MONASH University
Information Technology

MONASH University
Information Technology

FIT5047: Intelligent Systems

Probability
Chapter 13

Some slides are adapted from Stuart Russell, Andrew Moore or Dan Klein

So far

Agents did not consider:

- **Uncertainty about the world or the outcome of an action**
- **Learning their knowledge**

From Now On

- **Uncertainty**
  - Probability, Bayesian Networks
- **Machine Learning**
  - Classification, Regression
  - Clustering

Outline

- **Background:**
  - Random variables and probabilistic inference
  - Probabilistic models
  - Joint, marginal and conditional distributions
- **Inference by enumeration**
- **Product Rule, Chain Rule, Bayes’ Rule**
- **Independence and conditional independence**

Reasoning under Uncertainty

- **Uncertainty – the quality or state of being not clearly known**
  - distinguishes *deductive* knowledge from *inductive* belief
- **Sources of uncertainty**
  - Ignorance
  - Complexity
  - Physical randomness
  - Vagueness

Probability Calculus (I)

- **Classic approach to reasoning under uncertainty (origin: Pascal and Fermat)**
- **Definitions:**
  - **Experiment** – produces one of several possible outcomes
  - **Sample space** – the set of **all** possible outcomes
  - **Event** – a subset of the sample space
  - **Random variable** – a variable whose value is determined by the **outcome of an experiment**
  - **Probability function –** a function that assigns a probability to every possible outcome of an experiment
    ➔ Given a probability function we can define a probability for each value of a random variable

Random Variables

- **A random variable represents some aspect of the world about which we may have uncertainty**
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?
- **We denote random variables with capital letters**
- **Random variables have domains**
  - R in {true, false} (sometimes write as {+r, $\neg$r})
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}

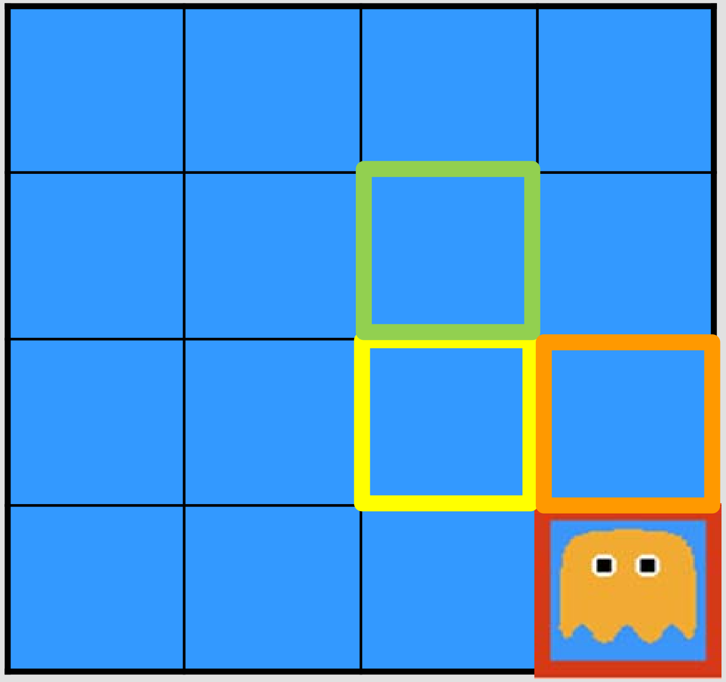
Probabilistic Inference

- **Probabilistic inference: compute a desired probability from other known probabilities**
- **We generally compute conditional probabilities**
  - They represent an agent's *beliefs* given the evidence
  - E.g., Pr(on time | no reported accidents) = 0.90
- **Probabilities change with new evidence:**
  - Observing new evidence causes *beliefs to be updated*
  - E.g., Pr(on time | no accidents, 5 a.m.) = 0.95

    Pr(on time | no accidents, 5 a.m., raining) = 0.80

Example – Inference in Ghostbusters

- **A ghost is somewhere in the grid**
- **Sensor readings tell how close a tile is to the ghost**
  - On the ghost: **red**
  - 1 away: **orange**
  - 2 away: **yellow**
  - 3+ away: **green**
- **Sensors are noisy, but we know Pr(Color|Distance)**

