Fifth Lecture

October 9, 2023

Consider the last homework of the previous lesson.

In order to calculate the conditional entropy
$H(Y|X) = \sum_x p(x)H(Y|X=x) = \frac{1}{2}H(Y|X=1) + \frac{1}{4}H(Y|X=2) + \frac{1}{8}H(Y|X=3) + \frac{1}{8}H(Y|X=4)$, we need the conditional distributions of Y. Those we can get from the joint and the marginal distributions. Eg. $Prob(Y=1|X=2) = \frac{Prob(X=2,Y=1)}{Prob(X=2)} = \frac{\frac{1}{16}}{\frac{1}{4}} = \frac{1}{4}$

Thus
$H(Y|X) = \frac{1}{2}H(Y|X=1) + \frac{1}{4}H(Y|X=2) + \frac{1}{8}H(Y|X=3) + \frac{1}{8}H(Y|X=4) =$
$\frac{1}{2}H\left(\frac{1}{4},\frac{1}{8},\frac{1}{8},\frac{1}{2}\right) + \frac{1}{4}H\left(\frac{1}{4},\frac{1}{2},\frac{1}{4},0\right) + \frac{1}{8}H\left(\frac{1}{4},\frac{1}{4},\frac{1}{2},0\right) + \frac{1}{8}H\left(\frac{1}{4},\frac{1}{4},\frac{1}{2},0\right) = \frac{1}{2}\cdot 1.75 + \frac{1}{4}\cdot 1.5 + \frac{1}{8}\cdot 1.5 + \frac{1}{8}\cdot 1.5 = \frac{13}{8}$

Check that $\frac{27}{8} = H(X,Y) = H(Y|X) + H(X) = \frac{13}{8} + \frac{7}{4}$

Let's prove that it is always true, the property that we stated, but didn't prove last time:

Property 2

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

Proof

$H(X,Y) =$

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

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

□

Property 3

$$0 \le H(X|Y) \le H(X).$$

Proof. $0 \le H(X|Y)$ follows from observing that $H(X|Y)$ is the expected value of entropies that are non-negative by $0 \le H(X)$ being valid in general. $H(X|Y) \le H(X)$ follows from the previous property: $H(X,Y) = H(X|Y) + H(Y)$, while we have already seen that $H(X,Y) \le H(X) + H(Y)$. This also gives that the condition of equality is exactly the same as it is in Property 1, namely that X and Y are independent. □

Now we prove a consequence of this.

Property 4 *For any function $g(X)$ of a random variable X we have*

$$H(g(X)) \le H(X).$$

Proof. Since $g(X)$ is determined by X we have $H(g(X)|X) = 0$. Thus using Theorem 3 a) we can write

$$H(X) = H(X) + H(g(X)|X) = H(X, g(X)) = H(g(X)) + H(X|g(X)) \ge H(g(X)).$$

We also see that the condition of equality is $H(X|g(X)) = 0$, which is equivalent to $g(X)$ determining X, i.e. to g being invertible. □

Then turn our attention back to the homework, and found that $H(X) - H(X|Y) = \frac{7}{4} - \frac{11}{8} = \frac{3}{8}$ and $H(Y) - H(Y|X) = 2 - \frac{13}{8} = \frac{3}{8}$, thus $H(X) - H(X|Y) = H(Y) - H(Y|X)$

It is always true, since from Property 2 we see that

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

This quantity is thus intuitively the difference of the amount of information X contains if we do and if we do not know Y. We can think about it as the amount of information Y carries about X. And we see that we get the same value if we exchange the role of X and Y. This interpretation is also consistent with the fact that the above value is 0 if and only if X and Y are independent. These thoughts motivate the following definition.

Def. For two random variables X and Y, their *mutual information* $I(X,Y)$ is defined as

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

By the foregoing we also have $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Obviously, $I(X,Y) = I(Y,X)$, and we get immediately from Property 1, that $I(X,Y) \geq 0$.

Later we will see that mutual information is a basic quantity that also comes up as a central value in certain coding theorems.

