Representing Uncertainty for Probabilistic Inference

Chapter 13

Uncertainty in the World

- An agent can often be uncertain about the state of the domain since there is often ambiguity and uncertainty
- Plausible/**probabilistic inference**
  - I've got this evidence; what's the chance that this conclusion is true?
    - I've got a sore neck; how likely am I to have meningitis?
    - A mammogram test is positive; what's the probability that the patient has breast cancer?

Probabilistic Inference

How do we use probabilities in AI?

- You wake up with a headache
- Do you have the flu?
- *H* = headache, *F* = flu

Logical Inference: if *H* then *F*
(but the world is usually not this simple)

Statistical Inference: compute the probability of a query/diagnosis/decision given (i.e., conditioned on) evidence/symptom/observation, i.e., $P(F \mid H)$

[Example from Andrew Moore]

Uncertainty

- Say we have a rule:

  if *toothache* ***then*** *problem is cavity*
- But not all patients have toothaches due to cavities, so we could set up rules like:

  if *toothache and ¬gum-disease and ¬filling and …*
  then *problem = cavity*
- This gets complicated; better method:

  if *toothache* ***then*** *problem is cavity with 0.8 probability*

  or $P(cavity \mid toothache) = 0.8$

  the probability of cavity is 0.8 given toothache is observed

Uncertainty in the World and our Models

- True uncertainty: rules *are* probabilistic in nature
  - quantum mechanics
  - rolling dice, flipping a coin
- Laziness: too hard to determine exception-less rules
  - takes too much work to determine *all* of the relevant factors
  - too hard to use the enormous rules that result
- Theoretical ignorance: don't know all the rules
  - problem domain has no complete, consistent theory (e.g., medical diagnosis)
- Practical ignorance: do know all the rules BUT
  - haven't collected all relevant information for a particular case

Logics

Logics are characterized by what they use as "primitives"

Logic	**What Exists in World**	**Knowledge States**
Propositional	facts	true/false/unknown
First-Order	facts, objects, relations	true/false/unknown
Temporal	facts, objects, relations, times	true/false/unknown
Probability Theory	facts	degree of belief 0..1
Fuzzy	degree of truth	degree of belief 0..1

Probability Theory

- Probability theory serves as a formal means for
  - Representing and reasoning with uncertain knowledge
  - Modeling **degrees of belief** in a proposition (event, conclusion, diagnosis, etc.)
- *Probability is the "language" of uncertainty*
  - A key modeling tool in modern AI

Sample Space

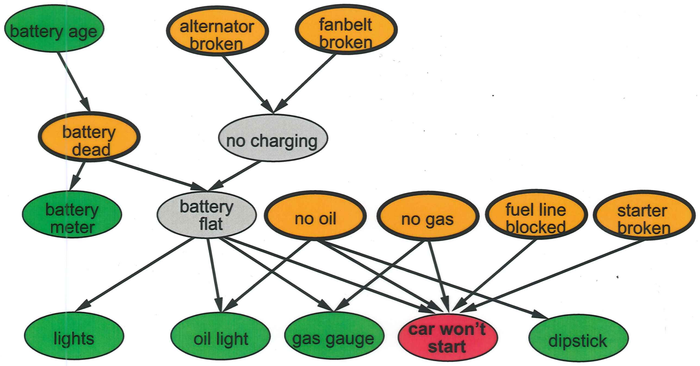
- A space of **events** in which we assign probabilities
- Events can be binary, multi-valued, or continuous
- Events are **mutually exclusive**
- Examples
  - Coin flip: {head, tail}
  - Die roll: {1, 2, 3, 4, 5, 6}
  - English words: a dictionary
  - High temperature tomorrow: {-100, …, 100}

Random Variable

- A variable, *X*, whose domain is a sample space, and whose value is (somewhat) uncertain
- Examples:
  - *X* = coin flip outcome
  - *X* = tomorrow's high temperature
- For a given task, the user defines a set of random variables for describing the world
- **Each variable has a set of mutually exclusive and exhaustive possible values**

Example: Car diagnosis

Initial evidence: car won't start
Testable variables (green), "broken, so fix it" variables (orange)
Hidden variables (gray) ensure sparse structure, reduce parameters

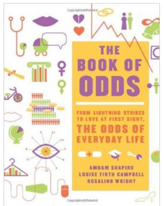


Probability for Discrete Events

- An agent's uncertainty is represented by $P(A{=}a)$ or simply $P(a)$
  - the agent's degree of belief that variable A takes on value a *given no other information* related to A
  - a single probability called an **unconditional** or **prior probability**

Probability for Discrete Events

- Examples
  - *P*(head) = *P*(tail) = 0.5   fair coin
  - *P*(head) = 0.51, *P*(tail) = 0.49   slightly biased coin
  - *P*(first word = "the" when flipping to a random page in R&N) = ?
- Book: *The Book of Odds*



Source of Probabilities

- Frequentists
  - probabilities come from data
  - if 10 of 100 people tested have a cavity, *P*(*cavity*) = 0.1
  - probability means the fraction that would be observed in the limit of infinitely many samples
- Objectivists
  - probabilities are real aspects of the world
  - objects have a propensity to behave in certain ways
  - coin has propensity to come up heads with probability 0.5
- Subjectivists
  - probabilities characterize an agent's belief
  - have no external physical significance

Probability Distributions

Given A is a RV taking values in $\langle a_1, a_2, \ldots, a_k \rangle$

e.g., if A is Sky, then value is one of $<clear, partly\_cloudy, overcast>$

- $P(a)$ represents a single probability where $A=a$

  e.g., if A is Sky, then $P(a)$ means any one of $P(clear)$, $P(partly\_cloudy)$, $P(overcast)$

