Coding and Information Theory

Chris Williams

School of Informatics, University of Edinburgh

November 2007

Overview

- What is information theory?
- Entropy
- Coding
- Rate-distortion theory
- Mutual information
- Channel capacity
- Reading: Bishop §1.6

Information Theory

Shannon (1948): Information theory is concerned with:

- **Source coding**, reducing redundancy by modelling the structure in the data
- **Channel coding**, how to deal with "noisy" transmission
- Key idea is **prediction**
  - Source coding: redundancy means predictability of the rest of the data given part of it
  - Channel coding: Predict what we want given what we have been given

Information Theory Textbooks

- Elements of Information Theory. T. M. Cover and J. A. Thomas. Wiley, 1991. [comprehensive]
- Coding and Information Theory. R. W. Hamming. Prentice-Hall, 1980. [introductory]
- Information Theory, Inference and Learning Algorithms D. J. C. MacKay, CUP (2003), available online (viewing only) `http://www.inference.phy.cam.ac.uk/mackay/itila`

Entropy

A discrete random variable X takes on values from an alphabet $\mathcal{X}$, and has probability mass function $P(x) = P(X = x)$ for $x \in \mathcal{X}$. The entropy $H(X)$ of X is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$$

convention: for $P(x) = 0$, $0 \times \log 1/0 \equiv 0$
The entropy measures the information content or "uncertainty" of X.
Units: $\log_2 \Rightarrow$ bits; $\log_e \Rightarrow$ nats.

Joint entropy, conditional entropy

$$\begin{aligned}
H(X,Y) &= -\sum_{x,y} P(x,y) \log P(x,y) \\
H(Y|X) &= \sum_{x} P(x) H(Y|X=x) \\
&= -\sum_{x} P(x) \sum_{y} P(y|x) \log P(y|x) \\
&= -E_{P(x,y)} \log P(y|x) \\
H(X,Y) &= H(X) + H(Y|X)
\end{aligned}$$

If X, Y are independent

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

Coding theory

A coding scheme C assigns a code $C(x)$ to every symbol x; $C(x)$ has length $\ell(x)$. The expected code length $L(C)$ of the code is

$$L(C) = \sum_{x \in \mathcal{X}} p(x) \ell(x)$$

Theorem 1: Noiseless coding theorem
The expected length $L(C)$ of any instantaneous code for X is bounded below by $H(X)$, i.e.

$$L(C) \geq H(X)$$

Theorem 2
There exists an instantaneous code such that

$$H(X) \leq L(C) < H(X) + 1$$

Practical coding methods

How can we come close to the lower bound ?

- Huffman coding

  $$H(X) \leq L(C) < H(X) + 1$$

  Use blocking to reduce the extra bit to an arbitrarily small amount.
- Arithmetic coding

Coding with the wrong probabilities

Say we use the wrong probabilities q_i to construct a code. Then

$$L(C_q) = -\sum_i p_i \log q_i$$

But

$$\sum_i p_i \log \frac{p_i}{q_i} > 0 \ \text{ if } q_i \neq p_i$$

$\Rightarrow$

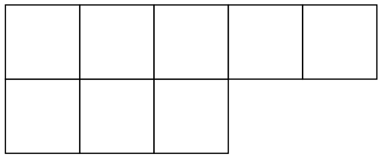
$$L(C_q) - H(X) > 0$$

i.e. using the wrong probabilities increases the minimum attainable average code length.

Coding real data

- So far we have discussed coding sequences if iid random variables. But, for example, the pixels in an image are not iid RVs. So what do we do ?
- Consider an image having N pixels, each of which can take on k grey-level values, as a single RV taking on k^N values. We would then need to estimate probabilities for all k^N different images in order to code a particular image properly, which is rather difficult for large k and N.
- One solution is to chop images into blocks, e.g. 8×8 pixels, and code each block separately.

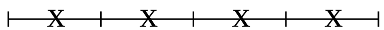
• Predictive encoding – try to predict the current pixel value given nearby context. Successful prediction reduces uncertainty.