The following theorem is called *Chain rule.*

Theorem 9 (**Chain rule**)

$$H(X_1,\ldots,X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \cdots + H(X_n|X_1,\ldots,X_{n-1}).$$

Proof: Goes by induction. Clear for $n=1$, and it is just Property 2 for $n=2$. Having it for $n-1$, apply Property 2 for $Y = X_n$ and $X = (X_1,\ldots,X_{n-1})$ in the form $H(X,Y) = H(X) + H(Y|X)$. It gives

$$H(X_1,\ldots,X_n) = H((X_1,\ldots,X_{n-1}),X_n) =$$

$$H(X_1,\ldots,X_{n-1}) + H(X_n|(X_1,\ldots,X_{n-1})) =$$

$$H(X_1) + H(X_2|X_1) + \cdots + H(X_{n-1}|X_1,\ldots,X_{n-2}) + H(X_n|X_1,\ldots,X_{n-1}).$$

□

Problems and exercises

Exercise (Taken from the book: Thomas Cover and Joy Thomas *Elements of Information Theory*, Second Edition, Wiley, 2006.)

The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; examples of possible values of X are $AAAA, BABABAB$, and $BBBAAAA$. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched (that is, both has a 50% chance to win at each game) and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, $H(X|Y)$, $H(X,Y)$ and $I(X,Y)$

Solution: Since X determines Y, we can immediately write $H(Y|X) = 0$.
Also, since $H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) = H(X)$, we get $H(X,Y)$ once we know $H(X)$ and we will be able to calculate $H(X|Y)$ as the difference between $H(X)$ and $H(Y)$.
Furthermore, $I(X,Y) = H(Y) - H(Y|X) = H(Y)$.
So it will be enough to calculate $H(X)$ and $H(Y)$.
The value of Y can be 4 in two different ways ($AAAA$ and $BBBB$), both with probability $\frac{1}{2^4}$, so $\text{Prob}(Y=4) = 1/8$.
Y can be 5 in 8 different ways: there are 4 ways to have one game won by B among the first four games the fifth of which is won by A, and there are four other options when we change the role of A and B.

Each of these options have probability $\frac{1}{2^5}$, thus $\text{Prob}(Y = 5) = 1/4$.
There are $\binom{5}{2} = 10$ ways to have exactly two games won by B among the first five while the sixth is won by A and we have 10 more options to have six games changing the role of A and B. Each of these have probability $\frac{1}{2^6}$. So $\text{Prob}(Y = 6) = \frac{20}{2^6} = 5/16$.
Similarly, we have $2 \cdot \binom{6}{3} = 40$ options to have $Y = 7$, each with probability $\frac{1}{2^7}$. So $\text{Prob}(Y = 7) = \frac{40}{2^7} = 5/16$.
Thus $H(Y) = \frac{1}{8}\log 8 + \frac{1}{4}\log 4 + 2 \cdot \frac{5}{16}\log\frac{16}{5} = \frac{3}{8} + \frac{1}{2} + \frac{40}{16} - \frac{5}{8}\log 5 = \frac{27}{8} - \frac{5}{8}\log 5$.
To calculate $H(X)$ note that each particular outcome of i games have probability $\frac{1}{2^i}$. So the distribution of X contains the value $\frac{1}{16}$ twice, $\frac{1}{32}$ 8 times, $\frac{1}{64}$ 20 times, and $\frac{1}{128}$ 40 times.
Thus $H(X) = \frac{2}{16}\log 16 + \frac{8}{32}\log 32 + \frac{20}{64}\log 64 + \frac{40}{128}\log 128 = 1/2 + 5/4 + 30/16 + 35/16 = \frac{93}{16}$.
Thus we obtained

$$H(X) = \frac{93}{16}, \quad H(Y) = \frac{27}{8} - \frac{5}{8}\log 5, \quad H(X, Y) = H(X) = \frac{93}{16}$$

$$H(X|Y) = \frac{93}{16} - (\frac{27}{8} - \frac{5}{8}\log 5) = \frac{39}{16} + \frac{5}{8}\log 5,$$

$$H(Y|X) = 0, \quad I(X, Y) = H(Y) = \frac{27}{8} - \frac{5}{8}\log 5.$$

◊

We continued with the other two home works that I had given the previous time.

