Probability Theory for Inference

Discrete random variables

- **A *random variable* can take on one of a set of different values, each with an associated probability. Its value at a particular time is *subject to random variation*.**
  - *Discrete* random variables take on one of a discrete (often finite) range of values
  - Domain values must be *exhaustive* and *mutually exclusive*

- **For us, random variables will have a discrete, countable (usually finite) domain of *arbitrary values*.**
  - Mathematical statistics usually calls these *random elements*
  - **Example: Weather is a discrete random variable with domain {sunny, rain, cloudy, snow}.**
  - ***Example: A Boolean random variable* has the domain {true,false},**

A word on notation

Assume *Weather* is a discrete random variable with domain {sunny, rain, cloudy, snow}.

• *Weather = sunny*	**abbreviated**	*sunny*
• *P(Weather=sunny)=0.72*	**abbreviated**	*P(sunny)=0.72*
• *Cavity = true*	**abbreviated**	*cavity*
• *Cavity = false*	**abbreviated**	*¬cavity*

Vector notation:

- **Fix order of domain elements:**
  <sunny,rain,cloudy,snow>
- **Specify the probability mass function (pmf) by a vector:**
  P*(Weather) = <0.72,0.1,0.08,0.1>*

13.2.3 Probability Axioms

- The axiomatization of probability theory by Kolmogorov (1933) based on three simple axioms

1. For any proposition a the probability is in between 0 and 1: $0 \leq P(a) \leq 1$

2. Necessarily true (i.e., valid) propositions have probability 1 and necessarily false (i.e., unsatisfiable) propositions have probability 0:
   $P(true) = 1 \qquad P(false) = 0$

3. The probability of a disjunction is given by the *inclusion-exclusion principle*
   $$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

Probability Theory

- **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=True ∧ Burglary=True ∧ Earthquake=False
  alarm ∧ burglary ∧ ¬earthquake
- P(Burglary) = .1
- P(Alarm, Burglary) =

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

Probability Theory: Definitions

- **Computing conditional prob**:
  - $P(a \mid b) = P(a \land b) / P(b)$
  - $P(b)$: **normalizing** constant
- **Product rule**:
  - $P(a \land b) = P(a \mid b)\, P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$ (**conditioning**)

Bayes' Rule & Diagnosis

$$\underbrace{P(a|b)}_{\text{Posterior}} = \frac{\overbrace{P(b|a)}^{\text{Likelihood}} * \overbrace{P(a)}^{\text{Prior}}}{\underbrace{P(b)}_{\text{Normalization}}}$$

- **Useful for assessing diagnostic probability from causal probability:**

$$P(Cause|Effect) = \frac{P(Effect|Cause) * P(Cause)}{P(Effect)}$$

Probability Summary

onditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$

oduct rule $$P(x,y) = P(x|y)P(y)$$

hain rule
$$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

Y independent if and only if: $\forall x, y : P(x,y) = P(x)P(y)$

and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y |$

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

Try It...

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

- P(alarm | burglary) = ??
- P(burglary | alarm) = ??
- P(burglary $\wedge$ alarm) = ??
- P(alarm) = ??

- **Computing conditional prob**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - 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**)

Probability Theory (cont.)

- **Conditional probability**: probability of effect given causes
- **Computing conditional probs**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **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**)

- $P(\text{burglary} \mid \text{alarm}) = .47$
  $P(\text{alarm} \mid \text{burglary}) = .9$
- $P(\text{burglary} \mid \text{alarm}) =$
  $P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm})$
  $= .09 / .19 = .47$
- $P(\text{burglary} \wedge \text{alarm}) =$
  $P(\text{burglary} \mid \text{alarm})\, P(\text{alarm}) =$
  $.47 * .19 = .09$
- $P(\text{alarm}) =$
  $P(\text{alarm} \wedge \text{burglary}) +$
  $P(\text{alarm} \wedge \neg\text{burglary}) =$
  $.09+.1 = .19$

Bayes Theorem
Application

Guilty or not?

> *A person is put in front of a jury. The jury finds the defendant guilty in* 98% *of the cases in which the defendant has committed a crime, and it finds the defendent not guilty in only* 97% *of the cases in which the defendant has not committed a crime. Furthermore, only* .008 *of the entire population has committed a crime.*

If a random person is found guilty by the jury, what's more likely: criminal or not?

Bayes Theorem
Application

Guilty or not?

