How much can we compress? - Shannon's Source Coding Theorem

On Probability and Entropy

Probability

- An **ensemble** X is a random variable x with a set of possible **outcomes** $\mathcal{A}_x$ with **probabilities** $\mathcal{P}_x$
- **Probability of a subset** T of $\mathcal{A}_x$

$$P(T) = \sum_{a_i \in T} P(x = a_i)$$

- A **joint ensemble** XY is an ensemble for which the outcomes are ordered pairs x,y where $x \in \mathcal{A}_x$ and $y \in \mathcal{A}_y$

Probability continued

- **Marginal probability** (from the joint probability P(x,y))

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

- **Conditional probability**

$$P(x = a_i \mid y = b_j) \equiv \frac{P(x = a_i, y = b_j)}{P(y = b_j)}$$

Probability continued

- **Product rule**

$$P(x,y \mid H) = P(x \mid y,H)P(y \mid H)$$

- **Sum rule**

$$P(x \mid H) = \sum_y P(x,y \mid H) = \sum_y P(x \mid y,H)P(y \mid H)$$

Bayes's theorem

$$P(y \mid x,H) = \frac{P(x \mid y,H)P(y \mid H)}{P(x \mid H)} = \frac{P(x \mid y,H)P(y \mid H)}{\sum_{y'} P(x \mid y',H)P(y' \mid H)}$$

Bayesian view of probability!

Information content

- First attempt: number of possible outcomes $|\mathcal{A}_X|$
  - ✓ not additive: for xy we have $|\mathcal{A}_X||\mathcal{A}_Y|$
- Perfect information content

$$H_0(X) = \log_2 |A_X|$$

  - ✓ additive, but no probabilistic element

Shannon information

- looking for an information content of the event $x=a_i$

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$

Information = decreased uncertainty

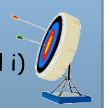
- Example: 4 outcomes a,b,c,d with probabilities p(a), p(b), p(c) and p(d)
- Sender knows the result, receiver doesn't
- Binary channel (yes/no questions)
- A lot of questions ⇨ a lot of information
- "code" = sequence of answers to questions
  - Is it a or b? Is it a (Is it c)?
  - Is it a? Is it b? Is it c?
- Case 1: P(a) = 1
- Case 2: P(a) = P(b) = P(c) = P(d) = 1/4
- Case 3: P(a)=1/2, P(b)=1/4, P(c)=P(d)=1/8

Entropy

- The entropy of X is a measure of the *expected* information content or "decreased uncertainty" of an event x

$$H(X) \equiv \sum_{x \in A_X} P(x) \log \frac{1}{P(x)}$$

  - ✓ $H(X) \geq 0$ (= iff $p_i=1$ for one i)
  - ✓ $H(X) \leq \log(|X|)$ (= iff $p_i=1/|X|$ for all i)



Binary entropy

$$H(X) \equiv \sum_i p_i \log_2 \frac{1}{p_i}$$

Information measure?

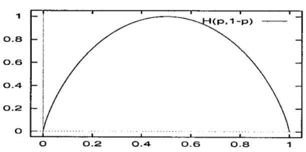


Figure 2.1. The binary entropy function $H_2(p) = H(p, 1-p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)}$ as a function of p.

Example: letter distribution

i	a_i	p_i	$\log_2 \frac{1}{p_i}$	i	a_i	p_i	$\log_2 \frac{1}{p_i}$	i	a_i	p_i	$\log_2 \frac{1}{p_i}$
1	a	0.06	4.1	10	j	0.00	10.7	19	s	0.06	4.1
2	b	0.01	6.3	11	k	0.01	6.9	20	t	0.07	3.8
3	c	0.03	5.2	12	l	0.04	4.9	21	u	0.03	4.9
4	d	0.03	5.1	13	m	0.02	5.4	22	v	0.01	7.2
5	e	0.09	3.5	14	n	0.06	4.1	23	w	0.01	6.4
6	f	0.02	5.9	15	o	0.07	3.9	24	x	0.01	7.1
7	g	0.01	6.2	16	p	0.02	5.7	25	y	0.02	5.9
8	h	0.03	5.0	17	q	0.01	10.3	26	z	0.00	10.4
9	i	0.06	4.1	18	r	0.05	4.3	27	-	0.19	2.4

$$\sum_i p_i \log_2 \frac{1}{p_i} \quad 4.11$$

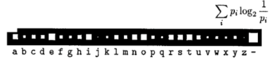
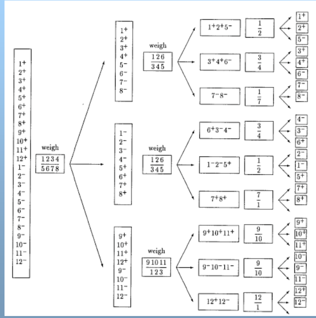


Figure 1.16. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The frequently asked questions manual for Linux*). The picture shows the probabilities by the sizes of white squares.