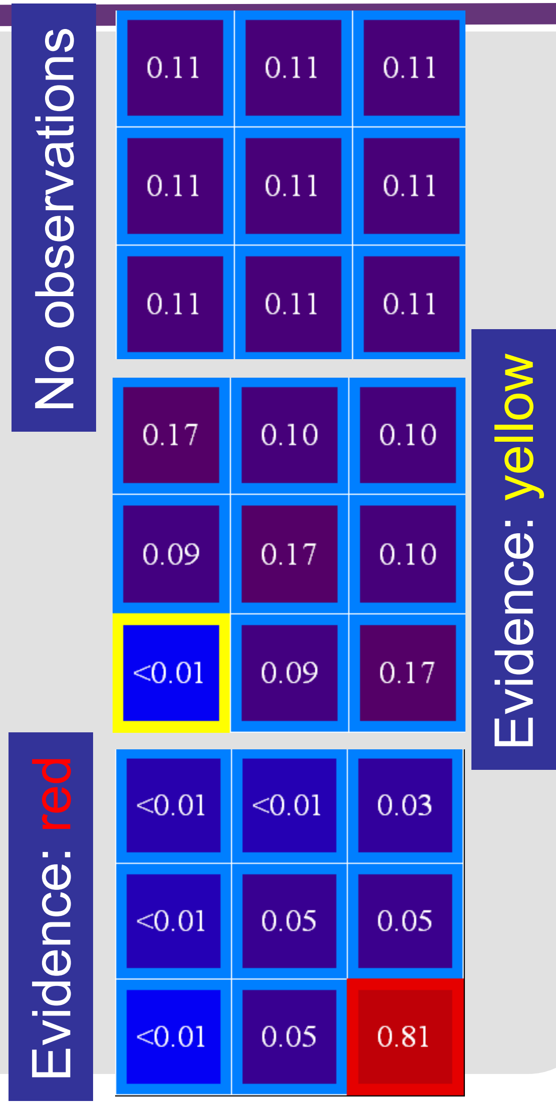


Pr(red|2)	Pr(orange|2)	Pr(yellow|2)	Pr(green|2)	TOTAL
0.05	0.17	0.46	0.32	1

We want to know: Pr(Location | Color)

Uncertainty and Probabilistic Inference

- **General situation:**
  - **Evidence**: Agent knows certain things about the state of the world
  - **Hidden variables**: Agent needs to reason about other aspects
  - **Model**: Agent knows something about how the known variables relate to the unknown variables
- **Probabilistic reasoning gives us a framework for managing our beliefs and knowledge**



MONASH University
Information Technology

Probabilistic Models (I)

- **Probabilistic models describe how (a portion of) the world works**
- **Models are always simplifications**
  - May not account for every variable
  - May not account for all interactions between variables
  - "*All models are wrong; but some are useful*."
    - George E. P. Box
- **What do we do with probabilistic models?**
  - We (or our agents) need to reason about unknown variables given evidence
    - > explanation (diagnostic reasoning)
    - > prediction (causal reasoning)
    - > value of information

Probability Distributions

- **Unobserved random variables have distributions that represent probabilities of value assignments**

Pr(Temp)

Temp	**Pr**
warm	0.5
cold	0.5

Pr(Weather)

Weather	**Pr**
sunny	0.6
rain	0.1
fog	0.3

- **A probability is a single number**

$\mathrm{Pr}(\mathrm{Weather}{=}\mathrm{rain}) = 0.1$ or $\mathrm{Pr}(\mathrm{rain}) = 0.1$

Probability Calculus (II)

Kolmogorov's axioms for finite discrete random variables – where $e_1, \ldots, e_n$ are the possible distinct values of random variable E

$$\Pr(e_i) \geq 0 \quad \forall i = 1, \ldots, n$$

$$\Pr(e_i) \leq 1 \quad \forall i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \Pr(e_i) = 1$$

$$\forall e_i, e_j \subseteq E$$

$$\text{if } e_i \cap e_j = \varnothing \text{ then } \Pr(e_i \vee e_j) = \Pr(e_i) + \Pr(e_j)$$

Joint Distributions

- **A *joint distribution* over a set of random variables $X_1, \ldots, X_n$ specifies a real number for each value assignment (or *outcome*):**

  $\mathbf{Pr}(X_1 = x_1, \ldots, X_n = x_n)$ **or** $\mathbf{Pr}(x_1, \ldots, x_n)$

  - Size of distribution of n variables with domain sizes d?

