COMS W4701: Artificial Intelligence

Lecture 8: Probability and Markov Chains

Tony Dear, Ph.D.

Department of Computer Science

School of Engineering and Applied Sciences

Today

- Probability, random variables, and distributions
- Joint and marginal distributions
- Conditional probabilities

- Probabilistic inference
- Product rule, conditioning chain rule
- Bayes’ theorem, independence

- Markov chains

AI Roadmap

- So far: Problem solving, decision making
- All problems have been fully observable (see states, rewards)

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

- In the real world, most situations are **partially observable**!
- Agents track their *uncertainty* using *belief states*

Uncertainty

- **Rationality** depends on both goals and degree of success

- One solution for uncertainty: Plan for *all* possible outcomes
- But we usually don't even know what outcomes are possible
  - Ex: How much do we need to know about a patient for an accurate diagnosis?

- Better way: *Summarize* uncertainty using probabilities
- Two interpretations: Problem uncertainty, *degree of belief*

- We still **maximize expected utility** (MEU) when making decisions

Random Variables

- A **random variable** $X: \Omega \to \mathbb{R}$ is a *function* that maps values in a domain Ω to a real value (a probability)
- Axioms: $\forall x \;\; P(X = x) \geq 0 \qquad \sum_x P(X = x) = 1$

- Any aspect of the world about which we are uncertain
  - R: Is it raining? (Boolean)
  - N: How many students predicted to come to class? (Nonnegative integer)
  - T: What is the temperature today? (Float, continuous)
  - L: Where is a robot on a 2D grid? (Tuples)

Probability Distributions

- Discrete RVs can be enumerated in a table (no continuous in this class)

- An **event** E is a *set* of outcomes, enumerated by logical propositions

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- $P(W = sun) = P(sun) = 0.6$
- $P(W \neq meteor) = P(\sim meteor) = 1.0$
- $P(rain \text{ OR } fog) = 0.4$

$P(W)$

W	Pr
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Joint Probability Distributions

- Probability distributions over *multiple* discrete RVs:

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

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

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

- **Joint distributions** are *Cartesian product* of RVs
- Size of table $= |X_1| \times |X_2| \times \cdots \times |X_n|$

- Events over joint distributions:
  - $P(T = hot, W = sun) = P(hot, sun) = 0.4$
  - $P(T = hot, W \neq sun) = P(hot, {\sim}sun) = 0.1$
  - $P(W = rain) = 0.4$
  - $P(T = hot \text{ OR } W = rain) = P(hot \text{ OR } rain) = 0.8$

$P(T, W)$

T	W	Pr
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Marginalization

- Given a joint distribution, we can find distributions over *subsets* of RVs
- We can sum out or **marginalize** irrelevant RVs

$$\boxed{P(Y) = \sum_z P(Y, Z = z)}$$

$P(T, W)$

T	W	Pr
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_w P(t, w)$$

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

T	Pr
hot	0.5
cold	0.5

$P(T)$

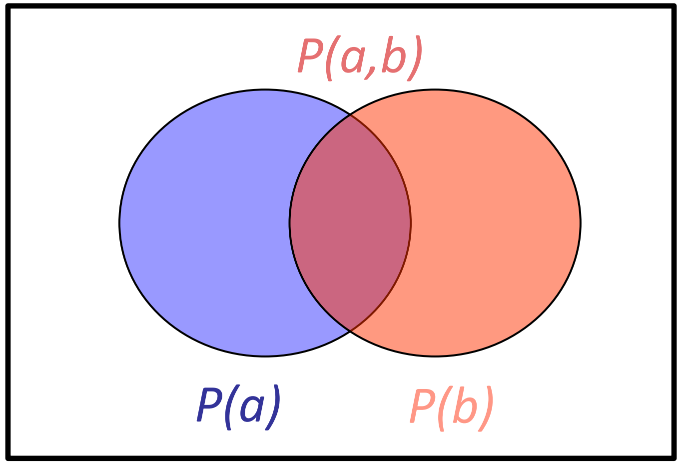
W	Pr
sun	0.6
rain	0.4

$P(W)$

Conditional Probabilities

- Marginal probabilities are at least as large as joint probabilities (why?)
- Their ratio is a ***conditional probability***
- Expresses a joint probability within a smaller space

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



$$P(T,W)$$

T	W	Pr
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(hot|sun) = \frac{P(hot,sun)}{P(sun)} = \frac{0.4}{0.6} = \frac{2}{3}$$

