

SHANNON'S PROBLEM

Given X , think of picking an element of $s \in S$ at random and reporting $X(s)$.

Can we give a number to the *expected surprise value* of this?

Basic Intuitions: the expected surprise of an unfair coin should be less than that of a fair coin.

The expected surprise of a random variable should depend only on the p 's, not on the values in V .

Expected surprise should be a kind of expected value.

AXIOMS FOR SURPRISAL

$U(X)$ is a function defined on random variables X .

Here are some conditions:

- 1 $U(X)$ only depends on the probabilities p_i and not on the values x_i or the underlying space S .
So we can write $U(p_1, \dots, p_n)$.

- 2 Among all partitions into the same number of pieces, the uniform rv's

$$U(1/n, \dots, 1/n)$$

should get the maximal values of U .

- 3 Again for partitions into the same number of pieces, U should be a continuous function of the p_i .
- 4 $U(p_1, \dots, p_n, 0) = U(p_1, \dots, p_n)$.

AN EXAMPLE TO MOTIVATE AN AXIOM

$S = 1$ million sentences, or rather the set $\{1, \dots, 1,000,000\}$.

$X(s)$ = the first word of S .

$Y(s)$ = the second word of S .

We write X, Y for the random variable giving the first two words of s .

Consider the random variable X, Y .

We don't want $U(X, Y) = U(X) + U(Y)$ because this would ignore any correlations between X and Y .

If $X(s) = \text{"President"}$, then $Y(s)$ is likely to be "Bush".

To get a handle on the correlation, we want to consider a random variable $Y|X = x_i$.

ANOTHER AXIOM

The idea is that after we hear the first word, the rest of the average surprise should be an appropriately weighted average of *further average surprisals*.

$$\begin{aligned}
 U(X, Y) &= U(X) + \sum_{i=1}^n p(x_i) \cdot U(Y|X = x_i) \\
 &= U(X) + U_X(Y)
 \end{aligned}$$

So $U_X(Y)$ measures “the uncertainty we feel about Y after we know that X has occurred but we don’t know which value it has taken.”

CHARACTERIZATION OF U

Theorem [Shannon-Khinchin]

Any U with the properties above is of the form

$$U(p_1, \dots, p_n) = -K \sum_{i=1}^n p_i \log p_i$$

for some $K \geq 0$.

For convenience we work with logarithms base 2, set $K = 1$, and redefine $0 \log 0 = 0$.

We set

$$H(p_1, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i$$

and call H the *entropy* function.

CONDITIONAL ENTROPY

We have two random variables, X and Y , on the same space.

For each y_j , we define **the conditional entropy of Y given $X = x_i$** to be

$$H(Y|X = x_i) = - \sum_j \Pr(Y = y_j|X = x_i) \log \Pr(Y = y_j|X = x_i).$$

We then define **the conditional entropy of Y given X** to be the weighted average

$$H(Y|X) = \sum_{x_i} \Pr(X = x_i) H(Y|X = x_i)$$

The idea is that $H(Y|X)$ measures the average uncertainty that remains about the value of Y when the value of X is known.

THE CHAIN RULE OF ENTROPY:

$$H(X, Y) = H(X) + H(Y|X)$$

Let $z_{i,j} = Pr(X = x_i \cap Y = y_j)$.

Let $w_{i,j} = Pr(Y = y_j | X = x_i)$. So $w_{i,j} = z_{i,j} / p_i$.

For each i ,

$$H(Y|X = x_i) = - \sum_{j=1}^m w_{i,j} \log w_{i,j}$$

And so $H(Y|X)$ is

$$\sum_{i=1}^n p_i H(Y|X = x_i) = - \sum_{i=1}^n \sum_{j=1}^m z_{i,j} \log w_{i,j}$$

VERIFICATION, CONTINUED

$$\begin{aligned}
 H(X, Y) &= - \sum_{(i,j)} z_{i,j} \log z_{i,j} \\
 &= - \sum_{i=1}^n \sum_{j=1}^m z_{i,j} \log z_{i,j} \\
 &= - \sum_{i=1}^n \sum_{j=1}^m z_{i,j} (\log w_{i,j} + \log p_i) \\
 &= H(Y|X) - \sum_{i=1}^n \sum_{j=1}^m z_{i,j} \log p_i \\
 &= H(Y|X) - \sum_{i=1}^n \log p_i \sum_{j=1}^m z_{i,j} \\
 &= H(Y|X) - \sum_{i=1}^n \log p_i \cdot p_i \\
 &= H(Y|X) + H(X)
 \end{aligned}$$

