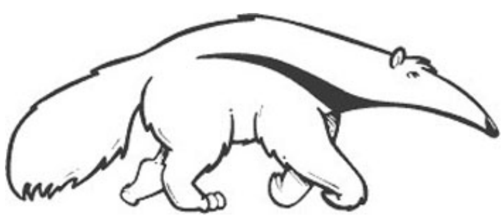
Probability

Introduction to Graphical Models

Prof. Alexander Ihler



BREN:ICS
INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

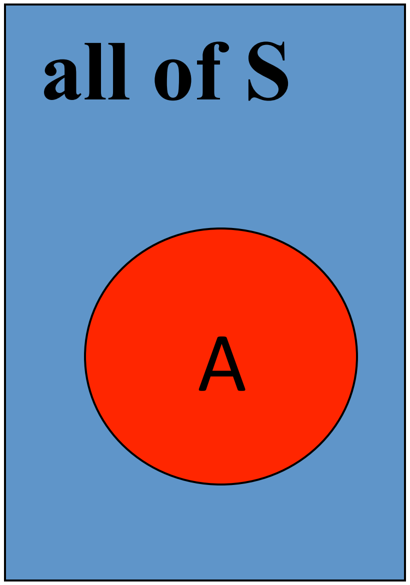
Uncertainty in the world

- Uncertainty due to
  - Randomness
  - Overwhelming complexity
  - Lack of knowledge
  - …
- Example: time to the airport
- Without representing & communicating uncertainty, it's easy to make and compound mistakes
- Probability gives
  - natural way to describe our assumptions
  - rules for how to combine information

Probability

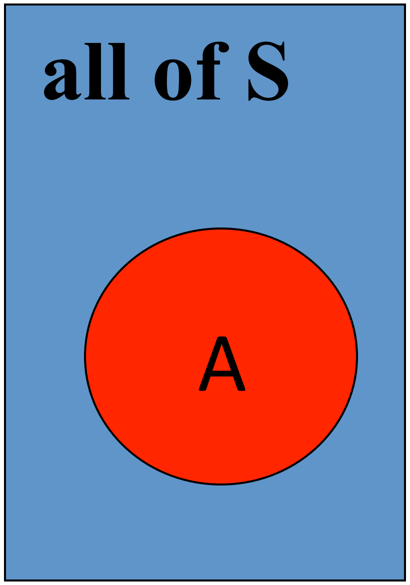
- Event "A" in event space "S"
  - Ex: "I have a headache"
  - Ex: "I have the flu"
  - Ex: "I have Ebola"

- Probability Pr[A]
  - Think of e.g. "# of worlds in which A happens"
  - This is a measure, like area
  - Can think of it in those terms



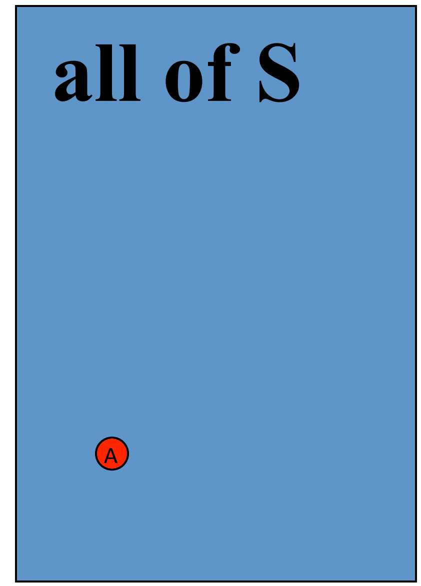
Probability

- Event "A" in event space "S"
- Probability Pr[A]
- Axioms of probability
  - $0 \leq \Pr[A] \leq 1$
  - $\Pr[\, S \,] = 1$
  - $\Pr\,[\, \emptyset \,] = 0$
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



