Fourth Lecture
September 25, 2023

Optimal codes

We have seen constructions giving average codeword length close to the lower bound $H_s(P)$, but nothing guaranteed that any of these codes would be best possible. So the question of how to find an optimal average length code comes up. This will be answered by constructing the so-called Huffman code. We will study this only for the binary case, i.e, when the size of the code alphabet is $s = 2$.

Def. A code f is *optimal* if $\mathbb{E}\,|f(X)| \leq \mathbb{E}\,|f'(X)|$ for all codes $f' : \mathcal{X} \to \mathcal{Y}^*$

We discussed that optimal code does exist, since there are finitely many possible codes, and there are more than one possibility for an optimal code, since eg. inverting the bits doesn't change the average codeword length but results in a different code. Similarly interchanging the codewords of the same length.

Assume $p_1 \geq \cdots \geq p_r > 0$, $p_i = p(x^{(i)})$ and having an optimal binary code $C = (f(x^{(1)}), \ldots, f(x^{(r)}))$, $l_i := |f(x^{(i)})|$. By the foregoing we can assume that the code is prefix. (Note that the $p_r > 0$ assumption is not a real restriction: if we have 0-probability events, they need not be encoded. Or they could even be encoded into long codewords, since their contribution to the average length will be zero anyway.)

Theorem 8 *(Properties of optimal code) If the prefix code $f : \mathcal{X} \to \{0,1\}^*$ is optimal then (there is a reordering of the source symbols and the codewords of the same length such that)*
(1) $l_1 \leq l_2 \leq \cdots \leq l_r$
(2) $l_r = l_{r-1}$
(3) the two longest codewords $f(x^{(r)})$ and $f(x^{(r-1)})$ differ only in the last bit.

Proof:
(1) $l_1 \leq l_2 \leq \cdots \leq l_r$. This is true, because if this is not satisfied, then we may exchange codewords without increasing the average length.

(2) Suppose that it is not true, then $l_r > l_{r-1}$ by *(1)* above and since the code is prefix, deleting the last digit of $f(x^{(r)})$ would result in a prefix code with smaller average length, so the original code was not optimal.

(3) Exchanging the last digit of the codeword $f(x^{(r)})$ we should get another codeword (otherwise this last digit could have been deleted without ruining the prefix property), and if this codeword is not $f(x^{(r-1)})$ but some $f(x^{(i)})$ with $i \neq r-1$, then we can simply exchange the two without effecting the average length as these two codewords both have the same length $|f(x^{(i)})| = |f(x^{(r)})| = |f(x^{(r-1)})|$. □

A very important observation follows that leads us to the optimal code construction Huffman algorithm.

Theorem*: Cutting the last digit of the two codewords $f(x^{(r-1)})$ and $f(x^{(r)})$ we obtain an optimal binary prefix code for the distribution $(p_1, p_2, \ldots, p_{r-2}, p_{r-1} + p_r)$.*

Proof:
This is true because the average length L of our code is $L' + p_{r-1} + p_r$, where L' is the average length of the code obtained by identifying the codewords $f(x^{(r-1)})$ and $f(x^{(r)})$ by cutting their last digit. If there was a better (i.e., one with smaller average length) prefix code for the distribution $(p_1, p_2, \ldots, p_{r-2}, p_{r-1}+p_r)$, then extending the codeword belonging to the probability $p_{r-1} + p_r$ source symbol once with a 0 digit and once with a 1 digit, we would obtain a better code than our original one, so its average length could have not been optimal. □

Huffman code

From these three observations the optimal code construction is immediate: add two smallest probabilities iteratively until only two distinct ones remain. Give these the (sub)words 0 and 1 and then follow the previous "adding up two probabilities" process backwards and put a 0 and a 1 at the end of the corresponding codewords.

Example:

$$P = (0.25, 0.14, 0.13, 0.12, 0.11.0.1, 0.1, 0.05)$$

The "merged" distributions are:

$(0.25, 0.14, 0.13, 0.12, 0.11, 0.1, 0.15 = \mathit{0.1 + 0.05})$

$(0.25, 0.14, 0.13, 0.12, 0.21 = \mathit{0.11 + 0.1}, 0.15 = \mathit{0.1 + 0.05})$

$(0.25, 0.14, 0.25 = \mathit{0.13 + 0.12}, 0.21 = \mathit{0.11 + 0.1}, 0.15 = \mathit{0.1 + 0.05})$

$(0.25, 0.29 = \mathit{0.14 + 0.15}(= \mathit{0.1 + 0.05}), 0.25 = \mathit{0.13 + 0.12}, 0.21 = \mathit{0.11 + 0.1})$

