\boxed{P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)}$

- Your patient comes in with a stiff neck.
- Is it meningitis?
- 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
- So probably not. But specifically?

39

Bayes Example: Diagnosing Meningitis

$\boxed{P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)}$

  - 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
  - $P(s|m) = 0.5$, $P(m) = 1/50000$, $P(s) = 1/20$
  - $P(m|s) = (P(s|m)\ P(m))/P(s)$
    $= (0.5 \times 1/50000) / 1/20 = .0002$
- So we expect that one in 5000 patients with a stiff neck to have meningitis.

40

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)

41

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?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_m) = \alpha\ P(E_1, \ldots, E_m \mid H_1 \wedge H_2)\ P(H_1 \wedge H_2)$
    $= \alpha\ P(E_1, \ldots, E_m \mid H_1 \wedge H_2)\ P(H_1)\ P(H_2)$
    $= \alpha \prod^{l}_{m=1} P(E_m \mid H_1 \wedge H_2)\ P(H_1)\ P(H_2)$
- How do we compute $P(E_j \mid H_1 \wedge H_2)$??

42

Limitations of Simple Bayesian Inference II

- Assume H1 and H2 are independent, given E1, …, Ej?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_j) = P(H_1 \mid E_1, \ldots, E_j)\ P(H_2 \mid E_1, \ldots, E_j)$
- 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: $A \rightarrow B \rightarrow C$
  - $P(C \mid B, A) = P(C \mid B)$
- Need a richer representation to model interacting hypotheses, conditional independence, and causal chaining
- Next time: conditional independence and Bayesian networks!

43