LET'S WORK OUT SOME EXAMPLES

Fl = favorite flavor of ice cream: chocolate, vanilla, strawberry, banana

		Fl			
		c	v	s	b
Age	5	$1/8$	$1/16$	$1/32$	$1/32$
	6	$1/16$	$1/8$	$1/32$	$1/32$
	7	$1/16$	$1/16$	$1/16$	$1/16$
	8	$1/4$	0	0	0

TRY IT OUT

If the last digit of your ID number is 0, compute $H(\text{Age})$.

If the last digit of your ID number is 1, compute $H(\text{Fl})$.

If the last digit of your ID number is 2, compute $H(\text{Age}|\text{Fl} = c)$.

If the last digit of your ID number is 3, compute $H(\text{Age}|\text{Fl} = v)$.

If the last digit of your ID number is 4, compute $H(\text{Age}|\text{Fl} = s)$.

If the last digit of your ID number is 5, compute $H(\text{Age}|\text{Fl} = b)$.

If the last digit of your ID number is 6, compute $H(\text{Fl}|\text{Age} = 5)$.

If the last digit of your ID number is 7, compute $H(\text{Fl}|\text{Age} = 6)$.

If the last digit of your ID number is 8, compute $H(\text{Fl}|\text{Age} = 7)$.

If the last digit of your ID number is 9, compute $H(\text{Fl}|\text{Age} = 8)$.

HERE'S ONE OTHER CALCULATION: $H(Fl, Age)$

		<i>Fl</i>			
		<i>c</i>	<i>v</i>	<i>s</i>	<i>b</i>
<i>Age</i>	5	1/8	1/16	1/32	1/32
	6	1/16	1/8	1/32	1/32
	7	1/16	1/16	1/16	1/16
	8	1/4	0	0	0

$$\begin{aligned}
 & -(1/8 \log(1/8) + 1/16 \log(1/16) + \dots + 0 \log 0) \\
 = & -(1/4 \log(1/4) + 2/8 \log(1/8) + 6/16 \log(1/16) + 4/32 \log(1/32)) \\
 = & 1/4(2) + 2/8(3) + 6/16(4) + 4/32(5) \\
 = & 108/32 \\
 = & 27/8
 \end{aligned}$$

LET'S NOW PUT SOME CALCULATIONS TOGETHER

$$H(FI|Age) = \sum_{a_i} \Pr(Age = a_i) H(FI|Age = a_i)$$

$$H(Age|FI) = \sum_{f_j} \Pr(FI = f_j) H(Age|FI = f_j)$$

We should get $H(Age, FI) = H(Age) + H(FI|Age)$
and also $H(Age, FI) = H(FI) + H(Age|FI)$

SOME PROPERTIES

Here are some basic mathematical facts about these notions. For the most part, they are good exercises with summations, and with the use of the Gibbs Inequality which we saw earlier:

$$\sum p_i \log p_i \geq \sum p_i \log q_i$$

For two sets of numbers of the same size which sum to 1.

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

$$H(X|Y) \leq H(X).$$

The difference $H(X) - H(X|Y)$ is called

mutual information $I(X; Y)$

and it satisfies $I(X; Y) = I(Y; X)$.

Also, $I(X; Y)$ is always ≥ 0 .

AN IMPORTANT APPLICATION

One important application of all of this is to the **EM algorithm**.

The proof that as we iterate the algorithm the likelihoods go up is based on **these information-theoretic inequalities**.

JAYNES' MAXIMUM ENTROPY PRINCIPLE

... in making inferences on the basis of partial information, we must use the probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to an arbitrary assumption of information which by hypothesis we do not have.

APPLICATIONS OF MAXIMUM ENTROPY

statistical mechanics
group behavior
stock market analysis
traffic networks
pattern recognition
commonsense reasoning

MAXIMUM ENTROPY IN COMPUTATIONAL LINGUISTICS

PP attachment problem:

- 1 I ate a pizza with anchovies.
- 2 I ate a pizza with friends.

Is the PP part of the NP or the VP?