$(0.46 = \mathit{0.25 + 0.21}(= \mathit{0.11 + 0.1}), 0.29 = \mathit{0.14 + 0.15}(= \mathit{0.1 + 0.05}), 0.25 = \mathit{0.13 + 0.12})$

$(0.46 = \mathit{0.25 + 0.21}(= \mathit{0.11 + 0.1}), 0.54 = \mathit{0.29}(= \mathit{0.14 + 0.15}(= \mathit{0.1 + 0.05})) + \mathit{0.25}(= \mathit{0.13 + 0.12}))$

And the code obtained writing it backwards for each stage of the construction:

$(0, 1)$

$(0, 10, 11)$

$(00, 01, 10, 11)$

$(00, 100, 101, 01, 11)$

$(00, 100, 110, 111, 101, 01)$

$(00, 100, 110, 111, 010, 011, 101)$

and finally

$$(00, 100, 110, 111, 010, 011, 1010, 1011)$$

We discussed that if there are more than one possibilities to choose and combine the two least likely symbols, then it is up to us which way we go on. Each leads us to an optimal Huffman code.

Also there is no rule in which order to assign the 0 and the 1 to the branches. We might as well assign 0 to the left branch and 1 to the right in the first step and do it the other way round in the next step.

Exercise: Two people made two different Huffman codes for the distribution $p_1 \geq p_2 \geq p_3 \geq p_4$. The codewords of these codes are $0, 10, 110, 111$ for one and $00, 01, 10, 11$ for the other. Determine the distribution if we know that $p_3 = 1/6$.

Remark 1: Huffman code can be constructed for any code alphabet size, not just for binary alphabet. If $s \neq 2$ then we combine the $s - 1$ least likely symbols in each step. However in this case, we might need to use "dummy symbols", symbols with probability 0, in order to ensure that there are $s - 1$ symbols even at the last step to be combined. There need to be $k \cdot (s - 1) + 1$ symbols at the beginning. We constructed the ternary Huffman code for the above distribution as an example, and we needed one dummy symbol for that.

Remark 2: If the distribution is unknown, we can estimate the probabilities by the relative frequencies. After reading K symbols, let w_i denote the number of occurrences of $x^{(i)}$. Then $\hat{p}_i = \frac{w_i}{K}$ is a good estimate of p_i. Moreover it is not necessary to divide w_i by K, since the construction doesn't use that the sum of the labels is one. So we might use the w_is in the Huffman construction.

Home-work

Exercise 1: Which ones of the following codes can and which ones cannot be a Huffman code?

a) $0, 10, 111, 101$

b) $00, 010, 011, 10, 110$

c) $1, 000, 001, 010, 011$

Exercise 2: Two people made two different Huffman codes for the distribution $p_1 \geq p_2 \geq p_3 \geq p_4$. The codewords of these codes are $0, 10, 110, 111$ for one and $00, 01, 10, 11$ for the other. Determine the distribution if we know that $p_3 = 1/6$.

More on the entropy function

Let us denote the *joint probability distribution* of random variable X and random variable Y:
$p(x, y) = Prob(X = x, Y = y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

The *marginal distribution* of X: $p(x) = Prob(X = x)$ for all $x \in \mathcal{X}$

The *marginal distribution* of Y: $p(y) = Prob(Y = y)$ for all $y \in \mathcal{Y}$

The *conditional distribution* of X given $Y = y$: $p(x|y) = Prob(X = x|Y = y)$ for all $x \in \mathcal{X}$

The *conditional distribution* of Y given $X = x$: $p(y|x) = Prob(Y = y|X = x)$ for all $y \in \mathcal{Y}$

Remember from Probability Theory that

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y) \qquad p(y) = \sum_{x \in \mathcal{X}} p(x, y) \qquad p(x|y) = \frac{p(x, y)}{p(y)} \qquad p(y|x) = \frac{p(x, y)}{p(x)}$$

and that X and Y are *independent* if for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$p(x, y) = p(x) \cdot p(y)$$

Def. The *joint entropy* of X and Y is simply the entropy of the joint distribution of the variable (X, Y)

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

Def. The *conditional entropy* is defined as:

$$H(X|Y) = \sum_y p(y) H(X|Y = y) =$$

$$= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) =$$

$$= -\sum_{x,y} p(x, y) \log p(x|y) =$$

$$= -\sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)}.$$

Property 1

$$H(X, Y) \leq H(X) + H(Y)$$

with equality iff X and Y are independent.