A person is put in front of a jury. The jury finds the defendant guilty in 98% of the cases in which the defendant has committed a crime, and it finds the defendent not guilty in only 97% of the cases in which the defendant has not committed a crime. Furthermore, only .008 of the entire population has committed a crime.

$$P(criminal) = 0.008 \qquad P(\neg criminal) = 0.992$$
$$P(guilty|criminal) = 0.98 \qquad P(\neg guilty|criminal) = 0.02$$
$$P(guilty|\neg criminal) = 0.03 \qquad P(\neg guilty|\neg criminal) = 0.97$$

If a random person is found guilty by the jury, what's more likely: criminal or not?

which is bigger? $P(criminal|guilty)$ or $P(\neg criminal|guilty)$?

Probabilities
Bayes Rule

$$P(a \wedge b) = P(a|b)P(b)$$

$$P(a \wedge b) = P(b|a)P(b)$$

$$P(b|a)P(a) = P(a|b)P(b)$$

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Bayes Theorem
Application

Guilty or not?

$$P(criminal) = 0.008 \qquad P(\neg criminal) = 0.992$$
$$P(guilty|criminal) = 0.98 \qquad P(\neg guilty|criminal) = 0.03$$
$$P(guilty|\neg criminal) = 0.02 \qquad P(\neg guilty|\neg criminal) = 0.97$$

If a random person is found guilty by the jury, what's more likely: criminal or not?

which is bigger? $P(criminal|guilty)$ or $P(\neg criminal|guilty)$?

$$P(criminal|guilty) = \frac{P(guilty|criminal)P(criminal)}{P(guilty)}$$

$$P(\neg criminal|guilty) = \frac{P(guilty|\neg criminal)P(\neg criminal)}{P(guilty)}$$

Calculating Conditional Probabilities

College students were asked if they have ever cheated on an exam. Results were broken down by gender.

Cheated on College Exam?

Gender	Yes	No	Total
Male	.32	.22	.54
Female	.28	.18	.46
Total	.60	.40	1.00

- Question: Given that a person has cheated, what is the probability he is male?
- Answer: $P(\text{Male}|\text{Cheater}) = \dfrac{P(\text{Male} \cap \text{Cheater})}{P(\text{Cheater})}$

$$= \frac{.32}{.60} = .5333$$

	Right-handed	Left-handed	Total
Male	0.41	0.08	0.49
Female	0.45	0.06	0.51
Total	0.86	0.14	1

Find the probability that a randomly selected person is:

(a) a male given that she is right-handed;

(b) right-handed given that she is a male;

(c) a female given that she is left-handed.

(d) Are the events *being a female* and *being left-handed* independent? Justify.

a) $P(M \mid R) = \dfrac{P(M \cap R)}{P(R)} = \dfrac{0.41}{0.86} \approx 0.477$

b) $P(R \mid M) = \dfrac{P(R \cap M)}{P(M)} = \dfrac{0.41}{0.49} \approx 0.837$

b) $P(R \mid M) = \dfrac{P(R \cap M)}{P(M)} = \dfrac{0.41}{0.49} \approx 0.837$

c) $P(F \mid L) = \dfrac{P(F \cap L)}{P(L)} = \dfrac{0.06}{0.14} \approx 0.429$

d) $\left.\begin{aligned} P(F \mid L) &\approx 0.429 \\ P(F) &= 0.51 \end{aligned}\right\} \neq \Rightarrow$ F and L are <u>not</u> independent

Joint probability distribution

- **Probability assignment to all combinations of values of random variables (i.e. all elementary events)**

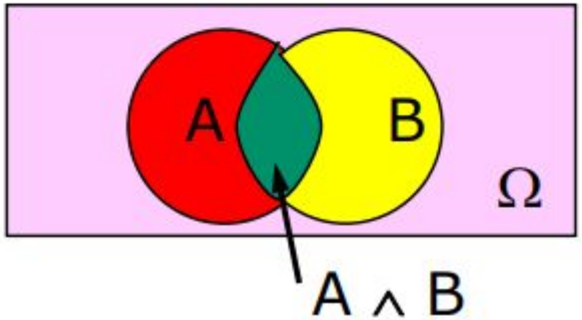
	toothache	¬ toothache
cavity	0.04	0.06
¬ cavity	0.01	0.89

- **The sum of the entries in this table has to be 1**
- ***Every question about a domain can be answered by the joint distribution***

- **Probability of a proposition is the sum of the probabilities of elementary events in which it holds**
  - P(cavity) = 0.1 [marginal of row 1]
  - P(toothache) = 0.05 [marginal of toothache column]

Conditional Probability

	toothache	¬ toothache
cavity	0.04	0.06
¬ cavity	0.01	0.89



- *P(cavity)=0.1* and *P(cavity ∧ toothache)=0.04* are both *prior* (unconditional) probabilities
- Once the agent has new evidence concerning a *previously unknown* random variable, e.g. Toothache, we can specify *a posterior* (conditional) probability e.g. *P(cavity | Toothache=true)*

$$P(a \mid b) = P(a \wedge b)/P(b)$$

[Probability of a with the Universe Ω restricted to b]

