Exercise Sheet 4

Probabilistic models of modern AI

vassilis.papadopoulos@epfl.ch – www.vassi.life/teaching/aiproba

References :

- Cover & Thomas, 'Elements of Information Theory', Chapter 5 [link]

We will be interested in more compression. In particular, we want to compress symbols that are drawn from a random variable X, which can take discrete values in $\mathcal{X}$. A symbol code $C : \mathcal{X} \to D^*$ is a mapping of the symbols in $\mathcal{X}$ to strings composed from elements in D (the codewords), where D is our alphabet (usually $D = \{0, 1\}$).

We can extend a code to encode sequences $s \in |\mathcal{X}|^n$ simply by concatening codewords, i.e. $C(x_1 x_2 \cdots x_n) = C(x_1)C(x_2) \cdots C(x_n)$. A code will be called uniquely decodable if the mapping it induces on sequences is injective (namely no two sequences are mapped to the same string).

Generally, you can set $D = 2$ for all exercises, and all proofs follow for any D. We will always write $H(X)$ and log, though when we are in base D, the correct quantities to consider are $H_D(X)$ and $\log_D$.

Exercise 1 Properties of Prefix Codes

A more restrictive class of symbol codes are 'prefix codes', which are such that no codeword is the prefix of any other.

1. Explain why prefix codes are also called 'instantaneously decodable'. For instantaneously decodable codes, one can decode the sequences symbol by symbol, rather than needing the entire sequence at once to decode it.
2. Prefix code satisfy a very simple condition known as the Kraft inequality:

$$\sum_i D^{-\ell_i} \leq 1 \tag{1}$$

   Where D is the number of coding symbols (usually $D = 2$ for binary), and ℓ_i are the lengths of the codes. The sum runs over all codes, equivalently all symbols of $\mathcal{X}$. We want to show this inequality. Let ℓ_{max} be the length of the longest codeword in a prefix code.
   a) Consider all code strings of length ℓ_{max}. If we have a codeword of length ℓ_i, how many of these strings become 'taken' due to the prefix property?
   b) Show that two sets of 'forbidden' ℓ_{max}-strings coming from any two codewords of lengths ℓ_j and ℓ_i are non-overlapping due to the prefix property.
   c) Deduce the Kraft inequality.
3. Actually, the Kraft inequality goes both ways! Given lengths ℓ_i satisfying the Kraft inequality, there exists a prefix code realizing these lengths.
   a) To prove it, perform an induction on ℓ_{max}, the length of the maximal codeword.
4. As it turns out, the Kraft inequality is even more powerful than that. It is also satisfied by uniquely decodable symbol codes! We will attempt to prove this here.
   a) Consider $\left(\sum_i D^{-\ell_i}\right)^k$, for an arbitrary integer k. Show

$$\left(\sum_i D^{-\ell_i}\right)^k = \sum_{m=1}^{k\times\ell_{max}} \rho(m) D^{-m} \tag{2}$$

   Where $\rho(m)$ is the number of sequences $x_{i_1} \cdots x_{i_k}$ whose encoding is of length m.

b) Using the uniquely decodable property, find a coarse bound on $\rho(m)$, and conclude by taking the limit $k \to \infty$.

So we know now that a uniquely decodable code satisfies the Kraft inequality and that conversely, given codeword lengths that satisfy it, we can craft a uniquely decodable code (even better, a prefix code).

Exercise 2 Optimal prefix codes

We would like to see how far we can push prefix codes in terms of their compression capabilities. We define the average length of a symbol code as:

$$\sum_i p(x_i)\ell_i \tag{3}$$

Where $p(x_i) = P(X = x_i)$, and ℓ_i is the length of the codeword for x_i.

1. Ignore for now the fact that the ℓ_i are integers. Argue that an optimal coding will be obtained when $\sum D^{-\ell_i} = 1$.
2. Setup a constrained optimization problem, and find the optimal values for the ℓ_i. (don't bother showing that it is indeed a global minimum for now)
3. What lower and upper bound do you obtain on the average length, assuming now you need to choose integer ℓ_i?

Exercise 3 Shannon coding theorem, part 2

In Exo 2, we found a code that yielded a code length smaller than $H(X) + 1$.
We will now show a lower bound.

1. Consider a symbol code with codeword lengths ℓ_i. Let $L = \sum_i p(x_i)\ell_i$ its average codelength. We remind $H_D(X) = -\sum_i p_i \log_D(p_i)$ is the entropy in base D.. Show one can write

$$L - H_D(X) = KL_D(p \parallel r) - \log_D(Z) \tag{4}$$

   Where $p_i = p(x_i)$, the r_i are some probability distribution you should exhibit, and Z is a number that depends on the ℓ_i, that you should also find. KL_D is the Kullback-Leibler divergence using $\log_D$.
2. Conclude that $L \geq H_D(X)$ if the source code satisfies the Kraft inequality
3. Deduce that $L \geq H_D(X)$ must hold for any uniquely decodable symbol code
4. For a uniquely decodable symbol code, show we have $L = H_D(X)$ iff $p_i = D^{-\ell_i}$.

You have shown that $H(X)$ is a lower bound for symbol codes. In exercise sheet 3, we considered a coding scheme that achieved $L \leq H(X) + \epsilon$ for long enough sequences. However, it was not a symbol code, as each symbol x_i was not assigned a codeword, rather we assigned codes to entire sequences instead. Could it be that this more powerful set of codes allows us to break the $H(X)$ barrier?

5. Show that the average bits per symbol of a coding scheme that assigns uniquely decodable codewords to sequences of n symbols is still bounded from below by $H(X)$.[1]

All in all, you have proved the Shannon coding theorem[2]: for sequences generated from an iid random variables X, the expected number of bits per symbol L of an optimal coding scheme satisfies

$$H(X) \leq L < H(X) + 1 \tag{5}$$

Namely we can always losslessly compress within less than one bit of the entropy. From Exercise sheet 3, we know we can get arbitrarily close to $H(X)$ when the sequence length grows, though not with a symbol code!

[1] Hint in rot13: Pbafvqre gur frdhraprf nf n arj frg bs flzobyf!
[2] Pbatengf!