- **Must obey:**

$$\forall x_i \quad \Pr(x_1, \ldots, x_n) \geq 0$$

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

$\mathbf{Pr}(W, T)$

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

- **For all but small distributions, impractical to write out**

Probabilistic Models (II)

- **A probabilistic model is a joint distribution over a set of variables**

$$\Pr(X_1, X_2, \dots, X_n)$$

- **Given a joint distribution, we can reason about unobserved variables given evidence**
- **General form of a query:**

$$\Pr(X_q | e_1, \dots, e_k)$$

Stuff you care about

Stuff you already know

- **This kind of *posterior distribution* is also called the *belief function* of an agent who uses this model**

Events in a Joint Distribution

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

- **From a joint distribution, we can calculate the probability of any event**
  - Probability that it is hot AND sunny
  - Probability that it is hot
  - Probability that it is hot OR sunny
- **Typically, the events we care about are *partial assignments*, like Pr(T=hot)**

$\mathbf{Pr}(\boldsymbol{W}, \boldsymbol{T})$

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

Marginal Distributions

- ***Marginal distributions*** **are sub-tables that eliminate variables**
- ***Marginalization*** **(summing out): Combine collapsed rows by adding**

$$\mathbf{Pr}(\boldsymbol{W},\boldsymbol{T})$$

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

$$\Pr(t) = \sum_{w\in\{sun,rain\}} \Pr(t,w)$$

$$\Pr(T)$$

T	Pr
hot	0.5
cold	0.5

$$\Pr(w) = \sum_{t\in\{hot,cold\}} \Pr(t,w)$$

$$\Pr(W)$$

W	Pr
sun	0.6
rain	0.4

Conditional Distributions (I)

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

Joint Distribution

$$\mathbf{Pr}(\boldsymbol{W}, \boldsymbol{T})$$

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

Conditional Distributions

$$\mathbf{Pr}(\boldsymbol{W}|\boldsymbol{T})$$

$$\mathbf{Pr}(\boldsymbol{W}|\boldsymbol{T} = \boldsymbol{hot})$$

W	Pr
sun	0.8
rain	0.2

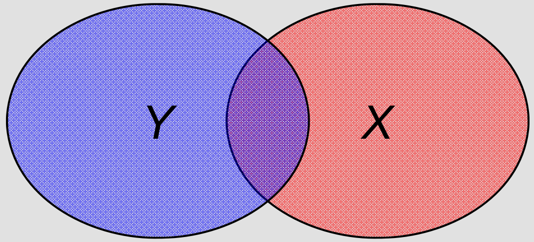
$$\mathbf{Pr}(\boldsymbol{W}|\boldsymbol{T} = \boldsymbol{cold})$$

W	Pr
sun	0.4
rain	0.6

Conditional Distributions (II)

$$\Pr(X \mid Y) = \frac{\Pr(X \wedge Y)}{\Pr(Y)}$$

$\Pr(X \cap Y)$



$\Pr(Y)$ $\Pr(X)$

$\mathbf{Pr}(\boldsymbol{W}, \boldsymbol{T})$

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

$$\Pr(W = rain | T = cold) = ?$$

Conditional Distributions (III)

- ***Conditional or posterior probabilities:***
  - E.g., Pr(*cavity* | *toothache*)=0.8, given that *toothache* is all I know
- **Notation for conditional distributions:**
  - Pr(*cavity* | *toothache*) = a single number
  - Pr(Cavity, Toothache) = 2x2 table sums to 1
  - Pr(Cavity | Toothache) = Two 2-element vectors, each sums to 1
- **If we know more:**
  - Pr(*cavity* | *toothache*, *catch*) = 0.9
  - Pr(*cavity* | *toothache*, *cavity*) = 1
- **Less specific beliefs remain *valid* after more evidence arrives, but are not always *useful***
- **New evidence may be irrelevant, allowing simplification:**
  - Pr(*cavity* | *toothache, traffic*) = Pr(*cavity* | *toothache*) = 0.8

Normalization Trick

- **A trick to get a whole conditional distribution at once:**
  - Select the joint probabilities matching the evidence
  - ***Normalize*** the selection (make it sum to 1)

$\mathbf{Pr}(\boldsymbol{W}, \boldsymbol{T})$

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

$\xrightarrow{\text{Select}}$

$\Pr(T, rain)$

T	R	Pr
hot	rain	0.1
cold	rain	0.3

$\xrightarrow{\text{Normalize}}$

$\Pr(T \mid rain)$

T	Pr
hot	0.25
cold	0.75

- **Why does this work?**

$$\Pr(x_1|x_2) = \frac{\Pr(x_1, x_2)}{\Pr(x_2)} = \frac{\Pr(x_1, x_2)}{\sum_{x_1} \Pr(x_1, x_2)}$$