- $\mathbf{P}(A)$ represents a probability distribution
  - the **set of values**: $\langle P(a_1), P(a_2), \ldots, P(a_k) \rangle$
  - If A takes n values, then $\mathbf{P}(A)$ is a set of n probabilities

    e.g., if A is Sky, then $\mathbf{P}(Sky)$ is the set of probabilities: $\langle P(clear), P(partly\_cloudy), P(overcast) \rangle$
  - Property: $\sum P(a_i) = P(a_1) + P(a_2) + \ldots + P(a_k) = 1$
    - sum over all values in the domain of variable A is 1 because the ***domain is mutually exclusive and exhaustive***

Probability Table

- *Weather*

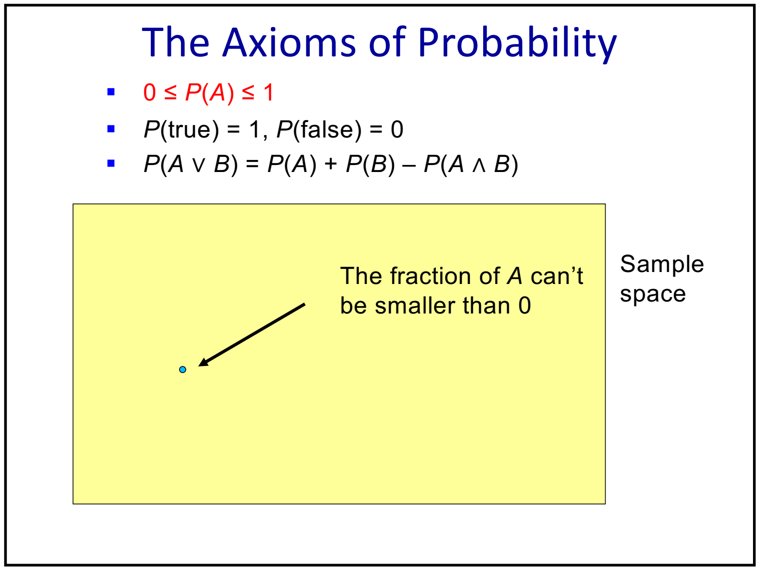
sunny	cloudy	rainy
200/365	100/365	65/365

- *P*(*Weather* = sunny) = *P*(sunny) = 200/365
- **P**(*Weather*) = 〈200/365, 100/365, 65/365〉
- We'll obtain the probabilities by counting frequencies from data

The Axioms of Probability

1. $0 \le P(A) \le 1$
2. $P(\text{true}) = 1$, $P(\text{false}) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Note: Here *P*(*A*) means *P*(*A*=*a*) for some value *a* and *P*(*A* ∨ *B*) means *P*(*A*=*a* ∨ *B*=*b*)

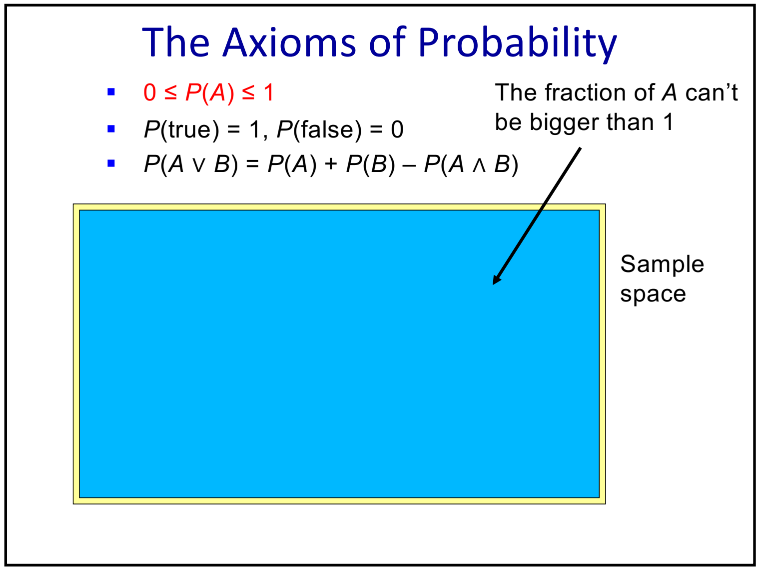


The Axioms of Probability

- $0 \le P(A) \le 1$
- $P(\text{true}) = 1$, $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

The fraction of *A* can't be smaller than 0

Sample space

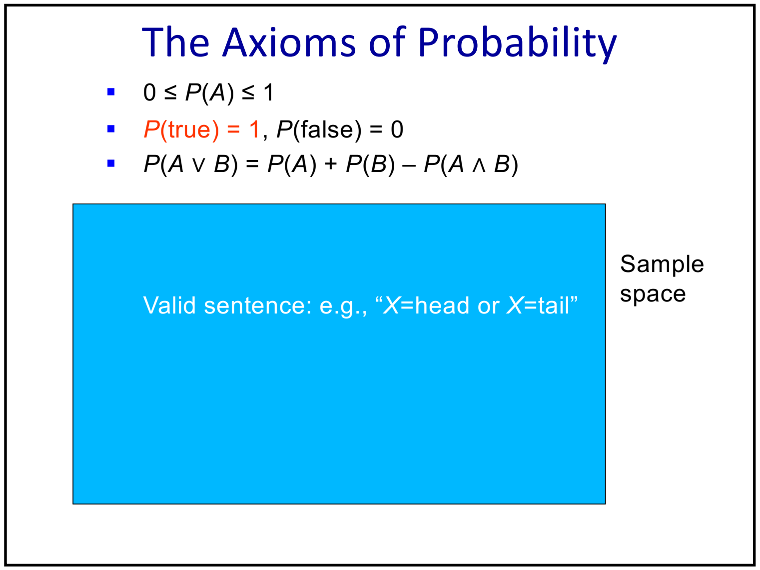


The Axioms of Probability

- $0 \le P(A) \le 1$
- $P(\text{true}) = 1$, $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