Human: 93.2%

Always NP: 55%

Maximum Likelihood: 79.7%

Maxent models: 81% - 85.%, depending on several factors.

cf. Ratnaparkhi Ph.D. 1998, McLauchlan MS 2001.

WHAT IS GOING ON IN THESE MODELS

Maximum Likelihood models suffer from:

- 1 Sparse data: *I ate nsima with a carrot.*
- 2 Errors and ambiguity. *I saw the man with a telescope.*

The high-level description of MaxEnt models is

- 1 One comes up with random variables whose *expectation* on a sample we trust more than the sample's conditional probabilities.
- 2 Then we maximize the entropy of some random variable(s) subject to the condition that the expectation matches the sample.

This kind of entropy maximization leads to *exponential models*.

BINARY RANDOM VARIABLES

A random variable X with range $\{0, 1\}$ is like an *atomic proposition*.

Given X_1, \dots, X_n , we generate (formal) propositions from X_1, \dots, X_n , using boolean connectives \wedge , \vee , and \neg .

We let ϕ , ψ , etc. denote such propositions.

Recall that up to logical equivalence, every ϕ is equivalent to a disjunction of *complete conjunctions*

$$\pm X_1 \wedge \pm X_2 \wedge \cdots \wedge \pm X_n$$

These are the *atoms* of the boolean algebra generated by the propositions X_i .

We let β range over these atoms.

PROBABILITY MEASURES ON THE ATOMS

Given rvs X_1, \dots, X_n on S the *joint distribution* is X_1, \dots, X_n .
 In the binary case, this is a rv with range $\{0, 1\}^n$ (\approx atoms).
 So it determines numbers $Pr(X_1 = x_1, \dots, X_n = x_n)$,
 where $(x_1, \dots, x_n) \in \{0, 1\}^n$. These sum to 1.
 It takes $2^n - 1$ numbers to specify a joint.

Such joint distributions correspond to *probability measures* w on the atoms.

Each w then induces a function from propositions ϕ to $[0, 1]$ by

$$w(\phi) = \sum \{w(\beta) : \models \phi \rightarrow \beta\}.$$

KNOWLEDGE BASES AND SATISFACTION

Again, we fix X_1, \dots, X_n .

A Knowledge Base (KB) is a set K of formal expressions

$$Pr(\theta) = \gamma \quad \text{and} \quad Pr(\phi|\psi) = \delta.$$

where the sentences are built from the X 's and $\gamma, \delta \in [0, 1]$.

The idea is that a KB models a rational agent's total knowledge.

K should be much smaller than a joint distribution.

Satisfaction We say that a joint distribution w satisfies K

if the formal sentences in K are true when interpreted by w .

w satisfies $Pr(\phi|\psi) = \delta$ iff

$$w(\psi) \cdot \delta = w(\phi).$$

INFERENCE PROCESSES

An Inference Process is a partial function N
input: X_1, \dots, X_n , and some KB K on these
output: Some probability function w satisfying K (if one exists).
When an agent uses K and estimates probabilities of new events,
she does so in a way consistent with K and with other events.
So these values should determine a joint distribution w .
In effect, we “identify an agent with an inference process.”

AXIOMS OF INFERENCE PROCESSES

Questions

What inference process should a rational agent use?

Are there principles that govern uncertain reasoning?

Paris and Vencovská formulate *Common Sense* Principles and then study what happens.

SOME HISTORY

Cox (1946): attempt to justify the laws of conditional probability

Shore and Johnson (1980): justify ME principles based on intuitions concerning reasonable updates. Axioms somewhat different.

Csiszár (1989)

Kern-Isberner (1997,1998)

Paris and Vencovská (1989,1990,1996,1997)

Paris (1994 book, 2000)

THE IRRELEVANT INFORMATION PRINCIPLE

If K_1 and K_2 are formulated in disjoint languages,
say X_1, \dots, X_n and Y_1, \dots, Y_m ,
and if ϕ is a sentence over the X 's,
then

$$N(X_1, \dots, X_n, K_1, \phi) = N(X_1, \dots, X_n, Y_1, \dots, Y_m, K_1 \cup K_2, \phi).$$

Suppose you are asked to estimate the probability that the IU basketball team wins the NCAA championship in 2004.

Based on your KB, you estimate it to be some number γ .

Suppose next that you add to your KB some facts about the value of the Malawi Kwacha during October, 2003.

Then you should again give γ .

EQUIVALENCE PRINCIPLE

Let K_1 and K_2 be KB's in the same language.