Inference by Enumeration (I)

- **Pr(sun)?**

- **Pr(sun | summer)?**

- **Pr(sun | winter, hot)?**

S	T	W	Pr
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration (II)

- **General case:**
  - Evidence variables: $E_1, \dots, E_k = e_1, \dots, e_k$
  - Query variable(s): Q
  - Unknown variables: $U_1, \dots, U_r$

  $\left.\right\}$ $X_1, \dots, X_n$ *All variables*
- **We want** $\Pr(\boldsymbol{Q} | \boldsymbol{e_1}, \dots, \boldsymbol{e_k})$
- **Procedure**
  1. Select the entries that are consistent with the evidence
  2. Sum out *U* to get the joint probability of Query and Evidence:
     $$\Pr(Q, e_1, \dots, e_k) = \sum_{u_1,\dots,u_r} \Pr\underbrace{(Q, u_1, \dots, u_r, e_1, \dots, e_k)}_{X_1, \dots, X_n}$$
  3. Normalize the remaining entries to conditionalize
- **Problems:**
  - Worst-case time complexity O(d^n)
  - Space complexity O(d^n) to store the joint distribution

Inference by Enumeration – Example

- **Pr(sun | summer)**
  - Evidence variables?
  - Query variables?
  - Unknown variables?
- **Procedure**
  1. Select entries consistent with the evidence
  2. Sum out *U* to get a joint probability of *Q* and *E*
  3. Normalize the remaining entries to conditionalize

S	T	W	Pr
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

summer, sun

summer, rain

The Product Rule

- **Sometimes we have conditional distributions but want the joint distribution**

$$\Pr(x|y) = \frac{\Pr(x,y)}{\Pr(y)} \quad \Longleftrightarrow \quad \Pr(x,y) = \Pr(x|y)\Pr(y)$$

- **Example:**

$\mathbf{Pr(\mathit{W})}$

W	Pr
sun	0.8
rain	0.2

$\mathbf{Pr(\mathit{T}|\mathit{W})}$

T	W	Pr
cold	sun	0.1
hot	sun	0.9
cold	rain	0.7
hot	rain	0.3

$\Longleftrightarrow$

$\mathbf{Pr(\mathit{T},\mathit{W})}$

T	W	Pr
cold	sun	0.08
hot	sun	0.72
cold	rain	0.14
hot	rain	0.06

The Chain Rule

- **We can always write a joint distribution as an incremental product of conditional distributions**

$$\Pr(x_1, \dots, x_n) = \prod_{i=1}^{n} \Pr(x_i | x_1, \dots, x_{i-1})$$

- **Example:**
  Pr(Traffic,Umbrella,Rain)=
  Pr(Umbrella|Rain,Traffic) x Pr(Traffic|Rain) x Pr(Rain)
- **Why is this true?**

$$\begin{aligned}\Pr(x_1, \dots, x_n) &= \Pr(x_n | x_1, \dots, x_{n-1}) \Pr(x_1, \dots, x_{n-1}) \\ &= \Pr(x_n | x_1, \dots, x_{n-1}) \Pr(x_{n-1} | x_1, \dots, x_{n-2}) \Pr(x_1, \dots, x_{n-2})\end{aligned}$$

Bayes Rule

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

$$\Pr(x, y) = \Pr(x|y)\Pr(y) = \Pr(y|x)\Pr(x)$$

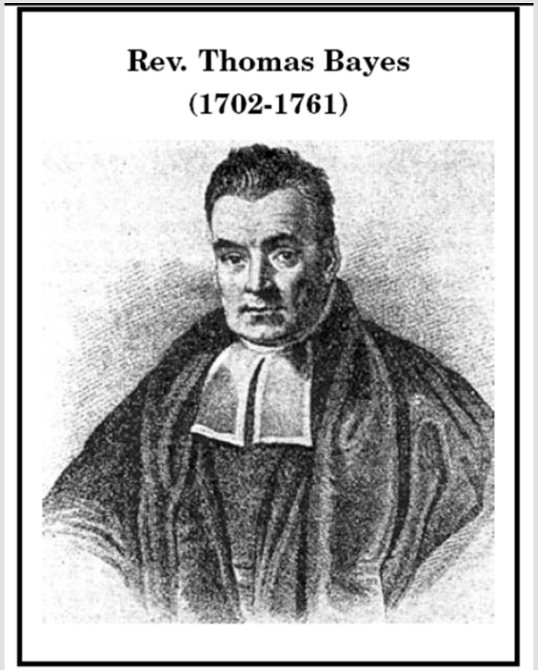
$$\Pr(x|y) = \frac{\Pr(y|x)\Pr(x)}{\Pr(y)}$$