→ *The new information restricts the set of possible worlds ω_i consistent with it, so changes the probability.*

- *So P(cavity | toothache) = 0.04/0.05 = 0.8*

Conditional Probability (continued)

- **Definition of Conditional Probability:**
  $P(a \mid b) = P(a \wedge b)/P(b)$

- **Product rule gives an alternative formulation:**
  $$P(a \wedge b) = P(a \mid b) * P(b)$$
  $$= P(b \mid a) * P(a)$$

- **A general version holds for whole distributions:**
  $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity) * \mathbf{P}(Cavity)$

- **Chain rule is derived by successive application of product rule:**
  $$\mathbf{P}(A,B,C,D,E) = P(A|B,C,D,E)\ P(B,C,D,E)$$
  $$= P(A|B,C,D,E)\ P(B|C,D,E)\ P(C,D,E)$$
  $$= \ldots$$
  $$= P(A|B,C,D,E)\ P(B|C,D,E)\ P(C|D,E)\ P(D|E)\ P(E)$$

Probabilistic Inference

- ***Probabilistic inference:*** **the computation**
  - from *observed evidence*
  - of *posterior probabilities*
  - for *query propositions*.
- **We use the *full joint distribution* as the "knowledge base" from which answers to questions may be derived.**
- **Ex: three Boolean variables *Toothache (T), Cavity (C), ShowsOnXRay (X)***

	t		$\neg t$	
	x	$\neg x$	x	$\neg x$
c	0.108	0.012	0.072	0.008
$\neg c$	0.016	0.064	0.144	0.576

- **Probabilities in joint distribution sum to 1**

Probabilistic Inference II

	t		¬t	
	x	¬x	x	¬ x
c	0.108	0.012	0.072	0.008
¬c	0.016	0.064	0.144	0.576

- **Probability of any proposition computed by finding atomic events where proposition is true and adding their probabilities**
  - *P(cavity ∨ toothache)*
    = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064
    = 0.28
  - *P(cavity)*
    = 0.108 + 0.012 + 0.072 + 0.008
    = 0.2
- ***P(cavity)* is called a *marginal probability* and the process of computing this is called *marginalization***

Probabilistic Inference III

	t		¬t	
	x	¬x	x	¬ x
c	0.108	0.012	0.072	0.008
¬c	0.016	0.064	0.144	0.576

- **Can also compute conditional probabilities.**
- *P(¬ cavity | toothache)*
  = P(¬ cavity ∧ toothache)/P(toothache)
  = (0.016 + 0.064) / (0.108 + 0.012 + 0.016 + 0.064)
  = 0.4
- **Denominator is viewed as a** *normalization constant***:**
  - Stays constant no matter what the value of Cavity is.
    (Book uses α to denote normalization constant 1/P(X), for random variable X.)

13.3 Inference Using Full Joint Distribution

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- E.g., there are six atomic events for cavity ∨ toothache: 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28
- Extracting the distribution over a variable (or some subset of variables), *marginal probability*, is attained by adding the entries in the corresponding rows or columns
- E.g., P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2
- We can write the following general marginalization (summing out) rule for any sets of variables **Y** and **Z**:

$$\underline{P}(\mathbf{Y}) = \sum_{\mathbf{z} \in \mathbf{Z}} \underline{P}(\mathbf{Y}, \mathbf{z})$$

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- Computing a conditional probability

  $P(cavity \mid toothache) =$

  $P(cavity \wedge toothache)/P(toothache) =$

  $(0.108 + 0.012)/(0.108 + 0.012 + 0.016 + 0.064) =$

  $0.12/0.2 = 0.6$

- Respectively

  $P(\neg cavity \mid toothache) =$

  $(0.016 + 0.064)/0.2 = 0.4$

- The two probabilities sum up to one, as they should

13.4 Independence

- If we expand the previous example with a fourth random variable Weather, which has four possible values, we have to copy the table of joint probabilities four times to have 32 entries together
- Dental problems have no influence on the weather, hence:

$$P(Weather = cloudy \mid toothache, catch, cavity) = P(Weather = cloudy)$$

- By this observation and product rule

$$P(toothache, catch, cavity, Weather = cloudy) = P(Weather = cloudy)\, P(toothache, catch, cavity)$$