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

Conditional Distributions

- A **conditional distribution** describes an *unobserved* variable given an *observed* one
- Equivalent to *normalizing* joint probabilities between both variables

$P(T,W)$

T	W	Pr
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W|cold) = \frac{P(W,cold)}{P(cold)}$$

$$P(cold) = 0.5$$

$P(cold,W)$

T	W	Pr
cold	sun	0.2
cold	rain	0.3

$P(W|cold)$

W	Pr
sun	0.4
rain	0.6

$$P(T|sun) = \frac{P(T,sun)}{P(sun)}$$

$$P(sun) = 0.6$$

$P(T,sun)$

T	W	Pr
hot	sun	0.4
cold	sun	0.2

$P(T|sun)$

T	Pr
hot	0.67
cold	0.33

Conditional Distributions

- Caution: Conditioning on *unobserved* variables does not produce a distribution
- Examples: $P(x|Y)$, $P(X|Y)$
- These tables have no guarantee of summing to 1!

$P(T, W)$

T	W	Pr
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(cold|W) = \frac{P(cold,W)}{P(W)}$ means $P(cold|w) = \frac{P(cold,w)}{P(w)}$ $\forall w \in W$

$P(cold, W)$

T	W	Pr
cold	sun	0.2
cold	rain	0.3

$\div$

$P(W)$

W	Pr
sun	0.6
rain	0.4

$=$

$P(cold|W)$

W	Pr
sun	0.33
rain	0.75

Probabilistic Inference

- We often want to *infer* knowledge about hidden variables given *evidence*
- $P(\text{unobserved variables} \mid \text{observed variables})$
  - Ex: What is $P(rain|puddle)$?
- Our beliefs generally change with new evidence
  - $P(rain|puddle, cold) \neq P(rain|puddle)$
- Our models usually give us $P(\text{evidence} \mid \text{hidden})$
  - Ex: Rain generally leads to puddles (not the other way)

Product Rule

- We know how to obtain marginal and conditional distributions from joint distributions
- We can also put together a marginal and conditional to recover a joint

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

Remember: Marginal RV must be same as the "conditioned" RV

$P(W)$

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

Conditioning

- We can combine the product rule with marginalization to find marginal probabilities from conditional probabilities

$$\sum_i P(x|y_i)P(y_i) = P(x|y_1)P(y_1) + P(x|y_2)P(y_2) + \cdots + P(x|y_n)P(y_n)$$

$$= P(x,y_1) + P(x,y_2) + \cdots + P(x,y_n) = \sum_i P(x,y_i) = P(x)$$

$P(W)$

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

Example: Chain Rule

Y	Z	$Pr(Y\|Z)$
$+y$	$+z$	0.2
$+y$	$-z$	0.5
$-y$	$+z$	0.8
$-y$	$-z$	0.5

Y	Z	$Pr(+x\|Y,Z)$
$+y$	$+z$	0.7
$+y$	$-z$	0.6
$-y$	$+z$	0.4
$-y$	$-z$	0.1

$$P(+x,+y\,|+z) = P(+y|+z)P(+x|+y,+z)$$
$$= 0.2 \times 0.7 = 0.14$$

$$P(+x,-y\,|+z) = P(-y|+z)P(+x|-y,+z)$$
$$= 0.8 \times 0.4 = 0.32$$

$$P(+x\,|+z) = \sum_y P(+x,y\,|+z)$$
$$= P(+x,+y\,|+z) + P(+x,-y\,|+z) = 0.46$$

Bayes' Theorem

- Chain rule takes us from conditional + marginal to a joint probability
- We can also convert from one conditional to another

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

- This allows us to "flip" a conditional probability around
- Can be useful for *inferring* or *diagnosing* hidden info given evidence

$$P(\text{hidden} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hidden})P(\text{hidden})}{P(\text{evidence})}$$

Example: Probabilistic Inference

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

- Suppose we have two 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 $P(+s|+m)$!

