Bayesian Reasoning

Chapter 13

Thomas Bayes, 1701-1761

Today's topics

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

Sources of Uncertainty

- Uncertain **inputs --** missing and/or noisy data
- Uncertain **knowledge**
  - Multiple causes lead to multiple effects
  - Incomplete enumeration of conditions or effects
  - Incomplete knowledge of causality in the domain
  - Probabilistic/stochastic effects
- Uncertain **outputs**
  - Abduction and induction are inherently uncertain
  - Default reasoning, even deductive, is uncertain
  - Incomplete deductive inference may be uncertain

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

Decision making with uncertainty

Rational behavior:

- For each possible action, identify the possible outcomes
- Compute the **probability** of each outcome
- Compute the **utility** of each outcome
- Compute the probability-weighted **(expected) utility** over possible outcomes for each action
- Select action with the highest expected utility (principle of **Maximum Expected Utility**)

Why probabilities anyway?

Kolmogorov showed that three simple axioms lead to the rules of probability theory

1. All probabilities are between 0 and 1:

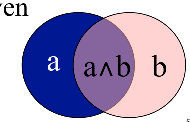
   $0 \leq P(a) \leq 1$

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 given by:

   $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



Probability theory 101

- **Random variables**
  - Domain
- **Atomic event**: complete specification of state
- **Prior probability**: degree of belief without any other evidence
- **Joint probability**: matrix of combined probabilities of a set of variables

- Alarm, Burglary, Earthquake
  - Boolean (like these), discrete, continuous
- Alarm=T∧Burglary=T∧Earthquake=F
  alarm ∧ burglary ∧ ¬earthquake
- P(Burglary) = 0.1
  P(Alarm) = 0.1
  P(earthquake) = 0.000003
- P(Alarm, Burglary) =

	alarm	**¬alarm**
burglary	.09	.01
¬burglary	.1	.8

Probability theory 101

	alarm	**¬alarm**
burglary	.09	.01
¬burglary	.1	.8

- **Conditional probability**: prob. of effect given causes
- **Computing conditional probs**:
  - P(a | b) = P(a ∧ b) / P(b)
  - P(b): **normalizing** constant
- **Product rule**:
  - P(a ∧ b) = P(a | b) * P(b)
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a) P(a)$ (**conditioning**)

- P(burglary | alarm) = .47
  P(alarm | burglary) = .9
- P(burglary | alarm) =
  P(burglary ∧ alarm) / P(alarm)
  = .09/.19 = .47
- P(burglary ∧ alarm) =
  P(burglary | alarm) * P(alarm)
  = .47 * .19 = .09
- P(alarm) =
  P(alarm ∧ burglary) +
  P(alarm ∧ ¬burglary)
  = .09+.1 = .19

Example: Inference from the joint

	alarm		**¬alarm**	
	earthquake	**¬earthquake**	**earthquake**	**¬earthquake**
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

P(burglary | alarm) = α P(burglary, alarm)
= α [P(burglary, alarm, earthquake) + P(burglary, alarm, ¬earthquake)
= α [(.01, .01) + (.08, .09)]
= α [(.09, .1)]

Since P(burglary | alarm) + P(¬burglary | alarm) = 1, α = 1/(.09+.1) = 5.26
(i.e., P(alarm) = 1/α = .19 – **quizlet**: how can you verify this?)

P(burglary | alarm) = .09 * 5.26 = .474

P(¬burglary | alarm) = .1 * 5.26 = .526

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*?**
- **Save these answers for next time! ☺**

9

Independence

- When sets of variables don't affect each others' probabilities, we call them **independent**, and can easily compute their joint and conditional probability:
  Independent(A, B) → P(A∧B) = P(A) * P(B), P(A | B) = P(A)
- {moonPhase, lightLevel} *might* be independent of {burglary, alarm, earthquake}
  - Maybe not: crooks may be more likely to burglarize houses during a new moon (and hence little light)
  - But if we know the light level, the moon phase doesn't affect whether we are burglarized
  - If burglarized, light level doesn't affect if alarm goes off
- Need a more complex notion of independence and methods for reasoning about the relationships