Note the intuitive plausibility of the statement. (The information content of the pair (X, Y) is not more than the sum of the information X and Y contain separately. And equality means that they "do not contain information about each other", that is, they are independent.)

Proof. Follows by applying Corollary 4 of Jensen's inequality for $p = p(x, y)$ and $q = p(x)p(y)$. In details:

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

$$-\sum_x \left(\sum_y p(x, y)\right) \log p(x) - \sum_y \left(\sum_x p(x, y)\right) \log p(y) + \sum_{x,y} p(x, y) \log p(x, y) =$$

$$\sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \geq 0.$$

Equality holds iff $p(x, y) = p(x)p(y) \forall x, y$, i.e. iff X and Y are independent. □

Example: Consider the random variables X and Y with the following joint distribution

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

We calculated the marginal distributions:

$Prob(Y = 1) =$
$Prob(X = 1, Y = 1) + Prob(X = 2, Y = 1) + Prob(X = 3, Y = 1) + Prob(X = 4, Y = 1) = \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{32} = \frac{1}{4}$
$Prob(Y = 2) =$
$Prob(X = 1, Y = 2) + Prob(X = 2, Y = 2) + Prob(X = 3, Y = 2) + Prob(X = 4, Y = 2) = \frac{1}{16} + \frac{1}{8} + \frac{1}{32} + \frac{1}{32} = \frac{1}{4}$
and similarly $Prob(Y = 3) = \frac{1}{4}$, $Prob(Y = 4) = \frac{1}{4}$.

$Prob(X = 1) =$
$Prob(X = 1, Y = 1) + Prob(X = 1, Y = 2) + Prob(X = 1, Y = 3) + Prob(X = 1, Y = 4) = \frac{1}{8} + \frac{1}{16} + \frac{1}{16} + \frac{1}{4} = \frac{1}{2}$
$Prob(X = 2) =$
$Prob(X = 2, Y = 1) + Prob(X = 2, Y = 2) + Prob(X = 2, Y = 3) + Prob(X = 2, Y = 4) = \frac{1}{16} + \frac{1}{8} + \frac{1}{16} + 0 = \frac{1}{4}$
and similarly $Prob(X = 3) = \frac{1}{8}$, $Prob(X = 4) = \frac{1}{8}$.

We discussed that these two random variables are NOT independent, since for example
$0 = Prob(X = 4, Y = 4) \neq Prob(X = 4) \cdot Prob(Y = 4) = \frac{1}{8} \cdot \frac{1}{4} = \frac{1}{32}$

Then we calculated the entropies:
$H(X) = H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) = 1.75$bits, $H(Y) = H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = \log 4 = 2$bits

$H(X, Y) = H\left(\frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{4}, 0, 0, 0\right) = -\frac{1}{4} \log \frac{1}{4} - \left(\frac{1}{8} \log \frac{1}{8}\right) \cdot 2 - \left(\frac{1}{16} \log \frac{1}{16}\right) \cdot 6 - \left(\frac{1}{32} \log \frac{1}{32}\right) \cdot 4 = \frac{27}{8}$bits

We checked that $\frac{27}{8} = H(X, Y) < H(X) + H(Y) = \frac{30}{8}$

In order to calculate the conditional entropy
$H(X|Y) = \sum_y p(y) H(X|Y = y) = \frac{1}{4} H(X|Y = 1) + \frac{1}{4} H(X|Y = 2) + \frac{1}{4} H(X|Y = 3) + \frac{1}{4} H(X|Y = 4)$,

we needed the conditional distributions of X. Those we can get from the joint and the marginal distributions. Eg. $Prob(X=1|Y=1) = \frac{Prob(X=1,Y=1)}{Prob(Y=1)} = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$

Thus
$H(X|Y) = \frac{1}{4}H(X|Y=1) + \frac{1}{4}H(X|Y=2) + \frac{1}{4}H(X|Y=3) + \frac{1}{4}H(X|Y=4) =$
$\frac{1}{4}H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4},\frac{1}{2},\frac{1}{8},\frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right) + \frac{1}{4}H(1,0,0,0) = \frac{1}{4}\cdot 1.75 + \frac{1}{4}\cdot 1.75 + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 0 = \frac{11}{8}$

We checked that $\frac{27}{8} = H(X,Y) = H(X|Y) + H(Y) = \frac{11}{8} + 2$

We stated but haven't proved the next property.

Property 2

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

Home-work

Find $H(Y|X)$, $H(Y) - H(Y|X)$ and $H(X) - H(X|Y)$