The fraction of *A* can't be bigger than 1

Sample space

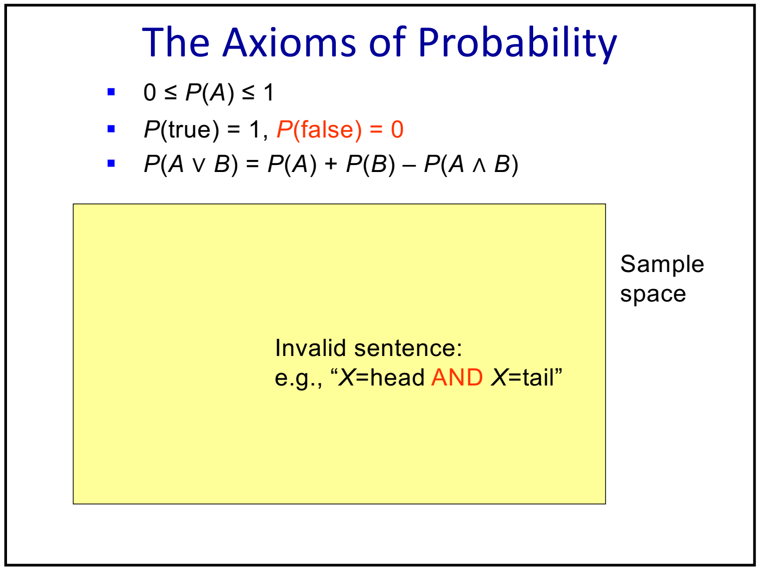


The Axioms of Probability

- $0 \le P(A) \le 1$
- $P(\text{true}) = 1$, $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Valid sentence: e.g., "*X*=head or *X*=tail"

Sample space



The Axioms of Probability

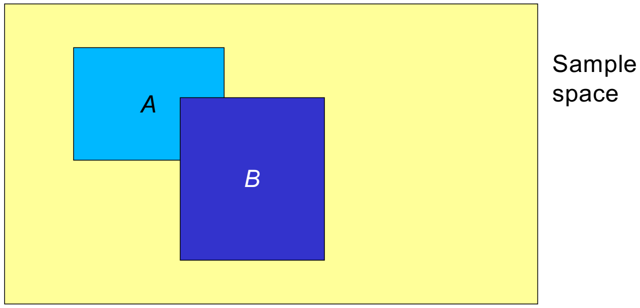
- $0 \le P(A) \le 1$
- $P(\text{true}) = 1$, $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Sample space

Invalid sentence:
e.g., "*X*=head AND *X*=tail"

The Axioms of Probability

- $0 \le P(A) \le 1$
- $P(\text{true}) = 1$, $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Some Theorems
Derived from the Axioms

- $P(\neg A) = 1 - P(A)$
- If A can take k different values $a_1, \ldots, a_k$:

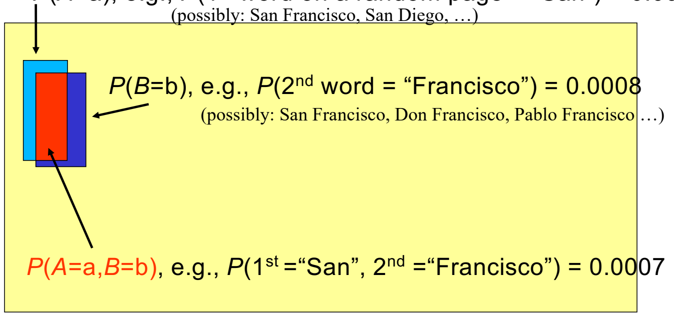
  $$P(A{=}a_1) + \ldots + P(A{=}a_k) = 1$$

- $P(B) = P(B \wedge \neg A) + P(B \wedge A)$, if A is a binary event
- $P(B) = \sum_{i=1 \ldots k} P(B \wedge A{=}a_i)$, if A can take k values

Called **Addition** or **Conditioning rule**

Joint Probability

- The **joint probability** $P(A{=}a, B{=}b)$ is shorthand for $P(A{=}a \wedge B{=}b)$, i.e., the probability of *both* A=a and B=b happening



$P(A{=}a)$, e.g., $P(1^{st}$ word on a random page = "San") = 0.001
(possibly: San Francisco, San Diego, …)

$P(B{=}b)$, e.g., $P(2^{nd}$ word = "Francisco") = 0.0008
(possibly: San Francisco, Don Francisco, Pablo Francisco …)

$P(A{=}a, B{=}b)$, e.g., $P(1^{st}$ ="San", 2^{nd} ="Francisco") = 0.0007

Full Joint Probability Distribution (FJPD)

Weather

		sunny	*cloudy*	*rainy*
Temp	*hot*	150/365	40/365	5/365
	cold	50/365	60/365	60/365

- $P(Temp{=}hot, Weather{=}rainy) = P(hot, rainy) =$ 5/365 = 0.014
- The **full joint probability distribution** table for n random variables, each taking k values, has k^n entries

Full Joint Probability Distribution (FJPD)

Bird	***Flier***	***Young***	**Probability**
T	T	T	0.0
T	T	F	0.2
T	F	T	0.04
T	F	F	0.01
F	T	T	0.01
F	T	F	0.01
F	F	T	0.23
F	F	F	0.5

3 Boolean random variables ⇒ $2^3 - 1 = 7$ “degrees of freedom” (DOF) or “independent values”

Sums to 1

Computing from the FJPD

- **Marginal Probabilities**
  - *P*(*Bird*=T) = *P*(*bird*) = 0.0 + 0.2 + 0.04 + 0.01 = 0.25
  - *P*(*bird*, ¬*flier*) = 0.04 + 0.01 = 0.05
  - *P*(*bird* ∨ *flier*) = 0.0 + 0.2 + 0.04 + 0.01 + 0.01 + 0.01 = 0.27
- Sum over all other variables
- “**Summing Out**”
- “**Marginalization**”

