Probability Review

slides adapted from

S. Russell, D. Klein, D. Koller, A. Moore, N. Onder

Outline

- Probability Review
  - Random Variables
  - Distributions
    - Joint
    - Marginal
    - Conditional
  - Rules
    - Product
    - Chain
    - Bayes’
  - Independence

Handling Uncertainty

- How to represent relations in the presence of uncertainty?
  - Build **models** to capture uncertainty in state of world, dynamics of system, and the observations
- How to use the representation to make inferences?
  - Use the model to **reason** about the world
- Questions
  - What formalism to use?
  - What queries can be asked of the model?
  - How can the model be constructed?

Approaches for Uncertainty

- Default or nonmonotonic reasoning
  - Optimistic reasoning – believe something until evidence to the contrary
- Fuzzy logic
  - Allows events to be “sort of true”
- Propositional Logic with certainty factors
  - Extension of propositional and first-order logic
- Probability

Probability

Probability theory is a well-defined framework for modeling and reasoning with uncertainty

- Has clear semantics
- Principled methods for different reasoning tasks
- Intuitive to humans

- Issues with efficient reasoning

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - *C*: Will/did a tossed coin come up heads or tails?
  - *B*: Will a customer buy a new iPhone?
  - *R*: Is it raining?
- A discrete random variable X takes value from a discrete set called the domain or sample space Ω_X
  - *C*: coin toss, Ω_C = { *heads, tails* }
  - *D:* roll of a die; Ω_D = { 1, 2, 3, 4, 5, 6 }
  - *B*: does a customer buy a phone; Ω_B = { *True, False* }
- An event is a subset of Ω_X
  - e_1 = { 1 } die roll of 1
  - e_2 = { 1, 3, 5 } odd value of a die roll

Probabilities

- For a discrete random variable *X* each value $x \in \Omega_X$ has a probability of occurring P(*X*=*x*) or P(*x*)
- Ex. *D* = roll of a fair die
  - P(*D=1*) = 1/6,
  - P(*D=2*) = 1/6,
  - ...
  - P(*D=6*) = 1/6

Unconditional or prior probabilities

- Ex. *S* = patient has sickness, pneumonia
  - P(*S=True*) = 0.001
  - P(*S=False*) = 0.999

Axioms of Probability

1. $0 \leq P(A) \leq 1$

2. *P*(*True*) = 1 and *P*(*False*) = 0

or

$$P(\Omega_A) = 1$$

3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

Probability Distribution

Defines probabilities for all values of a random variable

- S = patient has Pneumonia

Sickness, S	**P(*S*)**
True	0.001
False	0.999

- W = WBCcount

WBCcount, *W*	**P(*W*)**
high	0.005
normal	0.993
low	0.002

- Requirements

$$\forall x \;\; P(X = x) \geq 0$$

$$\sum_x P(X = x) = 1$$

Joint Probability Distribution

- Defines probabilities for all possible assignments of values of variables in a set, $X_1, X_2, \ldots, X_n$

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

$$P(x_1, x_2, \ldots, x_n)$$ shorthand for above statement

  - Must obey

$$P(x_1, x_2, \ldots, x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots, x_n)} P(x_1, x_2, \ldots, x_n) = 1$$

- Size of distribution, *n* variables w/ domains of size *d?*

Joint Probability Distribution

- Example: Medical Diagnosis
  - Two Variables: *S* – Pneumonia, *W* - WBCcount

		WBCcount, W		
		high	*normal*	*low*
S	*True*	0.0008	0.0001	0.0001
	False	0.0042	0.9929	0.0019

S	***W***	***P*(*S*, *W*)**
True	*High*	0.0008
True	*Normal*	0.0001
True	*Low*	0.0001
False	*Hight*	0.0042
False	*Normal*	0.9929
False	*Low*	0.0019

Joint Probability Distribution

Example: Recommendation Letters

- Intelligence, *I*
  - *low* i^0, *high* i^1
- Difficulty, *D*
  - *easy* d^0, *hard* d^1
- Grade, *G*
  - *A* g^1, *B* g^2, *C* g^3

superscript indicates different values of a variable

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.060
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Example from: Koller, Probabilistic Graphical Models

Check on Understanding

- $P(x^1, y^2)$

- $P(x^1)$

- $P(y^1 \text{ or } x^2)$

X	Y	P
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

Check on Understanding - Solutions

X	Y	P
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

- $P(x^1, y^2) = 0.4$

- $P(x^1) = P(x^1, y^1) + P(x^1, y^2) = 0.2 + 0.4$
  $= 0.6$

- $P(y^1 \text{ or } x^2) = P(y^1) + P(x^2) - P(y^1, x^2)$
  $= (0.2 + 0.3) + (0.3 + 0.1) - (0.3)$
  $= 0.6$