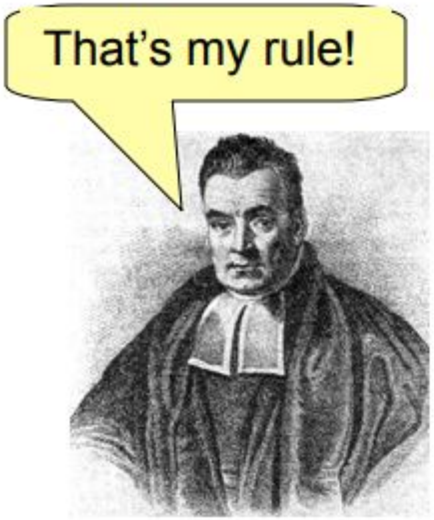
Conditional Independence

- Absolute independence:
  - A and B are **independent** 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:
  - $P(A \wedge B \wedge C) = P(A \mid C)\,P(B \mid C)\,P(C)$
- 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

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$



- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build a conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later

- In the running for most important AI equation!

ayes' Rule & Diagnosis

$$\underbrace{P(a|b)}_{\textbf{Posterior}} = \frac{\overbrace{P(b|a)}^{\textbf{Likelihood}} * \overbrace{P(a)}^{\textbf{Prior}}}{\underbrace{P(b)}_{\textbf{Normalization}}}$$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause) * P(Cause)}{P(Effect)}$$

Bayes' Rule For Diagnosis II

$$P(Disease \mid Symptom) = \frac{P(Symptom \mid Disease) * P(Disease)}{P(Symptom)}$$

Imagine:

- **disease = TB, symptom = coughing**
- ***P(disease | symptom)* is different in TB-indicated country vs. USA**
- ***P(symptom | disease)* should be the same**
  - It is more widely useful to learn *P(symptom | disease)*
- **What about P(symptom)?**
  - Use *conditioning* (next slide)
  - For determining, e.g., the *most likely* disease given the symptom, we can just ignore P(symptom)!!! (see slide 35)

Conditioning

- *Idea:* Use *conditional probabilities* instead of joint probabilities
- $P(a) = P(a \wedge b) + P(a \wedge \neg b)$
  $= P(a \mid b) * P(b) + P(a \mid \neg b) * P(\neg b)$

  <u>*Here*</u>:

  $P(symptom) = P(symptom \mid disease) * P(disease)$
  $P(symptom \mid \neg disease) * P(\neg disease)$

- More generally: $P(Y) = \sum_z P(Y|z) * P(z)$
- Marginalization and conditioning are useful rules for derivations involving probability expressions.

Conditional Independence

UT *absolute* independence is rare

entistry is a large field with hundreds of variables, one of which are independent. What to do?

and B are ***conditionally independent given C*** iff

- $P(A \mid B, C) = P(A \mid C)$
- $P(B \mid A, C) = P(B \mid C)$
- $P(A \wedge B \mid C) = P(A \mid C) * P(B \mid C)$

oothache (T), Spot in Xray (X), Cavity (C)

- None of these are independent of the other two
- But ***T and X are conditionally independent given C***

Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - $P(+catch \mid +toothache, +cavity) = P(+catch \mid +cavity)$
- The same independence holds if I don't have a cavity:
  - $P(+catch \mid +toothache, -cavity) = P(+catch \mid -cavity)$
- Catch is *conditionally independent* of Toothache given Cavity:
  - $P(Catch \mid Toothache, Cavity) = P(Catch \mid Cavity)$
- Equivalent statements:
  - $P(Toothache \mid Catch, Cavity) = P(Toothache \mid Cavity)$
  - $P(Toothache, Catch \mid Cavity) = P(Toothache \mid Cavity)\, P(Catch \mid Cavity)$
  - One can be derived from the other easily

Conditional Independence II *WHY??*

- **If I have a cavity, the probability that the XRay shows a spot doesn't depend on whether I have a toothache (and vice vers**

  $P(X|T,C) = P(X|C)$

- **From which follows:**

  $P(T|X,C) = P(T|C)$ **and** $P(T,X|C) = P(T|C) * P(X|$

- **By the chain rule), given conditional independence:**

  $$P(T,X,C) = P(T|X,C) * P(X,C) = P(T|X,C) * P(X|C) * P(C)$$

  $$= P(T|C) * P(X|C) * P(C)$$

- P(*Toothache, Cavity, Xray*) has $2^3 - 1 = 7$ independent entries
- **Given conditional independence, chain rule yields**

  2 + 2 + 1 = 5 independent numbers

Conditional Independence III

- **In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from *exponential* in *n* to *linear* in *n*.**

- ***Conditional independence is our most basic and robust form of knowledge about uncertain environments.***

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 later! ☺**

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*?**

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*?**