Unconditional / Prior Probability

- One’s uncertainty or original assumption about an event *prior* to having any data about it ***or anything else*** in the domain
- *P*(*Coin* = heads) = 0.5
- *P*(*Bird* = T) = 0.0 + 0.2 + 0.04 + 0.01 = 0.22
- Compute from the FJPD by marginalization

Marginal Probability

Weather

		sunny	*cloudy*	*rainy*
Temp	*hot*	150/365	40/365	5/365
	cold	50/365	60/365	60/365
	∑	200/365	100/365	65/365

P(*Weather*) = ⟨200/365, 100/365, 65/365⟩

Probability ***distribution*** for r.v. *Weather*

The name comes from the old days when the sums were written in the margin of a page

Marginal Probability

Weather

		sunny	cloudy	rainy	$\sum$
Temp	hot	150/365	40/365	5/365	195/365
	cold	50/365	60/365	60/365	170/365

P(*Temp*) = 〈195/365, 170/365〉

This is nothing but $P(B) = \sum_{i=1 \ldots k} P(B \wedge A=a_i)$,
where A can take k values

Conditional Probability

- Conditional probabilities
  - formalizes the process of accumulating evidence and updating probabilities based on new evidence
  - specifies the belief in a proposition (event, conclusion, diagnosis, etc.) that is *conditioned on* a proposition (evidence, feature, symptom, etc.) being true
- $P(a \mid e)$: conditional probability of $A=a$ *given* $E=e$ evidence is *all that is known true*

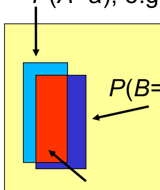
$$\boxed{P(a \mid e) = P(a \wedge e) \,/\, P(e) = P(a, e) \,/\, P(e)}$$

  - conditional probability can viewed as the joint probability $P(a, e)$ normalized by the prior probability, $P(e)$

Conditional Probability

The conditional probability $P(A=a \mid B=b)$ is the fraction of time $A=a$, within the region where $B=b$

$P(A=a)$, e.g. $P(1^{st}$ word on a random page = "San") = 0.001



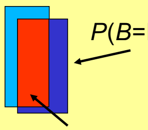
$P(B=b)$, e.g. $P(2^{nd}$ word = "Francisco") = 0.0008

$P(A=a \mid B=b)$, e.g. $P(1^{st}$="San" | 2^{nd} ="Francisco") = ?
(possibly: San, Don, Pablo …)

Conditional Probability

- P(san | francisco)
  = #(1^{st}=s and 2^{nd}=f) / #(2^{nd}=f)
  = P(san ∧ francisco) / P(francisco)
  = 0.0007 / 0.0008
  = 0.875

P(s)=0.001
P(f)=0.0008
P(s,f)=0.0007



$P(B=b)$, e.g. $P(2^{nd}$ word = "Francisco") = 0.0008

$P(A=a \mid B=b)$, e.g. $P(1^{st}$="San" | 2^{nd} ="Francisco") = **0.875**
(possibly: San, Don, Pablo …)

Although "San" is rare and "Francisco" is rare, given "Francisco" then "San" is quite likely!

Conditional Probability

Conditional probabilities behave exactly like standard probabilities; for example:

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

conditional probabilities are between 0 and 1 inclusive

$P(a_1 \mid e) + P(a_2 \mid e) + \ldots + P(a_k \mid e) = 1$

conditional probabilities sum to 1 where $a_1, \ldots, a_k$ are all values in the domain of random variable A

$P(\neg a \mid e) = 1 - P(a \mid e)$

negation for Boolean random variable A

Computing Conditional Probability

$P(\neg B|F) \quad = \ ?$

$P(F) \quad = \ ?$

Note: $P(\neg B|F)$ means $P(B\text{=false} \mid F\text{=true})$ and $P(F)$ means $P(F\text{=true})$

Full Joint Probability Distribution

Bird **(*B*)**	***Flier*** **(*F*)**	***Young*** **(*Y*)**	**Probability**
T	T	T	0.0
T	T	F	0.2
T	F	T	0.04
T	F	F	0.01
F	T	T	0.01
F	T	F	0.01
F	F	T	0.23
F	F	F	0.5

3 Boolean random variables ⇒ $2^3 - 1 = 7$ "degrees of freedom" or "independent values"

Sums to 1

Computing Conditional Probability

$P(\neg B|F) \quad = P(\neg B, F)/P(F)$

$= (P(\neg B, F, Y) + P(\neg B, F, \neg Y))/P(F)$

$= (0.01 + 0.01)/P(F)$

$P(F) \quad = P(F, B, Y) + P(F, B, \neg Y) + P(F, \neg B, Y) + P(F, \neg B, \neg Y)$

$= 0.0 + 0.2 + 0.01 + 0.01$

$= 0.22$

Marginalization

Computing Conditional Probability

- Instead of using Marginalization to compute P(F), can alternatively use **Normalization**:
- $P(\neg B|F) = .02/P(F)$ from previous slide
- $P(\neg B|F) + P(B|F) = 1$ by definition
- $P(B|F) = P(B,F)/P(F) = (0.0 + 0.2)/P(F)$
- So, $0.02/P(F) + 0.2/P(F) = 1$
- Hence, $P(F) = 0.22$