Marginal Probabilities

- Given a set of random variables $X_1, X_2, \ldots, X_n$ with a joint probability $P(x_1, x_2, \ldots, x_n)$
- The marginal probability of a random variables X_i is obtained by summing over all possible values of the other random variables

$$P(X_i = x_i) = \sum_{x_1, x_2, \ldots, x_{\{i-1\}}, x_{\{i+1\}}, \ldots, x_n} P(x_1, x_2, \ldots, x_n)$$

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): process of combining collapsed rows by adding

		WBCcount, W			
		high	*normal*	*low*	P(*S*)
Pneumonia, S	*True*	0.0008	0.0001	0.0001	0.001
	False	0.0042	0.9929	0.0019	0.999
	P(*W*)	0.005	0.993	0.002	

Check on Understanding (2)

X	Y	P(X,Y)
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

$$P(x) = \sum_{y} P(x, y)$$

X	P(X)
x^1	
x^2	

$$P(y) = \sum_{x} P(x, y)$$

Y	P(Y)
y^1	
y^2	

Check on Understanding (2) - Solutions

X	Y	P(X,Y)
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

X	P(X)
x^1	0.6
x^2	0.4

Y	P(Y)
y^1	0.5
y^2	0.5

Conditional Probabilities

- Conditional probability is defined as

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \quad s.t.\ P(B) \neq 0$$

- $P(X = x \mid Y = y)$ denotes the belief that event *X*=*x* occurs given that event *Y*=y has occurred

- An alternate formulation by product rule

$$P(x,y) = P(x \mid y)P(y) = P(y \mid x)P(x)$$

Check on Understanding (3)

X	Y	P(X,Y)
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

- $P(x^1 \mid y^1)$
- $P(x^2 \mid y^1)$
- $P(y^2 \mid x^1)$

Check on Understanding (3) - Solutions

X	Y	P
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

- $P(x^1 \mid y^1) = P(x^1, y^1) / P(y^1)$
  where $P(y^1) = P(x^1, y^1) + P(x^2, y^1) = 0.5$

  $= 0.2 / 0.5 = 0.4$

- $P(x^2 \mid y^1) = P(x^2, y^1) / P(y^1)$
  $= 0.3/0.5 = 0.6$

- $P(y^2 \mid x^1) = P(x^1, y^2) / P(x^1)$
  $= 0.4 / (0.2 + 0.4) = 0.33$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

$P(W \mid T)$

$P(W \mid T=hot)$

W	P
sunny	0.8
cloudy	0.2

$P(W \mid T=cold)$

W	P
sunny	0.4
cloudy	0.6

Conditional Distribution

$P(T, W)$

T	W	P
hot	sunny	0.4
hot	cloudy	0.1
cold	sunny	0.2
cold	cloudy	0.3

Joint Distribution

Conditioning

- Observe a variable
  - Grade = g^1

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.060
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Conditioning

- Observe a variable
  - Grade = g^1

I	D	G	Prob.
i^0	d^0	g^1	0.126
~~i^0~~	~~d^0~~	~~g^2~~	~~0.168~~
~~i^0~~	~~d^0~~	~~g^3~~	~~0.126~~
i^0	d^1	g^1	0.009
~~i^0~~	~~d^1~~	~~g^2~~	~~0.045~~
~~i^0~~	~~d^1~~	~~g^3~~	~~0.126~~
i^1	d^0	g^1	0.252
~~i^1~~	~~d^0~~	~~g^2~~	~~0.0224~~
~~i^1~~	~~d^0~~	~~g^3~~	~~0.0056~~
i^1	d^1	g^1	0.060
~~i^1~~	~~d^1~~	~~g^2~~	~~0.036~~
~~i^1~~	~~d^1~~	~~g^3~~	~~0.024~~

Conditioning: Reduction

- Observe a variable
  - Grade = g^1

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.060

Conditioning: Renormalization

$P(I, D, g^1)$

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.060

0.447

Normalize to get a conditional probability distribution

$P(I, D \mid g^1)$

I	D	G	Prob.
i^0	d^0	g^1	0.282
i^0	d^1	g^1	0.020
i^1	d^0	g^1	0.564
i^1	d^1	g^1	0.134

Check on Understanding (4)

- P(X | Y=y^2)

X	Y	P
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

Check on Understanding (4)

- P(X | Y=y^2)

X	Y	P
x^1	y^1	0.2
x^1	y^2	0.4
x^2	y^1	0.3
x^2	y^2	0.1

X	Y	P(X | y^2)
x^1	y^2	0.8
x^2	y^2	0.2

Product Rule

- For some problems, given conditional distributions and want to find the joint distribution

$$P(x, y) = P(x \mid y)P(y)$$

P(*W*)