We say that these are *equivalent* if the same w satisfy them.

Principle: If K_1 and K_2 are equivalent, then $N(K_1) = N(K_2)$.

Example: $K_1 = \{Pr(X_1 \wedge X_2) = .3\}$

$K_2 = \{Pr(X_2 \wedge X_1) = .3\}$.

RENAMING PRINCIPLE

Recall that our propositions are X_1, \dots, X_n .

Every bijection σ on $\{1, \dots, n\}$ induces a bijection on $\{X_1, \dots, X_n\}$,

hence an automorphism σ of the sentences. Write ϕ^σ .

This extends to KB's in the obvious way, keeping the numerical information.

Principle $N(K)(\phi) = N(K^\sigma)(\phi^\sigma)$.

THE RENAMING PRINCIPLE IN ACTION

Example: Suppose we consider $X_1 = \text{Heads}$ and $X_2 = \text{Tails}$.

Let $K = \{Pr(\text{Heads} \leftrightarrow \neg \text{Tails}) = 1\}$.

Let $\sigma(\text{Heads}) = \text{Tails}$, $\sigma(\text{Tails}) = \text{Heads}$.

Note that K is equivalent to K^σ .

Let $\phi = \text{Heads}$, so that $\phi^\sigma = \text{Tails}$.

Then by the principle, $N(K)(\text{Heads}) = N(K)(\text{Tails})$.

So $N(K)$ tells us that if K is our total knowledge, we should believe that a coin is fair.

In other words, *ceteris paribus*, we should believe that a coin is fair.

RELATIVIZATION PRINCIPLE, ROUGHLY

If $N(K) \models Pr(\phi|\psi) = \alpha$,
then $N(K \cup \{Pr(\chi|\neg\psi) = \beta\}) \models Pr(\phi|\psi) = \alpha$,
and conversely.

“Very roughly, the probabilities one would give if some event occurred should only depend on the knowledge one would have if that event occurred.

“The knowledge that I have about the world if the event does not occur is irrelevant.”

OBSTINACY PRINCIPLE

If $N(K_1)$ satisfies K_2 , then $N(K_1) = N(K_1 \cup K_2)$.

WEAK INDEPENDENCE PRINCIPLE

$$N(\{Pr(X) = a, Pr(Y) = b\})(X \wedge Y) = \\ N(\{Pr(X) = a, Pr(Y) = b, Pr(Z) = c, Pr(X \wedge Z) = 0\})(X \wedge Y).$$

Suppose one believes that

- ① the probability of winning the lottery is $1/200,000$.
- ② the probability of being murdered is $1/200$.

Then one makes some estimate as to the probability of both.

Someone comes and tells us $Pr[\text{die in nuclear accident}] = 1/50,000$.

We should not change our estimate.

CONTINUITY PRINCIPLE

$N(K)$ should be a continuous function of the real numbers in K .
Small changes in the numerical values of (subjective) probabilities should only produce small changes in the inference process.

PARIS & VENCOVSKÁ'S THEOREM

There is only one N that observes the principles of Renaming, Irrelevant Information, Relativization, Obstinacy, Equivalence, Weak Independence, and Continuity. This N is the *Maximum entropy solution*: Given K , find m to maximize the entropy

$$-\sum_{\beta} Pr(\beta) \log Pr(\beta).$$

Here β ranges over the atoms (\approx complete conjunctions).

IS THERE ONLY ONE COMMON SENSE PRINCIPLE?

van Fraassen's Symmetry Principle:

Essentially similar problems should have essentially similar solutions.

There is an attempt to formalize this,
so as to get one master "Common Sense" Principle.
But it didn't work out, and the matter is open.

OTHER WORK THAT I DIDN'T GET TO TALK ABOUT

Cross-Entropy: $H(X, Y) = \sum p_i \log(p_i/q_i)$.

The idea is that if we are given X as a prior belief, and also some constraints, we might update to Y satisfying the constraints such that $H(X, Y)$ is *minimum*.

Compare also with semantics for conditionals.

I didn't get to talk about use of ME in *default reasoning*:

- ① Penguins are birds.
- ② Birds fly.
- ③ Penguins do not fly.
- ④ Penguins live in the arctic.

Can we “deduce” that penguins who do not live in the arctic to not fly?

Cf. Pearl and Goldszmidt (1996).