Normalization

Addition rule

- In general, $P(A \mid B) = \alpha\, P(A, B)$
  where $\alpha = 1/P(B) = 1/(P(A, B) + P(\neg A, B))$
- $P(Q \mid E_1, \ldots, E_k) = \alpha\, P(Q, E_1, \ldots, E_k)$

$$= \alpha \textstyle\sum_Y P(Q, E_1, \ldots, E_k, Y)$$

Conditional Probability with Multiple Evidence

$$P(\neg B \mid F, \neg Y) = P(\neg B, F, \neg Y) / P(F, \neg Y)$$
$$= P(\neg B, F, \neg Y) / (P(\neg B, F, \neg Y) + P(B, F, \neg Y))$$
$$= .01 /(.01 + .2)$$
$$= 0.048$$

Conditional Probability

- $P(X_1=x_1, \ldots, X_k=x_k \mid X_{k+1}=x_{k+1}, \ldots, X_n=x_n) =$
  sum of all entries in FJPD where $X_1=x_1, \ldots, X_n=x_n$ divided by sum of all entries where $X_{k+1}=x_{k+1}, \ldots, X_n=x_n$
- But this means in general we need the *entire* FJPD table, requiring an *exponential number of values* to do probabilistic inference (i.e., compute conditional probabilities)

The Chain Rule

- From the definition of conditional probability we have

$$P(A, B) = P(B) * P(A \mid B) = P(A \mid B) * P(B)$$

- It also works the other way around:

$$P(A, B) = P(A) * P(B \mid A) = P(B \mid A)\ P(A)$$

- It works with more than 2 events too:

$$P(A_1, A_2, \ldots, A_n) =$$
$$P(A_1) * P(A_2 \mid A_1) * P(A_3 \mid A_1, A_2) * \ldots$$
$$* P(A_n \mid A_1, A_2, \ldots, A_{n-1})$$

Called "**Product Rule**"

Called "**Chain Rule**"

Probabilistic Reasoning

How do we use probabilities in AI?

- You wake up with a headache
- Do you have the flu?
- *H* = headache, *F* = flu

Logical Inference: if *H* then *F*
(but the world is usually not this simple)

Statistical Inference: compute the probability of a query/diagnosis/decision given (i.e., conditioned on) evidence/symptom/observation, i.e., $P(F \mid H)$

[Example from Andrew Moore]

Example

Statistical Inference: Compute the probability of a diagnosis, *F*, given symptom, *H*, where *H* = "has a headache" and *F* = "has flu"

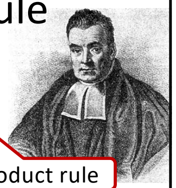
That is, compute $P(F \mid H)$

You know that

- $P(H) = 0.1$ "one in ten people has a headache"
- $P(F) = 0.01$ "one in 100 people has flu"
- $P(H \mid F) = 0.9$ "90% of people who have flu have a headache"

[Example from Andrew Moore]

Inference with Bayes's Rule



Thomas Bayes, "Essay Towards Solving a Problem in the Doctrine of Chances," 1764

$$P(F \mid H) = \frac{P(F,H)}{P(H)} = \frac{P(H \mid F)P(F)}{P(H)}$$

Def of cond. prob.

Product rule

- $P(H) = 0.1$ "one in ten people has a headache"
- $P(F) = 0.01$ "one in 100 people has flu"
- $P(H|F) = 0.9$ "90% of people who have flu have a headache"

- $P(F|H) = (0.9 * 0.01) / 0.1 = 0.09$
- So, there's a 9% chance you have flu – much less than 90%
- But it's higher than $P(F) = 1\%$ since you have a headache

Bayes's Rule

- Bayes's Rule is the basis for probabilistic reasoning given a prior model of the world, P(Q), and a new piece of evidence, E, Bayes's rule says how this piece of evidence decreases our ignorance about the world
- Initially, know P(Q) ("prior")
- Update after knowing E ("posterior"):

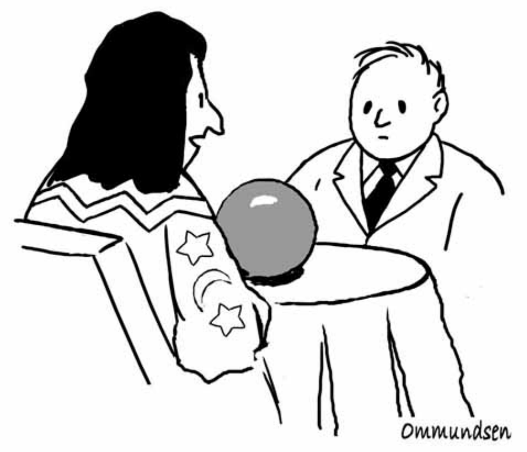
$$P(Q|E) = P(Q)\frac{P(E|Q)}{P(E)}$$

Inference with Bayes's Rule

***P*(*A*|*B*) = *P*(*B* | *A*)*P*(*A*) / *P*(*B*)** **Bayes's rule**

- Why do we make things this complicated?
  - Often *P*(*B*|*A*), *P*(*A*), *P*(*B*) are easier to get
- Some terms:
  - **Prior**: ***P*(*A*)**: probability of *A before* any evidence
  - **Likelihood**: ***P*(*B*|*A*)**: assuming *A*, how likely is the evidence *B*
  - **Posterior**: ***P*(*A*|*B*)**: probability of *A* after knowing evidence *B*
  - **(Deductive) Inference**: deriving an unknown probability from known ones

Bayes's Rule in Practice



"Is this needed for a Bayesian analysis?"