Normalization

- We computed the dominator $P(+s) = \sum_m P(+s, m) = P(+s, +m) + P(+s, -m)$
- First term is the numerator of $P(+m| + s)$
- Second term is the numerator of $P(-m| + s)$

- The denominator is a **normalization constant** for the distribution $P(M| + s)$

$$P(M| + s) \propto_M P(M, +s)$$

- To find $P(M| + s)$, we can simply compute all the numerator terms of the distribution
- These give us relative likelihoods, which are sufficient in many cases
- If we want probabilities, just divide by the sum (normalize)

Independence

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

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

- Knowing something about X tells us nothing about Y

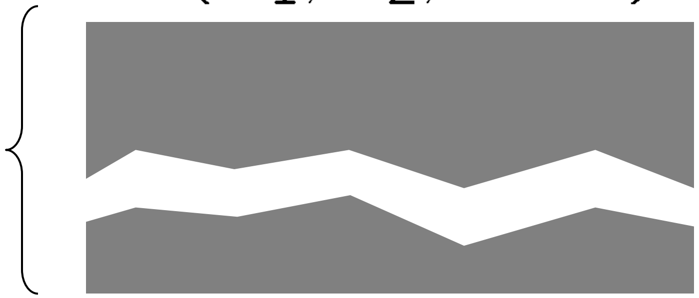
- This is the *only case* in which we can put together marginal distributions to reconstruct a joint distribution!
- Second identity also useful for simplifying chain rule

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:

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

- "Given Z, knowing something about X tells us nothing more about Y"

Example: Conditional Independence

- Fire F, smoke S, alarm A
- Fire and alarm probably aren't independent, but…
  - P(alarm | smoke) = P(alarm | smoke, fire)
  - P(fire | smoke) = P(fire | smoke, alarm)
  - P(alarm, fire | smoke) = P(alarm | smoke) P(fire | smoke)

- Caution: Independent RVs can *lose* independence conditioned on a third!
  - Ex: Sprinkler S, rain R, wet grass W
  - S and R probably independent
  - What if we know something about W?

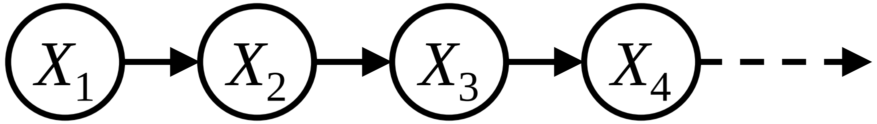
$$X \perp\!\!\!\perp Y \not\Rightarrow X \perp\!\!\!\perp Y | Z$$

Temporal and Spatial Reasoning

- MDP agents *take actions* over time/space in static, fully observable envs
- Let's now *reason* over time/space in *dynamic*, *partially observable* envs
- The world is changing around us, and we want to maintain and update **belief states** about the world
- Applications often process *sequences* of observations
  - Speech recognition, robot localization, medical monitoring, …

Markov Chains

- **States** are RVs X_t associated with a timestep t:



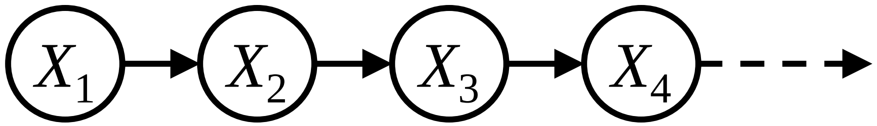
$P(X_1)$ $\quad$ $P(X_t|X_{t-1})$

- **Transition model** is given by probabilities $P(X_t \mid X_{t-1})$

- **Markov** assumption: Transitions only depend on finite previous states
  - Can be higher-order, e.g. second-order transition would be $P(X_t \mid X_{t-1}, X_{t-2})$
- **Stationarity** assumption: Transition model same for all t

Conditional Independence

- Lots of conditional independences here!



- Given the present state, past and future states are independent

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \qquad X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2$$

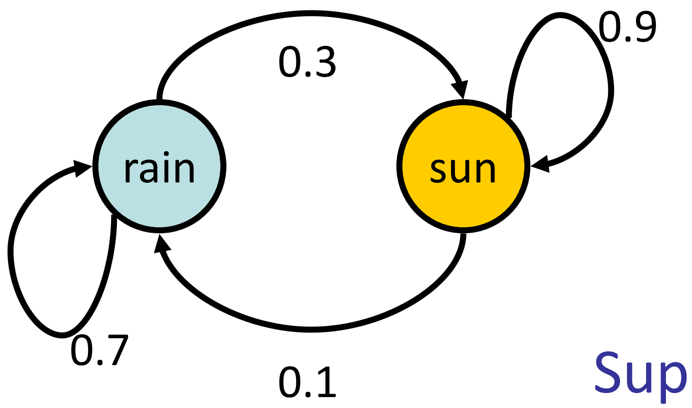
$$X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3 \qquad X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$$