Probability

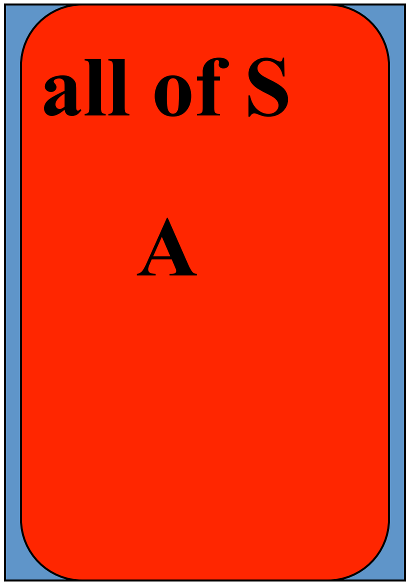
- Event "A" in event space "S"
- Probability Pr[A]
- Axioms of probability
  - $0 \leq \Pr[A] \leq 1$
  - $\Pr[\, S \,] = 1$
  - $\Pr[\, \emptyset \,] = 0$
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



"A" can't get any smaller than size zero…
No worlds in which "A" is true

Probability

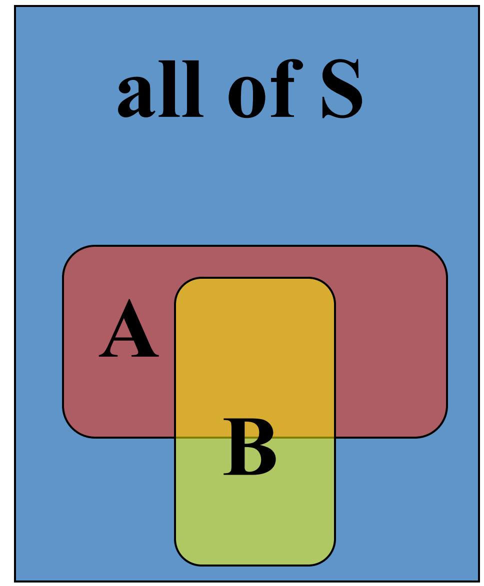
- Event “A” in event space “S”
- Probability Pr[A]
- Axioms of probability
  - $0 \leq Pr[A] \leq 1$
  - $Pr[\ S\ ] = 1$
  - $Pr\ [\ \emptyset\ ] = 0$
  - $Pr[A \cup B] = Pr[A] + Pr[B] – Pr[A \cap B]$

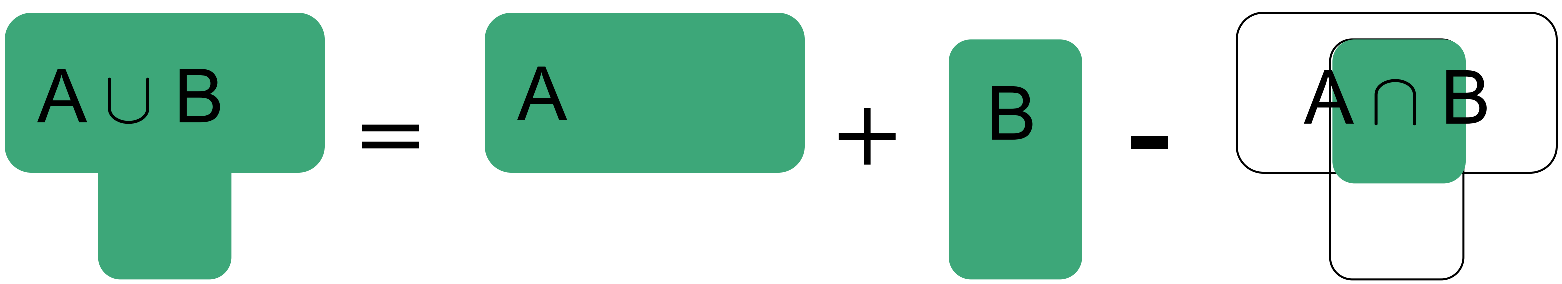


“A” can’t get any larger than all worlds: 100% of worlds have “A” true

Probability

- Event "A" in event space "S"
- Probability Pr[A]
- Axioms of probability
  - $0 \leq \Pr[A] \leq 1$
  - Pr[S] = 1
  - Pr [$\emptyset$] =0
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$