Summary of Important Rules

- **Conditional Probability**: $P(A|B) = P(A,B)/P(B)$
- **Product rule**: $P(A,B) = P(A|B)P(B)$
- **Chain rule**: $P(A,B,C,D) = P(A|B,C,D)P(B|C,D)P(C|D)P(D)$
- **Conditionalized version of Chain rule**:

  $$P(A,B|C) = P(A|B,C)P(B|C)$$
- **Bayes's rule**: $P(A|B) = P(B|A)P(A)/P(B)$
- **Conditionalized version of Bayes's rule**:

  $$P(A|B,C) = P(B|A,C)P(A|C)/P(B|C)$$
- **Addition / Conditioning rule**: $P(A) = P(A,B) + P(A,\neg B)$

  $$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$$

Common Mistake

- $P(A) = 0.3$ so $P(\neg A) = 1 – P(A) = 0.7$

- $P(A|B) = 0.4$ so $P(\neg A|B) = 1 – P(A|B) = 0.6$
  because $P(A|B) + P(\neg A|B) = 1$

  but $P(A|\neg B) \neq 0.6$ (in general)
  because $P(A|B) + P(A|\neg B) \neq 1$ in general

Quiz

- A doctor performs a test that has 99% reliability, i.e., 99% of people who are sick test positive, and 99% of people who are healthy test negative. The doctor estimates that 1% of the population is sick.
- Question: A patient tests positive. What is the chance that the patient is sick?
- 0-25%, 25-75%, 75-95%, or 95-100%?

Quiz

- A doctor performs a test that has 99% reliability, i.e., 99% of people who are sick test positive, and 99% of people who are healthy test negative. The doctor estimates that 1% of the population is sick.
- Question: A patient tests positive. What is the chance that the patient is sick?
- 0-25%, 25-75%, 75-95%, or 95-100%?
- Common answer: 99%; Correct answer: 50%

Given:

$P(TP \mid S) = 0.99$

$P(\neg TP \mid \neg S) = 0.99$

$P(S) = 0.01$

TP = "tests positive"
S = "is sick"

Query:

$P(S \mid TP) = ?$

$P(TP \mid S) = 0.99$

$P(\neg TP \mid \neg S) = 0.99$

$P(S) = 0.01$

$P(S \mid TP) =$

$P(TP \mid S)\, P(S) \,/\, P(TP)$

$= (0.99)(0.01) \,/\, P(TP) = 0.0099/P(TP)$

$P(\neg S \mid TP) = P(TP \mid \neg S)P(\neg S) \,/\, P(TP)$

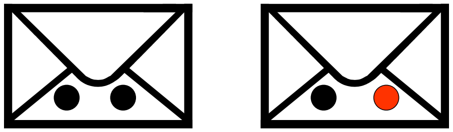
$= (1 – 0.99)(1 – 0.01) \,/\, P(TP) = 0.0099/P(TP)$

$0.0099/P(TP) + 0.0099/P(TP) = 1$, so $P(TP) = 0.0198$

So, $P(S \mid TP) = 0.0099 \,/\, 0.0198 = 0.5$

Inference with Bayes's Rule

- In a bag there are two envelopes
  - one has a red ball (worth \$100) and a black ball
  - one has two black balls. Black balls are worth nothing



- You randomly grab an envelope, and randomly take out one ball – it's **black**
- At this point you're given the option to switch envelopes. Should you switch or not?

Similar to the "Monty Hall Problem"

Inference with Bayes's Rule

E: envelope, 1 = (R,B), 2 = (B,B)

B: the event of drawing a *black* ball

Given: $P(B|E=1) = 0.5$, $P(B|E=2) = 1$, $P(E=1) = P(E=2) = 0.5$

Query: Is $P(E=1 \mid B) > P(E=2 \mid B)$?

Use Bayes's rule: $P(E|B) = P(B|E)*P(E) \,/\, P(B)$

Conditioning rule

$P(B) = P(B|E=1)P(E=1) + P(B|E=2)P(E=2) = (.5)(.5) + (1)(.5) = .75$

$P(E=1|B) = P(B|E=1)P(E=1)/P(B) = (.5)(.5)/(.75) = 0.33$

$P(E=2|B) = P(B|E=2)P(E=2)/P(B) = (1)(.5)/(.75) = 0.67$

After seeing a black ball, the posterior probability of this envelope being #1 (thus worth \$100) is *smaller* than it being #2

Thus you should switch!

Another Example

- 1% of women over 40 who are tested have breast cancer. 85% of women who really *do* have breast cancer have a positive mammography test (true positive rate). 8% who do *not* have cancer will have a positive mammography (false positive rate).
- Question: A patient gets a positive mammography test. What is the chance she has breast cancer?

- Let Boolean random variable M mean "positive mammography test"
- Let Boolean random variable C mean "has breast cancer"
- Given:

$P(C) = 0.01$

$P(M|C) = 0.85$

$P(M|\neg C) = 0.08$

Compute the posterior probability: $P(C|M)$

- $P(C|M) = P(M|C)P(C)/P(M)$ by Bayes's rule
  $= (.85)(.01)/P(M)$
- $P(M) = P(M|C)P(C) + P(M|\neg C)P(\neg C)$ by the Conditioning rule
- So, $P(C|M) = .0085/[(.85)(.01) + (.08)(1-.01)]$
  $= 0.097$
- So, there is only a 9.7% chance that if you have a positive test you really have cancer!

Independence

Two events A, B are **independent** if the following hold:

- $P(A, B) = P(A) * P(B)$
- $P(A, \neg B) = P(A) * P(\neg B)$

  …
- $P(A \mid B) = P(A)$
- $P(B \mid A) = P(B)$
- $P(A \mid \neg B) = P(A)$

  …

Independence

- Independence is a kind of domain knowledge
  - Needs an understanding of ***causation***
  - Very strong assumption
- Example: P(burglary) = 0.001 and P(earthquake) = 0.002
  - Let's say they are *independent*
  - The full joint probability table = ?