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

Solution: The code in part (a) is not even prefix (the fourth codeword starts with the second one), so it cannot be a Huffman code.

The code in part (b) cannot have optimal average length because deleting the last digit of the last codeword we still have a prefix code. Therefore this code is also cannot be a Huffman code.

The code in part (c) can be a Huffman code, to prove this we give a distribution P for which it can belong. Let $P = (1/2, 1/8, 1/8, 1/8, 1/8)$. Performing the construction for this distribution shows that it results in the code given in (c) if we choose the binary digits accordingly. (One can also easily check that the average length is exactly $H(P)$ so it must be optimal as it cannot be less. ◊

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 $p_3 = 1/6$.

Solution: When constructing the code, in the first step both people had to create the distribution $(p_1, p_2, p_3 + p_4)$. In the next step, however, one of them had to create $(p_1 + p_2, p_3 + p_4)$, while the other created $(p_1, p_2 + p_3 + p_4)$. If both led to Huffman codes, that means both steps are optimal, so it did not matter whether we add p_1 or $p_3 + p_4$ to p_2. This implies $p_1 = p_3 + p_4$. All the probability values can be obtained from this as follows.
Since $p_4 \leq p_3 = 1/6$, we have $p_1 = p_3 + p_4 \leq 1/6 + 1/6 = 1/3$. On the other hand, $p_1 \geq \frac{1}{2}(p_1 + p_2) = \frac{1}{2}(1 - (p_3 + p_4) \geq \frac{1}{2}(1 - 1/3) = 1/3$. So p_1 is not more and not less than $1/3$, thus $p_1 = 1/3$. Thus $p_4 = p_1 - p_3 = 1/3 - 1/6 = 1/6$ and $p_2 = 1 - 1/3 - 1/6 - 1/6 = 1/3$. So the distribution is $(1/3, 1/3, 1/6, 1/6)$. Note that the above two codes give indeed the same optimal average length 2 for this distribution. ◊

Then some more excercises followed:

Exercise 3 What is the largest integer value of ℓ for which a prefix code of 8 codewords with respective lengths $1, 2, 3, 4, 5, 6, 7$, and ℓ over a binary alphabet does not exist? (Notice that we do not assume anything about the relation between the value of ℓ and the other given lengths.)

Solution: By the theorems of McMillan and Kraft such a code does not exist if and only if

$$\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^6} + \frac{1}{2^7} + \frac{1}{2^\ell} > 1.$$

This is the case if and only if $\ell \leq 6$ (we would have equality for $\ell = 7$). So the requested largest number is $\ell = 6$. $\Diamond$

Exercise 4 We have two dice with 1 dot on two faces, 2 dots on two faces, and 3 dots on two faces. We roll the two dice together and want to encode the total number of dots we see on the rolled faces of the two dice. Give the Shannon-Fano code for alphabet size 2 and also for alphabet size 3 for this problem. Construct also a binary and a ternary code that has shortest average length, that is one, for which the expected number of bits (ternary digits) needed to encode the result of many rolls is as small as possible.

Solution: The result can be $2, 3, 4, 5$, or 6 dots on the two faces seen, and their probabilities can be calculated by the number of elementary events giving the corresponding number. So the probabilities are $1/9, 2/9, 3/9, 2/9, 1/9$, respectively. Then the probabilities have to be arranged in decreasing order, and the corresponding w_i values are $0, 3/9, 5/9, 7/9, 8/9$ in the Shannon-Fano code construction.
The binary Shannon-Fano code we obtain from these values is $(00, 01, 10, 110, 111)$. More precisely the codeword for 4 is 00, the codeword for 3 is 01, the codeword for 5 is 10, the codeword for 1 is 110 and the codeword for 6 is 111.
The ternary Shannon-Fano code is: $(0, 10, 12, 21, 22)$, of course 4 has the shortest codeword.
To obtain shortest average length we have to construct a Huffman code for the above distribution. Doing this we can get $00, 10, 11, 010, 011$ in the binary case, and we can get $0, 1, 20, 21, 22$ in the ternary case.
(This shows that the binary Shannon-Fano code also has optimal average length in this case, while the ternary doesn't.) $\Diamond$