Weighting problem



Entropy continued

- The joint entropy of X,Y

$$H(X,Y) \equiv \sum_{xy \in A_X A_Y} P(x,y) \log \frac{1}{P(x,y)}$$

- The conditional entropy of X given Y

$$H(X \mid Y) \equiv \sum_{y \in A_Y} P(y) \left[\sum_{x \in A_X} P(x \mid y) \log \frac{1}{P(x \mid y)} \right]$$

$$= \sum_{xy \in A_X A_Y} P(x,y) \log \frac{1}{P(x \mid y)}$$

"Average uncertainty that remains about x when y is known"

Entropy continued

- Chain rule for entropy

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

- Mutual information

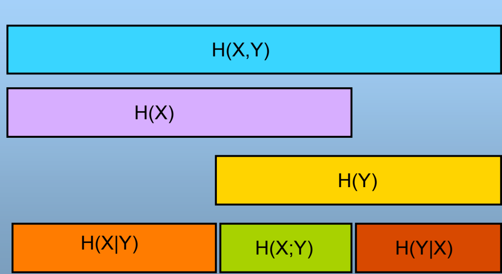
"Average reduction in uncertainty of x when learning the value of y

H(X;Y) ≡ H(X) - H(X|Y)

- Entropy distance

D_H(X,Y) ≡ H(X,Y) - H(X;Y)

Entropy relationships



Kullback-Leibler divergence

- Also known as "relative entropy"

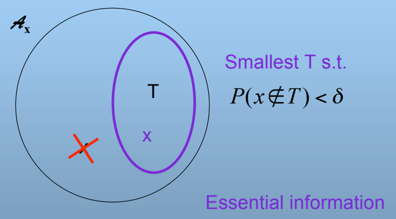
$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Not strictly a "distance"

Idea

- Some symbols have a smaller probability
- gamble that the rare symbols won't occur
- encode the observations in a smaller code (alphabet) C_X
- measure $\log_2 |C_X|$
- the larger the risk, the smaller the alphabet

Formalize the idea



$$H_\delta(X) = \log_2 \min\{|T| : T \subseteq A_X, P(x \in T) \geq 1 - \delta\}$$

Block coding

- assume that $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ i.i.d.
- independent variables, thus $H(X^N) = NH(X)$
- $H_\delta(X^N)$ depends on the value of δ, so where is the theory?
- N grows, $H_\delta(X^N)$ becomes almost independent of δ!

Shannon's source coding theorem

Let X be an ensemble with entropy H(X) bits. Given $\varepsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 s.t.
For $N > N_0$,

$$\left| \frac{1}{N} H_\delta(X^N) - H(X) \right| < \varepsilon$$

Typical set

- for long strings

$$p(\mathbf{x})_{typical} \cong p_1^{(p_1 N)} p_2^{(p_2 N)} \dots p_j^{(p_j N)}$$

- the information content of a typical string is

$$\log \frac{1}{p(\mathbf{x})} \cong N \sum_i p_i \log_2 \frac{1}{p_i} \cong NH$$

- the typical set $T_{N\beta} \equiv \left\{ x \in A_X^N : \left| \frac{1}{N} \log_2 \frac{1}{P(\mathbf{x})} - H(\mathbf{x}) \right| < \beta \right\}$

AEP and source coding

Asymptotic Equipartition Principle: for N i.i.d. random variables $X^N = \{X_1, \dots, X_N\}$, with N sufficiently large, the outcome $\mathbf{x} = \{x_1, \dots, x_N\}$ is almost certain to belong to a subset of $\mathbf{A_x}^N$ having only $2^{NH(X)}$ members all having probability close to $2^{-NH(X)}$