10

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 *study*?**
- **Is *prepared* independent of *study*?**

11

Conditional independence

- Absolute independence:
  - A and B are **independent** if P(A ∧ B) = P(A) * P(B); equivalently, P(A) = P(A | B) and P(B) = P(B | A)
- A and B are **conditionally independent** given C if
  - P(A ∧ B | C) = P(A | C) * P(B | C)
- This lets us decompose the joint distribution:
  - P(A ∧ B ∧ C) = P(A | C) * P(B | C) * P(C)
- Moon-Phase and Burglary are ***conditionally independent given*** Light-Level
- Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint probability distribution

12

Exercise: Conditional independence

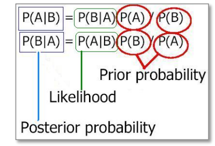
p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

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

13

Bayes' rule



- Derived from the product rule:
  - P(C | E) = P(E | C) * P(C) / P(E)
- Often useful for diagnosis:
  - If E are (observed) effects and C are (hidden) causes,
  - We may have a model for how causes lead to effects (P(E | C))
  - We may also have prior beliefs (based on experience) about the frequency of occurrence of effects (P(C))
  - Which allows us to reason abductively from effects to causes (P(C | E))

14

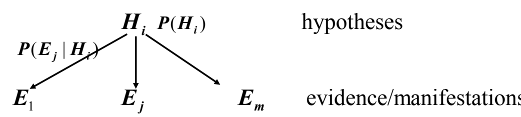
Ex: meningitis and stiff neck

- Meningitis (M) can cause a a stiff neck (S), though here are many other causes for S, too
- We'd like to use S as a diagnostic symptom and estimate p(M|S)
- Studies can easily estimate p(M), p(S) and p(S|M)
  p(M)=0.7, p(S)=0.01, p(M)=0.00002
- Applying Bayes' Rule:
  p(M|S) = p(S|M) * p(M) / p(S) = 0.0014
- We can also do this w/o p(S) if we know p(S|~M)
  α <p(S|M)*P(m), p(S|~M)*p(~M)>

15

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' s theorem (formula 1):

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

16

Simple Bayesian diagnostic reasoning

- Also known as: Naive Bayes classifier
- Knowledge base:
  - Evidence / manifestations: $E_1, \ldots E_m$
  - Hypotheses / disorders: $H_1, \ldots H_n$

    Note: 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): $E_1, \ldots, E_l$
- Goal: Find the hypothesis H_i with the highest posterior
  - $\mathrm{Max}_i\ P(H_i \mid E_1, \ldots, E_l)$

Simple Bayesian diagnostic reasoning

- 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 evidence E_i is conditionally independent of the others, *given* a hypothesis H_i, then:

  $P(E_1 \ldots E_m \mid H_i) = \prod^m_{j=1} P(E_j \mid H_i)$
- If we only care about relative probabilities for the H_i, then we have:

  $P(H_i \mid E_1 \ldots E_m) = \alpha\ P(H_i) \prod^m_{j=1} P(E_j \mid H_i)$

Limitations

- 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's 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 $P(E_j \mid H_1 \wedge H_2)$?

Limitations

- 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)
- Another limitation is that simple application of Bayes's rule doesn't allow us to 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: conditional independence and Bayesian networks!

Summary

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Can answer queries by summing over atomic events
- But we must find a way to reduce the joint size for non-trivial domains
- Bayes' rule lets unknown probabilities be computed from known conditional probabilities, usually in the causal direction
- Independence and conditional independence provide the tools

21

Postscript: Frequentists vs. Bayesians

- Frequentist inference draws conclusions from sample data based on the frequency or proportion of the data
- Bayesian inference uses Bayes' rule to update probability estimates for a hypothesis as additional evidence is learned
- The differences are often subtle, but can be consequential

22

Frequentists vs. Bayesians
http://xkcd.com/1132/