Information Theory, Pattern Recognition and Neural Networks

Handout 3 March 1, 2006

1 Course summary: central chapters

Data compression and noisy channel coding (Chapters 1–6, 8–10, 14).
Inference and data modelling. (Chapters 20 and 22).

2 Supervisions to come

5: Thursday 9th March, Ryle Seminar Room, 2pm and 5.30pm.
6: Thursday 16th March, HEP Seminar Room, 2pm and 5.30pm.

3 Exercises that have been recommended

1: Invent a code. 1.3 (p.8), 1.5-7 (p.13), **1.9**, & 1.11 (p.14).
2–3: Invent a compressor. **ex 5.29** (p. 103), 5.22, 5.27, 6.7, 6.17.
then if you need more practice, 5.26, 5.28, 6.15, 15.3 (p. 233)
4: Invent a channel. 9.17 (p. 155) 10.12 (172) 15.12 (235); then if you need more practice, 15.11, 15.13, 15.15.
5–6: See question on handout 2.
Examples 22.1-4 (p. 300) and exercise 22.8.
Ex 3.10 (p57) (children); 8.10, black and white cards; 9.19 TWOS; 9.20, birthday problem; 15.5, 15.6, (233) magic trick; 8.3 (140), 8.7; 22.11 sailor.
Ex 22.5.

4 What's on the exam

Data compression. Evaluating entropy, conditional entropy, mutual information. Symbol codes. Huffman algorithm. 'How well would arithmetic coding do?'

Noisy channels. Evaluating conditional entropy, mutual information. Definition of capacity. Evaluating capacity. Finding optimal input distributions. Inference of input given output. Connection to reliable communication.

Inference problems. Inferring parameters. Comparing two hypotheses. Sketching posterior distributions. Finding error bars.

5 For the final supervision

Invent a supervision. Send me an email suggesting what you would like to happen in the final supervision.

Possible exercise for supervision 6:

Here are two approaches that have been suggested to the problem in handout 2:

Bayes' theorem, in which the log likelihood ratio is

$$\log \frac{P(\mathbf{x} \mid \mathcal{H}_P)}{P(\mathbf{x} \mid \mathcal{H}_Q)} = \sum_i F_i \log \frac{p_i}{q_i},$$

where i runs over characters in the alphabet, F_i is the number of times character i actually occurred in the data string $\mathbf{x}$, and the two models $\mathcal{H}_P$ and $\mathcal{H}_Q$ state that the symbols come i.i.d. from the distributions $\mathbf{p}$ and $\mathbf{q}$ respectively.

Chi-squared. In a chi-squared approach, we compute the two measures of goodness of fit,

$$\chi^2_P = \sum_i \frac{(F_i - p_i N)^2}{p_i N}$$

$$\chi^2_Q = \sum_i \frac{(F_i - q_i N)^2}{q_i N},$$

where N is the number of characters received; then go for the hypothesis with smaller χ^2.

These two approaches do not always make the same decision. (Notice that the log likelihood ratio is a *linear* function of $\{F_i\}$ whereas $\chi^2_P - \chi^2_Q$ has a *quadratic* dependence on $\{F_i\}$.)

Task: seek out examples that magnify the differences between these two approaches. (a) Can you find an example data set, and pair of hypotheses $\mathbf{p}$ and $\mathbf{q}$, for which the two approaches give completely different answers? (b) Can you find two data sets that are intuitively equivalent from the point of view of comparing $\mathbf{p}$ and $\mathbf{q}$, but for which one of the approaches gives different answers?