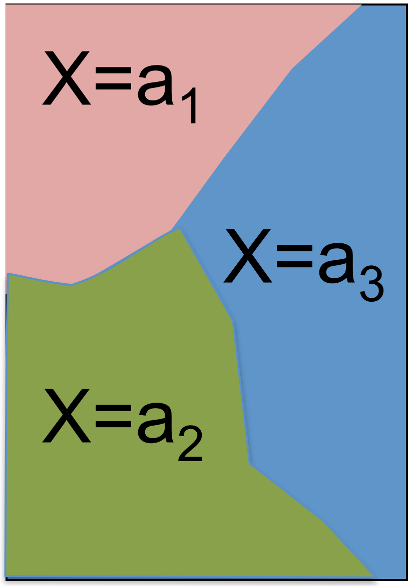




Discrete random variables

- X takes on finite set of values $S=\{a_1 \ldots a_d\}$
  - *Disjoint* and *Exhaustive*
- Probability mass functions (pmfs)
  - Define a measure on subsets of S
- $\Pr[X=a_i]$ defined for each value a_i

$$\Pr[X \in A \subseteq S] = \sum_{a_i \in A} Pr[X = a_i]$$



- Constraints:

$$0 \leq \Pr[X = a_i] \leq 1 \qquad \sum_i \Pr[X = a_i] = 1$$

Examples

- Bernoulli RV (coin toss)
  - $X \in \{0,1\}$ $Pr[X=1] = p$ $Pr[X=0] = 1-p$
- Binomial (p,n) – toss the coin n times
  - $Y = \sum X_i$ is binomial
- Discrete(d) – die roll
  - $X \in \{1 \ldots d\}$ $Pr[X=1 \ldots X=d] = [p_1 \ldots p_d]$
  - Multinomial(d,n): roll the die n times

Joint distributions

- Often, we want to reason about multiple variables
- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity
- Joint distribution
  - Assigns each event (T=t, D=d, C=c) a probability
  - Probabilities sum to 1.0
- Law of total probability:

$$p(C=1) = \sum_{t,d} P(T=t, D=d, C=1)$$

= 0.008 + 0.072 + 0.012 + 0.108 = 0.20

  - *Some* value of (T,D) must occur; values disjoint
  - "Marginal probability" of C; "marginalize" or "sum over" T,D

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Conditional probability

- Chain rule:

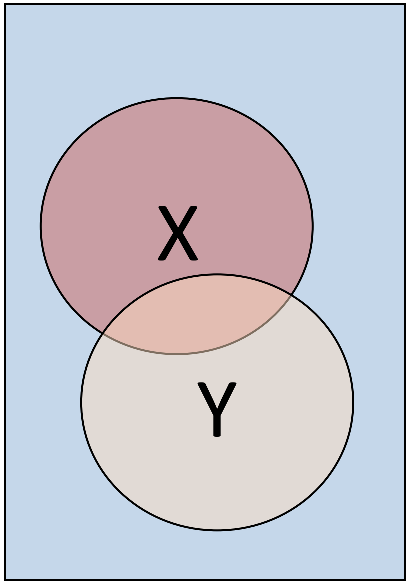
$$p(X=x, Y=y) = p(X=x)p(Y=y|X=x)$$

  - p(X=x,Y=y) : probability that both X=x and Y=y
  - p(X=x) : probability that X=x (and some Y)
  - P(Y=y|X=x): probability that Y=y given X=x already
  - If p(X) > 0 : $p(Y|X) = \frac{p(X,Y)}{p(X)}$

- More generally:

$$p(X,Y,Z) = p(X)\ p(Y|X)\ p(Z|X,Y)$$

$$p(W,X,Y,Z) = p(X)\ p(Y|X)\ p(Z|X,Y)\ p(W|X,Y,Z)$$



Example from Russell & Norvig

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Recall p(C=1) = 0.20
- Suppose we observe D=0, T=0?

$$p(C=1|D=0,T=0) = \frac{p(C=1,D=0,T=0)}{p(D=0,T=0)}$$

$$= \frac{0.008}{0.576 + 0.008} = 0.012$$

- Observe D=1, T=1?

$$= \frac{0.108}{0.016 + 0.108} = 0.871$$

Called *posterior probabilities*

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Combining these rules:

$$p(C=1|T=1) = \frac{p(C=1,T=1)}{p(T=1)}$$

$$= \frac{0.012 + 0.108}{0.064 + 0.012 + 0.016 + 0.108} = 0.60$$

$p(T=1) =$ 0.20

Called the *probability of evidence*

Computing posteriors

- Sometimes easiest to normalize last

$$p(C|T=1) \;=\; \frac{1}{p(T=1)}\, p(C,T=1) \;\propto\; p(C,T=1) \;=\; \sum_d p(C,d,T=1)$$