- Chain rule for joint distribution can be vastly simplified!

$$\begin{aligned} P(X_1, X_2, \ldots, X_T) &= P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1}) \\ &= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1}) \end{aligned}$$

Example Markov Chain: Weather

State diagram representation:



Transition matrix representation:

$$T = \begin{array}{cc} \text{sun} & \text{rain} \\ \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \end{array} \begin{array}{l} \text{sun} \\ \text{rain} \end{array}$$

Suppose $P(X_1 = \text{sun}) = 0.8$, $P(X_1 = \text{rain}) = 0.2$

$$P(X_2 = \text{sun}) = \sum_{x_1} P(X_2 = \text{sun} | x_1) P(x_1)$$

$$= P(\text{sun} \mid \text{sun}) P(\text{sun}) + P(\text{sun} \mid \text{rain}) P(\text{rain})$$

$$= 0.9(0.8) + 0.3(0.2) = 0.78$$

$$P(X_2) = T \cdot P(X_1)$$

$$= \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}$$

$$= \begin{pmatrix} 0.78 \\ 0.22 \end{pmatrix}$$

State Evolution

- We can find the state distribution at any time t:
  - $P(X_1)$ given
  - $P(X_t) = \sum_{x_{t-1}} P(x_{t-1}, X_t) = \sum_{x_{t-1}} P(X_t | x_{t-1}) P(x_{t-1})$

- Conditional independences allow us to massively simplify the chain rule!

- If using the transition matrix: $P(X_t) = T^{t-1} P(X_1)$
- Computation complexity simply linear in t

Stationary Distributions

- Stationary Markov chains eventually "forget" the initial distribution X_1
- In the limit, we always end up with the same **stationary distribution**

$$P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X|x) P_\infty(x)$$

- Weather example:

$$\underset{P(X_1)}{\left\langle \begin{matrix} 1.0 \\ 0.0 \end{matrix} \right\rangle} \quad \underset{P(X_2)}{\left\langle \begin{matrix} 0.9 \\ 0.1 \end{matrix} \right\rangle} \quad \underset{P(X_3)}{\left\langle \begin{matrix} 0.84 \\ 0.16 \end{matrix} \right\rangle} \quad \underset{P(X_4)}{\left\langle \begin{matrix} 0.804 \\ 0.196 \end{matrix} \right\rangle} \Rightarrow \underset{P(X_\infty)}{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}$$

$$\underset{P(X_1)}{\left\langle \begin{matrix} 0.0 \\ 1.0 \end{matrix} \right\rangle} \quad \underset{P(X_2)}{\left\langle \begin{matrix} 0.3 \\ 0.7 \end{matrix} \right\rangle} \quad \underset{P(X_3)}{\left\langle \begin{matrix} 0.48 \\ 0.52 \end{matrix} \right\rangle} \quad \underset{P(X_4)}{\left\langle \begin{matrix} 0.588 \\ 0.412 \end{matrix} \right\rangle} \Rightarrow \underset{P(X_\infty)}{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}$$

Example: Finding Stationary Distributions

$$P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X|x)P_\infty(x)$$

$$T = \begin{array}{cc} & \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ & \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \begin{array}{l} \text{sun} \\ \text{rain} \end{array} \end{array}$$

$$P_\infty(sun) = P(sun|sun)P_\infty(sun) + P(sun|rain)P_\infty(rain)$$
$$P_\infty(rain) = P(rain|sun)P_\infty(sun) + P(rain|rain)P_\infty(rain)$$

$$P_\infty(sun) = 0.9P_\infty(sun) + 0.3P_\infty(rain)$$
$$P_\infty(rain) = 0.1P_\infty(sun) + 0.7P_\infty(rain)$$
$$P_\infty(sun) + P_\infty(rain) = 1$$

$$P_\infty(sun) = 3/4$$
$$P_\infty(rain) = 1/4$$

*$P_\infty(X)$ is the unit eigenvector of T corresponding to eigenvalue 1

Summary

- Probability is the language of uncertainty
- An agent’s belief states are represented by random variables
- Tools and concepts: Joint / conditional distributions, product / chain rule, Bayes’ rule, (conditional) independence
- One way to reason spatio-temporal domains: Markov chains