COMS W4701: Artificial Intelligence

Lecture 16: Uncertainty and Probability

Tony Dear, Ph.D.
Department of Computer Science
School of Engineering and Applied Sciences

Today

- Probability, random variables, and distributions
- Joint and conditional probabilities and distributions
- Product rule, chain rule
- Bayes' theorem, independence

Uncertainty

- So far: Planning and decision making in fully observable environments
- How do we reason in uncertain and *partially observable* environments?

- **Belief state**: A probability distribution over the entire state space
- Represent both uncertainty in the problem as well as *degree of belief*
- We can avoid the hard requirements of logic-based approaches

- Recall 90s AI resurgence relied heavily on *probabilistic approaches*
  - Diagnosis, speech and image recognition, tracking, mapping, error correction, etc.

Probabilities

- **Sample space**: Set Ω of all possible outcomes of a random experiment
- **Event**: Subset of a sample space (often described by a logical proposition)

- **Probability model (function)**: $P: \Omega \rightarrow [0,1]$ s.t. $\sum_{\omega \in \Omega} P(\omega) = 1$
- Probability of an event ϕ: $P(\phi) = \sum_{\omega \in \phi} P(\omega)$
  - Properties: $P(\emptyset) = 0$, $P(\Omega) = 1$, $P(\bar{\phi}) = 1 - P(\phi)$

- *Uniform probability model*: $P(\omega) = 1/|\Omega|\ \forall \omega$ and $P(\phi) = |\phi|/|\Omega|$
- Probabilities may represent *frequencies* or subjective *degrees of belief*

Random Variables

- A **random variable** $X: \Omega \to R$ maps sample space outcomes to some *range* R
- Ranges may be discrete/continuous, finite/infinite, ordered/unordered

- The **probability distribution** of a RV X enumerates range value probabilities
- *Categorical distributions* describe discrete and finite RVs in a table or vector
- Can use logical operators to combine different outcomes

W	P(W)
sun	0.6
rain	0.1
cloud	0.29
snow	0.01

- $P(W = sun) = P(sun) = 0.6$
- $P(sun \text{ OR } rain) = 0.6 + 0.1 = 0.7$
- $P(cloud \text{ OR } {\sim}snow)$

$$= P(cloud) + P({\sim}snow) - P(cloud\ AND\ {\sim}snow) = 0.29 + 0.99 - 0.29 = 0.99$$

Joint Probability Distributions

- **Joint distributions** enumerate probabilities of *combinations* of multiple RVs together
- Size of full categorial joint distribution $= |X_1| \times |X_2| \times \cdots \times |X_n|$

- Given a joint distribution, we can also find distributions over *subsets* of RVs
- **Marginalization**: Sum out irrelevant RVs

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

T	W	Pr(T,W)
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(w) = \sum_t P(t, w) \longrightarrow$$

W	P(W)
sun	0.6
rain	0.4

Conditional Probability Distributions

- **Conditional probability:** Probability of an event *given* that another one occurred
- Ratio between joint probability and marginal probability of known event

T	W	Pr(T,W)
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(sun|hot) = \frac{P(sun, hot)}{P(hot)} = \frac{0.4}{0.5} = \frac{4}{5}$$

$$P(sun|cold) = \frac{P(sun, cold)}{P(cold)} = \frac{0.2}{0.5} = \frac{2}{5}$$

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

- A **conditional distribution** contains the probabilities of an *unobserved* variable, all conditioned on one outcome
- Equivalent to *normalizing* all joint probabilities with the conditioned outcome values

W	P(W|hot)
sun	0.8
rain	0.2

W	P(W|cold)
sun	0.4
rain	0.6

Product Rule

- The **product rule** yields joint probability $P(x, y)$ from a marginal $P(y)$ and conditional $P(x|y)$

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

- We can also follow with marginalization to find the "other" marginal $P(x)$

$P(W)$

W	Pr
sun	0.8
rain	0.2

$P(D|W)$

D	W	Pr
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$P(D, W)$

D	W	Pr
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

$P(D)$

D	Pr
wet	0.22
dry	0.78

Chain Rule

- The product rule can be extended to more than two RVs
- Idea: Successively build up larger joint probabilities