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Assign T=1

D	C	F(D,C)
0	0	0.064
0	1	0.012
1	0	0.016
1	1	0.108

Sum over D

C	G(C)
0	0.08
1	0.120

Normalize

C	P(C|T=1)
0	0.40
1	0.60

```
P = gm.Factor( [T,D,C] )
P[0,0,0] = 0.576
...   # define joint distribution
```

```
F = P.condition( [T] , [1] )     # assign T=1
G = P.sum( [D] )                 # sum over D
H = G / G.sum()                  # normalize
```

Bayes rule

- Lets us calculate posterior given evidence

$$p(Y|X)\ p(X) = p(X,Y) = p(X|Y)\ p(Y)$$

$$\Rightarrow \quad p(Y|X) = \frac{p(X|Y)\ p(Y)}{p(X)}$$

"Bayes rule"

- Example: flu
  - P(F), P(H|F)
  - P(F=1 | H=1) = ?

F	P(F)
0	0.95
1	0.05

F	H	P(H|F)
0	0	0.80
0	1	0.20
1	0	0.50
1	1	0.50

$$= \frac{0.50 * 0.05}{0.50 * 0.05 + 0.20 * 0.95} = 0.116$$

Independence

- X, Y independent:
  - $p(X=x,Y=y) = p(X=x)\, p(Y=y)$ for all x,y
  - Shorthand: $p(X,Y) = P(X)\, P(Y)$
  - Equivalent: $p(X|Y) = p(X)$ or $p(Y|X) = p(Y)$ (if $p(Y), p(X) > 0$)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

This reduces representation size!

Joint:

A	B	C	P(A,B,C)
0	0	0	.4 * .7 * .1
0	0	1	.4 * .7 * .9
0	1	0	.4 * .3 * .1
0	1	1	…
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Independence

- X, Y independent:
  - $p(X=x,Y=y) = p(X=x)\ p(Y=y)$ for all x,y
  - Shorthand: $p(X,Y) = P(X)\ P(Y)$
  - Equivalent: $p(X|Y) = p(X)$ or $p(Y|X) = p(Y)$ (if $p(Y), p(X) > 0$)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

Joint:

A	B	C	P(A,B,C)
0	0	0	0.028
0	0	1	0.252
0	1	0	0.012
0	1	1	0.108
1	0	0	0.042
1	0	1	0.378
1	1	0	0.018
1	1	1	0.162

This reduces representation size!

Note: it is hard to "read" independence from the joint distribution.
We can "test" for it, however.

Conditional Independence

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\, p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$ (if all > 0)
  - Intuition: X has no additional info about Y beyond Z's

- Example

X = height p(height|reading, age) = p(height|age)

Y = reading ability p(reading|height, age) = p(reading|age)

Z = age

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

Conditional Independence

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\ p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$
  - Intuition: X has no additional info about Y beyond Z's

- Example: Dentist

Again, hard to "read" from the joint probabilities; only from the conditional probabilities.

Like independence, reduces representation size!

Joint prob:

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Conditional prob:

T	D	C	P(T|D,C)
0	0	0	0.90
0	0	1	0.40
0	1	0	0.90
0	1	1	0.40
1	0	0	0.10
1	0	1	0.60
1	1	0	0.10
1	1	1	0.60

Entropy and Information

- "Entropy" is a measure of randomness
  - How hard is it to communicate a result to you?
  - Depends on the probability of the outcomes

- Communicating fair coin tosses
  - Output: H H T H T T T H H H H T …
  - Sequence takes n bits – each outcome totally unpredictable

- Communicating my daily lottery results
  - Output: 0 0 0 0 0 0 …
  - Most likely to take one bit – I lost every day.
  - Small chance I'll have to send more bits (won & when)

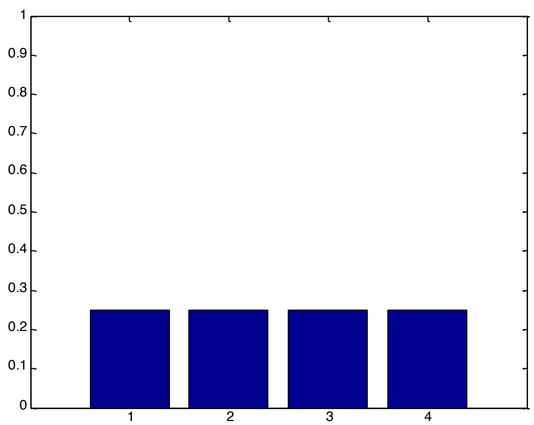
Lost: 0
Won 1: 1(…)0
Won 2: 1(…)1(…)0

- Takes less work to communicate because it's less random
  - Use a few bits for the most likely outcome, more for less likely ones`