W	P
sunny	0.8
cloudy	0.2

P(*D* | *W*)

D	W	P
wet	sunny	0.1
dry	sunny	0.9
wet	cloudy	0.7
dry	cloudy	0.3

P(*D*, *W*)

D	W	P
wet	sunny	0.08
dry	sunny	0.72
wet	cloudy	0.14
dry	cloudy	0.06

Chain Rule

- In general, the joint distribution can be written as an incremental product of conditional distributions
  - Derived by successive applications of the product rule

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i \mid x_1, \dots, x_{i-1})$$

Bayes' Rule

- Two ways to factor a joint distribution over two variables

$$P(x, y) = P(x \mid y)P(y) = P(y \mid x)P(x)$$

- Dividing, to get

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

- Why is this useful?
  - Many times one conditional may be easy to estimate and the other hard
  - Can calculate one from the other

Inference with Bayes' Rule

- Bayes' rule is often used for diagnostic reasoning
- Form a hypothesis about the world based on observable variables; Bayes' rule in terms of belief of hypothesis *H* given evidence *e*

$$P(H \mid e) = \frac{P(e \mid H)P(H)}{P(e)}$$

  - $P(H \mid e)$ – posterior probability
  - $P(H)$ - prior probability
  - $P(e \mid H)$ – likelihood of the evidence
  - $P(e)$ normalizing constant $P(e) = \sum_h P(e \mid h)P(h)$

Can write as $P(H \mid e) \propto P(e \mid H)P(H)$

Bayes' Reasoning

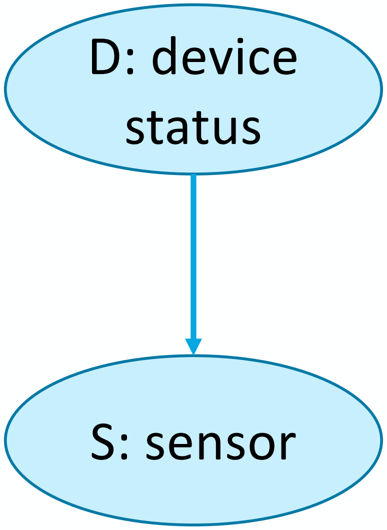
- Medical Diagnosis
  - *M*: meningitis; *S*: stiff neck
  - Given:
    - P(*s* | *m*) = 0.8
    - P(*m*) = 0.0001
    - P(*s*) = 0.1

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 * 0.0001}{0.1} = 0.0008$$

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)P(Cause)}{P(Effect)}$$

Diagnostic Inference with Bayes'

- Some equipment has a status of *normal* or *error*
- The equipment operation is sensed indirectly via a sensor: *high* or *low*



P(D: device status)

normal	error
0.9	0.1

P(S: sensor | D: device status)

Sensor	normal	error
high	0.1	0.6
low	0.9	0.4

Diagnostic Inference with Bayes'

- Diagnostic inference: compute the probability of the device operating normally or in error given a sensor reading

$$P(D \mid S = high) = \begin{pmatrix} P(D = normal \mid S = high) \\ P(D = error \mid S = high) \end{pmatrix}$$

- Use Bayes' rule to reverse conditioning variables

Independence

- Two variables *X* and *Y* are independent if

$$P(X, Y) = P(X)P(Y)$$

$$\forall x, y \; P(x, y) = P(x)P(y)$$

$X \perp Y \quad X \perp\!\!\!\perp Y$

Symbols for independence

- Can also be written as:

$$P(X \mid Y) = P(X)$$

$$P(Y \mid X) = P(Y)$$

Independence Example

- Which of the joint probability P1 or P2 illustrates independence of S and T?

P1

T	S	P
hot	sunny	0.3
hot	cloudy	0.2
cold	sunny	0.3
cold	cloudy	0.2

S	P
sunny	0.6
cloudy	0.4

T	P
hot	0.5
cold	0.5

P2

T	S	P
hot	sunny	0.4
hot	cloudy	0.1
cold	sunny	0.2
cold	cloudy	0.3

Answer: P1

Conditional Independence

- Random variables *X* and *Y* are conditionally independent given *Z* iff

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z) \qquad X \perp Y \mid Z$$

$$\forall x, y, z \;\; P(x, y \mid z) = P(x \mid z)P(y \mid z)$$

- Equivalent form:

$$P(X \mid Y, Z) = P(X \mid Z)$$

Cond. Independence vs. Independence

- Conditional independence does not imply independence
- Example
  - A = height
  - B = reading ability
  - C = age

$$P(reading\ ability \mid age, height) = P(reading\ ability \mid age)$$

$$P(height \mid reading\ ability, age) = P(height \mid age)$$

- Height and reading ability are dependent (not independent), but are conditionally independent given age