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

$$= P(x_1,x_2)\frac{P(x_1,x_2,x_3)}{P(x_1,x_2)} = P(x_1,x_2,x_3)$$

- In general: $P(x_1,\ldots,x_n) = P(x_1)P(x_2|x_1)\cdots P(x_n|x_1,\ldots,x_{n-1})$

$$= \prod_i P(x_i|x_1,\ldots,x_{i-1})$$

Chain Rule

- The chain rule can also be applied when all probabilities are conditioned on the same observation:

$$P(x_1|x_0)P(x_2|x_1,x_0)P(x_3|x_1,x_2,x_0)$$

$$= \frac{P(x_0,x_1)}{P(x_0)}\frac{P(x_0,x_1,x_2)}{P(x_0,x_1)}\frac{P(x_0,x_1,x_2,x_3)}{P(x_0,x_1,x_2)}$$

$$= \frac{P(x_0,x_1,x_2,x_3)}{P(x_0)} = P(x_1,x_2,x_3|x_0)$$

- In general: $P(x_1,\ldots,x_n|y_1,\ldots,y_m) = \prod_i P(x_i|x_1,\ldots,x_{i-1},y_1,\ldots,y_m)$

Example: Chain Rule

- Given: $P(a) = 0.5, P(b|a) = 0.2, P(c|a,b) = 0.7$

- Product rule: $P(a,b) = P(a)P(b|a) = 0.5 \times 0.2 = 0.1$
- (Also) product rule: $P(b,c|a) = P(b|a)P(c|a,b) = 0.2 \times 0.7 = 0.14$

- Chain rule: $P(a,b,c) = P(a)P(b|a)P(c|a,b) = 0.5 \times 0.2 \times 0.7$
  $= P(a,b)P(c|a,b) = 0.1 \times 0.7$
  $= P(a)P(b,c|a) = 0.5 \times 0.14$

- What if we were given $P(c|a)$ or $P(c|b)$ instead of $P(c|a,b)$?
- Can compute $P(a,c) = P(a)P(c|a)$, but we can't do anything with $P(c|b)$!

Bayes' Theorem

- We can combine conditional probability with the product rule to express a *posterior* probability given *evidence*:

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

- $P(x)$ is the *prior* and $P(y|x)$ is the *likelihood* of the evidence

- As with chain rule, this also holds if all terms are conditioned on another variable(s) z:

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

Example: Probabilistic Inference

- Bayes' theorem can be used to *infer* hidden information given evidence

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

Binary random variables:
- M: meningitis
- S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\}$$

Known probabilities

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)}$$

$$= \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999} = 0.008$$

Much smaller than $P(+s|+m)$!

Independence

- Two variables are **independent** if we can *factor* their joint distribution
- Breaks down a large joint distribution into smaller marginal ones

$$X \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad \forall x, y:\ P(x, y) = P(x)P(y);\ \ P(x|y) = P(x)$$

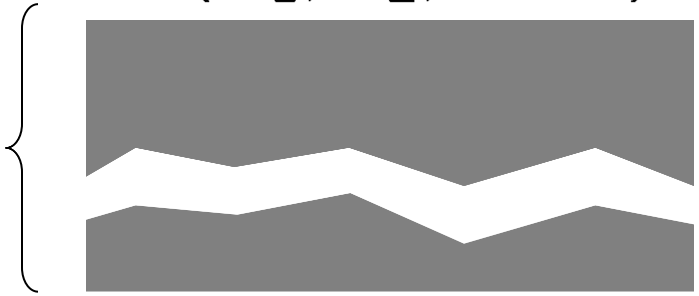
- Knowing something about X tells us nothing about Y

- This is the *only case* in which we can put together marginal distributions to reconstruct a joint distribution!
- Second identity also useful for simplifying chain rule

Example: Independence

- Suppose we have N binary RVs
- Joint distribution would have size $O(2^N)$ (rows)

$P(X_1, X_2, \ldots X_n)$



2^n

- What if we can assert independence?

- We can represent the *same information* using N 2-row tables ($O(2N)$)

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

$\ldots$

$P(X_n)$

H	0.5
T	0.5

Conditional Independence

- Absolute / marginal independence is often difficult to assert
- It is easier to assert this relationship given some *evidence*

- Two variables can be **conditionally independent** *given* a third variable:

$$X \perp\!\!\!\perp Y | Z \iff \begin{array}{l} \forall x, y, z : P(x, y|z) = P(x|z)P(y|z) \\ \forall x, y, z : P(x|z, y) = P(x|z) \end{array}$$

- Given Z, knowing something about X does not affect our belief about Y

Summary

- Probability is the language of uncertainty
- Belief states are probability distributions, usually over random variables

- Given a joint distribution, we can do find marginal and conditional probs
- For inference, use conditioning, product/chain rule, Bayes' theorem

- Independence and conditional independence assert relationships between variables, can help simplify models