Entropy and Information

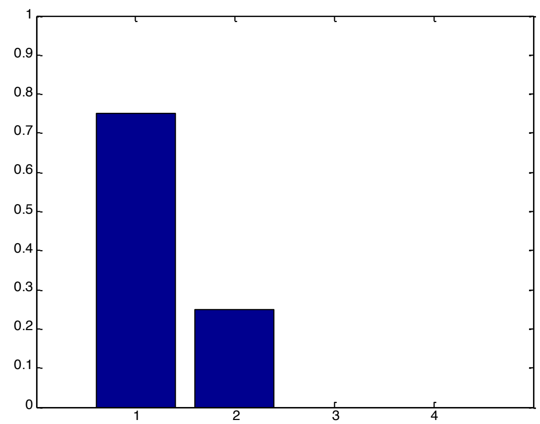
- Entropy $H(X) \equiv -\mathbb{E}_X\big[\log p(X)\big] = -\sum_x p(x)\log p(x)$
  - Log base two, units of entropy are "bits"
  - Natural log, units are "nats"

- Examples:

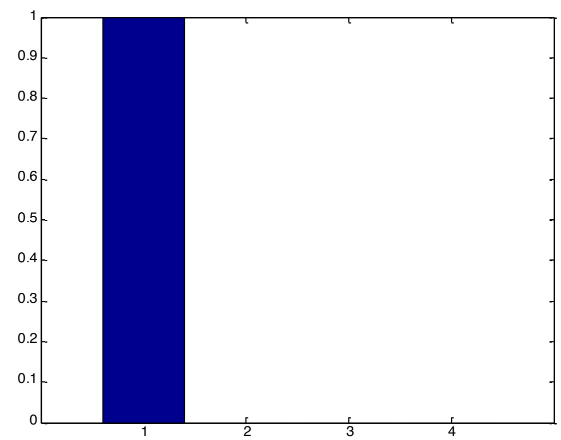


H(x) = .25 log 4 + .25 log 4 +
.25 log 4 + .25 log 4
= log 4 = 2 bits

Max entropy for 4 outcomes



H(x) = .75 log 4/3 + .25 log 4
≈ .8133 bits



H(x) = 1 log 1
= 0 bits

Min entropy

KL Divergence

- Measures dissimilarity of two distributions

$$D(\,p\,\|\,q\,) \;=\; \sum_x p(x) \log \Big[\frac{p(x)}{q(x)}\Big]$$

- "Pseudo-distance":
  - Nonnegative: $D(\,p\,\|\,q\,) \geq 0$

    $D(\,p\,\|\,q\,) = 0 \;\;\Leftrightarrow\;\; p(x) = q(x) \;\text{ a.e.}$
  - But, asymmetric: $D(\,p\,\|\,q\,) \neq D(\,q\,\|\,p\,)$

- Mutual information
  - KL divergence between true distribution and independent model:

$$I(X,Y) \;=\; D(\,p(X,Y)\,\|\,p(X)\,p(Y)\,)$$

Mutual information

- MI measures co-dependence

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$= \sum_{x,y} p(x,y) \log \left[\frac{p(x,y)}{p(x)\,p(y)} \right]$$

  - How much randomness is in X and Y individually?
  - How much randomness is in the vector (X,Y) ?
  - Also equals the KL-divergence between joint & independent model:

$$I(X,Y) = D(\,p(X,Y) \,\|\, p(X)\,p(Y)\,) \qquad \geq 0$$

- Extreme cases:
  - X,Y independent: MI = 0 (knowing X tells us 0 bits about Y)
  - X=Y: MI = H(X) (knowing X tells us H(X) bits about Y)

Summary

- Discrete random variables
- Probability distributions
  - Law of total probability; marginal probability
  - Chain rule; conditional probability
- Observing evidence
  - Posterior probabilities
  - Bayes rule
- Independence
  - Conditional independence
- Information theory
  - Entropy, mutual information, KL-divergence