Independence

- Given: $P(B) = 0.001$, $P(E) = 0.002$, $P(B|E) = P(B)$
- The full joint probability distribution table (FJPD) is:

Burglary	Earthquake	Prob.
B	E	$= P(B)P(E)$
B	$\neg E$	
$\neg B$	E	
$\neg B$	$\neg E$	

- Need only 2 numbers to fill in entire table
- Now we can do anything, since we have the FJPD

Independence

- Given n independent, Boolean random variables, the FJPD has 2^n entries, but we only need n numbers (degrees of freedom) to fill in entire table
- Given n independent random variables, where each can take k values, the FJPD table has:
  - k^n entries
  - Only $n(k\text{-}1)$ numbers needed (DOFs)

Conditional Independence

- Random variables can be ***dependent***, but **conditionally independent**
- Example: Your house has an alarm
  - Neighbor John calls when he hears the alarm
  - Neighbor Mary calls when she hears the alarm
  - Assume John and Mary don't talk to each other
- Is *JohnCall independent* of *MaryCall*?
  - **No** – If John called, it is likely the alarm went off, which increases the probability of Mary calling
  - $P(MaryCall \mid JohnCall) \neq P(MaryCall)$

Conditional Independence

- But, if we *know* the status of the *Alarm*, *JohnCall* will ***not*** affect whether or not Mary calls

  $P(MaryCall \mid Alarm, JohnCall) = P(MaryCall \mid Alarm)$
- We say *JohnCall* and *MaryCall* are **conditionally independent** given *Alarm*
- In general, "***A* and *B* are conditionally independent given *C***" means:

  $P(A \mid B, C) = P(A \mid C)$

  $P(B \mid A, C) = P(B \mid C)$

  $P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$

Independence vs. Conditional Independence

- Say Alice and Bob each toss ***separate coins***. *A* represents "Alice's coin toss is heads" and *B* represents "Bob's coin toss is heads"
- *A* and *B* are **independent**
- Now suppose Alice and Bob toss the ***same coin***. Are *A* and *B independent*?
  - *No*. Say the coin may be biased towards heads. If *A* is heads, it will lead us to increase our belief in *B* beings heads. That is, $P(B|A) > P(A)$

- Say we add a new variable, *C*: "the coin is biased towards heads"
- The values of *A* and *B* are *dependent on C*
- But if we know *for certain* the value of *C* (true or false), then any evidence about *A* cannot change our belief about *B*
- That is, $P(B|C) = P(B|A, C)$
- *A* and *B* are **conditionally independent** given *C*

Revisiting Earlier Example

- Let Boolean random variable M mean "positive mammography test"
- Let Boolean random variable C mean "has breast cancer"
- Given:

$P(C) = 0.01$

$P(M|C) = 0.85$

$P(M|\neg C) = 0.08$

Bayes's Rule with Multiple Evidence

- Say the same patient goes back and gets a *second* mammography and it too is positive. Now, what is the chance she has Cancer?
- Let $M1$, $M2$ be the 2 positive tests
- $M1$ and $M2$ are ***not*** independent
- Compute posterior: $P(C|M1, M2)$

Bayes's Rule with Multiple Evidence

- $P(C|M1, M2) = P(M1, M2|C)P(C)/P(M1, M2)$ by Bayes's rule

$= P(M1|M2, C)P(M2|C)P(C)/P(M1, M2)$

Conditionalized Chain rule

- $P(M1, M2) = P(M1, M2|C)P(C) + P(M1, M2|\neg C)P(\neg C)$ by Conditioning rule

$= P(M1|M2, C)P(M2|C)P(C) + P(M1|M2, \neg C)P(M2|\neg C)P(\neg C)$

by Conditionalized Chain rule

Cancer "causes" a positive test, so ***M*1 and *M*2 are conditionally independent given *C***, so

- $P(M1|M2, C) = \text{P}(M1 \mid C) = 0.85$
- $\text{P}(M1, M2) = P(M1|M2, C)P(M2|C)P(C) + P(M1|M2, \neg C)P(M2|\neg C)P(\neg C)$

$= P(M1|C)P(M2|C)P(C) + P(M1|\neg C)P(M2|\neg C)P(\neg C)$ by cond. indep.

$= (.85)(.85)(.01) + (.08)(.08)(1-.01)$

$= 0.01356$

So, $P(C|M1, M2) = (.85)(.85)(.01)/ .01356$

$= 0.533$ or 53.3%

Example

- Prior probability of having breast cancer:

  $P(C) = 0.01$
- Posterior probability of having breast cancer after 1 positive mammography:

  $P(C|M1) = 0.097$
- Posterior probability of having breast cancer after 2 positive mammographies (and cond. independence assumption):

  $P(C|M1, M2) = 0.533$

Bayes's Rule with Multiple Evidence

- Say the same patient goes back and gets a second mammography and it is **negative**. Now, what is the chance she has cancer?
- Let $M1$ be the positive test and $\neg M2$ be the negative test
- Compute posterior: $P(C|M1, \neg M2)$

Bayes's Rule with Multiple Evidence

- $P(C|M1, \neg M2) = P(M1, \neg M2|C)P(C)/ P(M1, \neg M2)$

  by Bayes's rule

  $= P(M1|C)P(\neg M2|C)P(C)/P(M1, \neg M2)$

  $= (.85)(1-.85)(.01)/P(M1, \neg M2)$
- $P(M1, \neg M2) = P(M1, \neg M2|C)P(C) +$

  $P(M1, \neg M2|\neg C)P(\neg C)$ by Conditioning rule

  $= P(M1|\neg M2, C)P(\neg M2|C)P(C) +$

  $P(M1|\neg M2, \neg C)P(\neg M2|\neg C)P(\neg C)$

  by Conditionalized Chain rule

Cancer "causes" a positive test, so ***M*1 and ¬*M*2 are conditionally independent given *C***, so

$P(M1|\neg M2, C)P(\neg M2|C)P(C) +$

$P(M1|\neg M2, \neg C)P(\neg M2|\neg C)P(\neg C)$

$= P(M1|C)P(\neg M2|C)P(C) +$

$P(M1|\neg C)P(\neg M2|\neg C)P(\neg C)$ by cond. indep.