$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

Rate-distortion theory

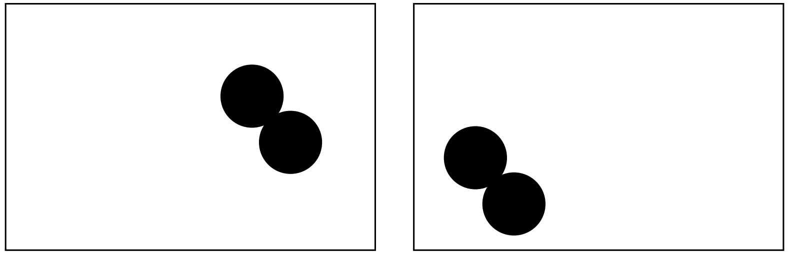
What happens if we can't afford enough bits to code all of the symbols exactly ? We must be prepared for *lossy* compression, when two different symbols are assigned the same code.
In order to minimize the errors caused by this, we need a *distortion function* $d(x_i, x_j)$ which measures how much error is caused when symbol x_i codes for x_j.



The k-means algorithm is a method of choosing code book vectors so as to minimize the expected distortion for $d(x_i, x_j) = (x_i - x_j)^2$

Source coding

- Patterns that we observe have a lot of structure, e.g. visual scenes that we care about don't look like "snow" on the TV
- This gives rise to **redundancy**, i.e. that observing part of a scene will help us predict other parts
- This redundancy can be exploited to code the data efficiently—*loss less* compression



- Q: Why is coding so important?
- A: Because of the lossless coding theorem: the best probabilistic model of the data will have the shortest code
- Source coding gives us a way of comparing and evaluating different models of data, and searching for good ones
- Usually we will build models with *hidden variables*— a new *representation* of the data

Mutual information

$$
\begin{aligned}
I(X;Y) &= KL(p(x,y), p(x)p(y)) \geq 0 \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = I(Y;X) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= H(X) - H(X|Y) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

- Mutual information is a measure of the amount of information that one RV contains about another. It is the reduction in uncertainty of one RV due to knowledge of the other.
- Zero mutual information if X and Y are independent

Mutual Information

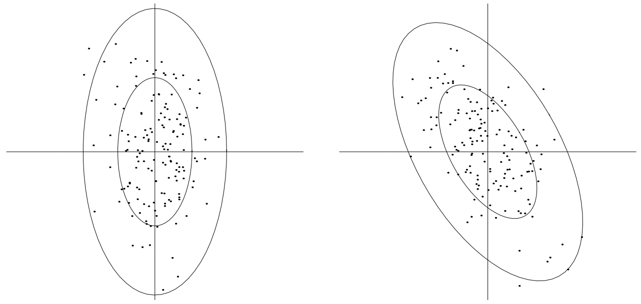
- Example 1:

		Y_1	
		smoker	non smoker
Y_2	lung cancer	1/3	0
	no lung cancer	0	2/3

- Example 2:

		Y_1 smoker	Y_1 non smoker
Y_2	lung cancer	1/9	2/9
Y_2	no lung cancer	2/9	4/9

Continuous variables



$$I(Y_1;Y_2) = \int\int P(y_1,y_2)\log\frac{P(y_1,y_2)}{P(y_1)P(y_2)}\,dy_1\,dy_2 = -\frac{1}{2}\log(1-\rho^2)$$

PCA and mutual information

Linsker, 1988, Principle of maximum information preservation
Consider a random variable $Y = \mathbf{a}^T\mathbf{X} + \epsilon$, with $\mathbf{a}^T\mathbf{a} = 1$.
How do we maximize $I(Y;\mathbf{X})$?

$$I(Y;\mathbf{X}) = H(Y) - H(Y|\mathbf{X})$$

But $H(Y|\mathbf{X})$ is just the entropy of the noise term ϵ. If $\mathbf{X}$ has a joint multivariate Gaussian distribution then Y will have a Gaussian distribution. The (differential) entropy of a Gaussian $N(\mu,\sigma^2)$ is $\frac{1}{2}\log 2\pi e\sigma^2$. Hence we maximize information preservation by choosing $\mathbf{a}$ to give Y maximum variance subject to the constraint $\mathbf{a}^T\mathbf{a} = 1$.

Channel capacity

The channel capacity of a discrete memoryless channel is defined as

$$C = \max_{p(x)} I(X;Y)$$

Noisy channel coding theorem
(Informal statement) Error free communication above the channel capacity is impossible; communication at bit rates below C is possible with arbitrarily small error.