- **Why is this helpful?**
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems (e.g., ASR, MT)

Bayes Rule: Conditionalization

- **Attributed to Rev. Thomas Bayes**

$$\Pr(h \mid e) = \frac{\Pr(e \mid h)\Pr(h)}{\Pr(e)}$$



- **Also called *Conditionalization*:**

$$\Pr'(h) = \Pr(h \mid e)$$

- **Also read as**

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Prob of evidence}}$$

- **Assumptions:**
  - Joint priors over $\{h_i\}$ and e exist
  - Total evidence: e is observed

Inference with Bayes Rule – Example

Diagnosis of breast cancer (hypothesis), given xray (evidence)

- **Let $\Pr(h)=0.01$, $\Pr(e|h)=0.8$ and $\Pr(e|{\sim}h)=0.1$**
- **Bayes theorem yields**

$$\Pr(h \mid e) = \frac{\Pr(e \mid h)\Pr(h)}{\Pr(e)}$$

$$= \frac{\Pr(e \mid h)\Pr(h)}{\Pr(e \mid h)\Pr(h) + \Pr(e \mid \sim h)\Pr(\sim h)}$$

$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99}$$

$$= \frac{0.008}{0.008 + 0.099} = \frac{0.008}{0.107} \approx 0.075$$

Ghostbusters Revisited (I)

- **We have two distributions:**
  - **Prior distribution** over ghost location: $\Pr(L)$
  - **Sensor model**: $\Pr(\underbrace{R}_{\text{reading}} \mid \underbrace{D}_{\text{distance}})$
    - Given by some "black box" process
    - Assume reading is at the lower left corner
    - E.g., $\Pr(yellow|D \geq 3) = 0.27$

      $\Pr(yellow|D = 2) = 0.46$

      $\Pr(yellow|D = 1) = 0.25$

      $\Pr(yellow|D = 0) = 0.03$

Should these probabilities sum to 1?

	1	2	3
1	0.11	0.11	0.11
2	0.11	0.11	0.11
3	0.11	0.11	0.11

0.17	0.10	0.10
0.09	0.17	0.10
<0.01	0.09	0.17

- **The posterior distribution Pr(L|R) over ghost locations given a reading**

$$\Pr(l = (3,1)|yellow)$$
$$\propto \Pr(yellow|l = (3,1))\Pr(l = (3,1))$$
$$\propto 0.03 * 0.11 = 0.0033$$

MONASH University
Information Technology

Ghostbusters Revisited (II)

- **The posterior distribution Pr(L|R) over ghost locations given a reading**

$$\Pr(l = (3,1)|yellow)$$
$$= \alpha \Pr(yellow|l = (3,1)) \Pr\big(l = (3,1)\big)$$
$$= \alpha 0.03 * 0.11 = \alpha 0.0033$$
$$\Pr(l = (2,1)|yellow) = \Pr(l = (3,2)|yellow)$$
$$= \alpha \Pr(yellow|l = (2,1)) \Pr\big(l = (2,1)\big)$$
$$= \alpha 0.25 * 0.11 = \alpha 0.0275$$
$\Pr(l = (i,i)|yellow)$ for $i=1,2,3$
$$= \alpha \Pr(yellow|l = (i,i)) \Pr\big(l = (i,i)\big)$$
$$= \alpha 0.46 * 0.11 = \alpha 0.0506$$
$$\Pr(l = (1,2)|yellow) = \Pr(l = (1,3)|yellow)$$
$$= \Pr(l = (2,3)|yellow)$$
$$= \alpha \Pr(yellow|l = (1,2)) \Pr\big(l = (1,2)\big)$$
$$= \alpha 0.27 * 0.11 = \alpha 0.0297$$
$$\alpha\,(0.0033 + 0.0275*2 + 0.0506*3 + 0.0297*3) = 1$$
$$\boldsymbol{\alpha = 1/0.2992 = 3.342}$$