$= (.85)(1 - .85)(.01) + (1 - .08)(.08)(1 - .01)$

$= 0.074139$ $(= P(M1, \neg M2))$

So, $P(C|M1, \neg M2) = (.85)(1 - .85)(.01)/ .074139$

$= 0.017$ or 1.7%

Bayes's Rule with Multiple Evidence and Conditional Independence

- Assume all evidence variables, B, C and D, are conditionally independent given the diagnosis variable, A
- $P(A|B,C,D) = P(B,C,D|A)P(A)/P(B,C,D)$
  $= \underbrace{P(B|A)P(C|A)P(D|A)}_{\text{Conditionalized Chain rule + conditional independence}}P(A)/\underbrace{P(D|B,C)P(C|B)P(B)}_{\text{Chain rule}}$

$$= P(A)\frac{P(B|A)}{P(B)}\frac{P(C|A)}{P(C|B)}\frac{P(D|A)}{P(D|B,C)}$$

Inference Ignorance

- "Inferences about Testosterone Abuse Among Athletes," 2004
  - Mary Decker Slaney doping case
- "Justice Flunks Math," 2013
  - Amanda Knox trial in Italy

Naïve Bayes Classifier

- Classification problem: Find the value of class/decision/diagnosis variable *Y* that is most likely given evidence/measurements/attributes $X_i = v_i$
- Use Bayes's rule and conditional independence:

$$P(Y = c \mid X_1 = v_1, X_2 = v_2, \ldots, X_n = v_n)$$
$$= P(Y = c)P(X_1 = v_1 \mid Y = c) \ldots P(X_n = v_n \mid Y = c) / P(X_1 = v_1, \ldots, X_n = v_n)$$

- Try all possible values of *Y* and pick the value that gives the maximum probability
- But denominator, $P(X_1=v_1, \ldots, X_n=v_n)$, is a constant for all values of *Y*, so it won't affect which value of *Y* is best

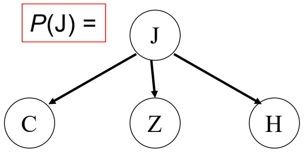
Naive Bayes Classifier Testing Phase

- For a given test instance defined by $X_1=v_1, \ldots, X_n=v_n$, compute

$$argmax_c \underbrace{P(Y=c)}_{\text{Class variable}} \prod_{i=1}^{n} \underbrace{P(X_i=v_i \mid Y=c)}_{\text{Evidence variable}}$$

- Assumes all evidence variables are conditionally independent of each other given the class variable
- Robust because it gives the right answer as long as the correct class is more likely than all others

Naïve Bayes Classifier Training Phase



P(J) =

P(C|J) =
P(C|¬J) =

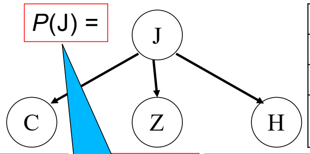
P(Z|J) =
P(Z|¬J) =

P(H|J) =
P(H|¬J) =

J	Person is Junior
C	Brought coat to class
Z	Lives in zipcode 53706
H	Saw “Hunger Games 1” more than once

Compute from the Training set all the necessary Prior and Conditional probabilities

Naïve Bayes Classifier



P(J) =

P(C|J) =
P(C|¬J) =

Z|J) =
¬J) =

P(H|J) =
P(H|¬J) =

J	Person is Junior
C	Brought coat to class
Z	Lives in zipcode 53706
H	Saw “Hunger Games 1” more than once

\# Juniors

\# people in database

\# Juniors who saw H>1

\# Juniors

\# non-juniors who saw H>1

\# non-juniors

Naïve Bayes Classifier

- Assume k classes and n evidence (i.e., attribute) variables, each with m possible values
- k-1 values needed for computing $P(Y=c)$
- (m-1)k values needed for computing $P(X_i=v_i \mid Y=c)$ for each evidence variable X_i
- So, (k-1) + n(m-1)k values needed instead of exponential size FJPD table

Naïve Bayes Classifier

- Conditional probabilities can be very, very small, so instead use logarithms to avoid underflow:

$$\arg\max_c \log P(Y=c) + \sum_{i=1}^{n} \log P(X_i = v_i \mid Y=c)$$

Add-1 Smoothing

- **Unseen event problem**: Training data may *not* include some cases
  - – flip a coin 3 times, all heads → one-sided coin?
  - – Conditional probability = 0
  - – Just because a value doesn't occur in the training set doesn't mean it will never occur
- "**Add-1 Smoothing**" ensures that *every* conditional probability > 0 by pretending that you've seen each attribute's value 1 extra time

Add-1 Smoothing

- Compute **Conditional probabilities** as

number of times attribute *X* has value v_i in all training instances with class *c*

$$P(X = vi \,|Y = c) = \frac{count(X = vi, Y = c) + 1}{count(Y = c) + m}$$

number of training instances with class *c*

- Note: $\sum_{i=1}^{m} P(X = vi|Y = c) = 1$

where *m* = number of possible values for attribute *X*

Add-1 Smoothing

- Compute **Prior probabilities** as

$$P(Y = c) = \frac{count(Y = c) + 1}{N + k}$$

where *N* = size of the training set, *k* = number of classes

Laplace Smoothing

- aka **Add-δ Smoothing**
- Instead of adding 1, add δ (a positive real number)
- Compute **conditional probabilities** as

$$P(X{=}v_i \mid Y{=}c) = \frac{count(X{=}vi,Y{=}c)+ \delta}{count(Y{=}c)+ \delta m}$$

where *m* = number of possible values for attribute *X*