	1	2	3
1	0.11	0.11	0.11
2	0.11	0.11	0.11
3	0.11	0.11	0.11

0.17	0.10	0.10
0.09	0.17	0.10
<0.01	0.09	0.17

Example Problems

- Suppose a murder occurs in a town of population 10,000 (10,001 before the murder). A suspect is brought in and DNA tested. The probability that there is a DNA match given that a person is innocent is 1/100,000; the probability of a match on a guilty person is 1. What is the probability he is guilty given a DNA match?
- Doctors have found that people with Creutzfeldt–Jakob disease (CJ) almost invariably ate lots of hamburgers, thus Pr(HamburgerEater|CJ) = 0.9. CJ is a rare disease: about 1 in 100,000 people get it. Eating hamburgers is widespread: Pr(HamburgerEater) = 0.5. What is the probability that a regular hamburger eater will have CJ disease?

Independence

- **Two variables are *independent* if:**

$$\Pr(X, Y) = \Pr(X)\Pr(Y)$$

$$\forall x,y \ \ \Pr(x,y) = \Pr(x)\Pr(y) \ \ \text{or} \ \ \Pr(x|y)=\Pr(x)$$

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

- **Independence is a simplifying *modeling assumption***
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?

Independence – Example

- **N fair, independent coin flips:**

$\Pr(X_1)$

H	0.5
T	0.5

$\Pr(X_2)$

H	0.5
T	0.5

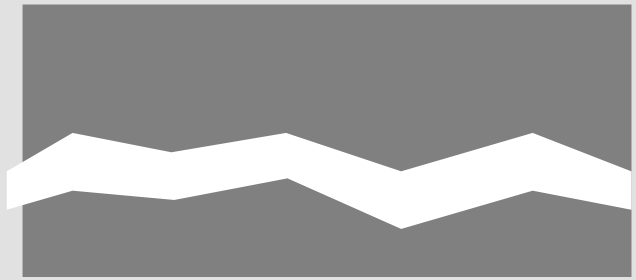
…

$\Pr(X_n)$

H	0.5
T	0.5

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

2^n



Which Variables are Independent?

$\mathbf{Pr}(\boldsymbol{T})$

T	Pr
warm	0.5
cold	0.5

$\mathbf{Pr}_1(\boldsymbol{T}, \boldsymbol{W})$

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

$\mathbf{Pr}_2(\boldsymbol{T}, \boldsymbol{W})$

T	W	Pr
warm	sun	0.3
warm	rain	0.2
cold	sun	0.3
cold	rain	0.2

$\mathbf{Pr}(\boldsymbol{W})$

W	Pr
sun	0.6
rain	0.4

Conditional Independence (I)

- **Employs domain knowledge to simplify probabilistic models**
- **Example: Pr(Toothache,Cavity,Catch)**
  If I know whether I have a cavity, the probability that the probe catches in the tooth doesn't depend on whether I have a toothache:
  – Pr(+catch | +toothache, +cavity) = Pr(+catch | +cavity)
  – Pr(+catch | +toothache, ¬cavity) = Pr(+catch| ¬cavity)
  ➔ Catch is ***conditionally independent*** of Toothache given Cavity

  Pr(Catch | Toothache, Cavity) = Pr(Catch | Cavity)

  > Pr(Toothache | Catch, Cavity) = Pr(Toothache | Cavity) or

  > Pr(Toothache, Catch | Cavity) =
  Pr(Toothache | Cavity) x Pr(Catch | Cavity)

Conditional Independence (II)

- **Unconditional (absolute) independence is rare**
- **Conditional independence is our most basic and robust form of knowledge about uncertain environments:**

$$\forall x, y, z \qquad \Pr(x, y|z) = \Pr(x|z)\Pr(y|z) \text{ or}$$
$$\Pr(x|y, z) = \Pr(x|z)$$
$$\Pr(X, Y|Z) = \Pr(X|Z)\Pr(Y|Z)$$
$$\Pr(X|Y, Z) = \Pr(X|Z)$$
$$X \perp\!\!\!\perp Y \mid Z$$

- **Example**

  Pr(Traffic|Umbrella,Rain)=Pr(Traffic|Rain) or

  Pr(Traffic,Umbrella|Rain)= Pr(Umbrella|Rain) x Pr(Traffic|Rain)

- **Bayesian networks / graphical models help us express conditional independence assumptions**

Reading

- **Russell, S. and Norvig, P. (2010), *Artificial Intelligence – A Modern Approach* (3nd ed), Prentice Hall**
  - Chapter 13

Next Lecture Topic

- **Lecture Topic 6**
  - Bayesian Networks