Information Theory

Problem Set 03 - The Source Coding Theorem

Luís Felipe Ramos Ferreira

lframos_ferreira@outlook.com

1. (a) The Shannon information content $h(x)$ of the outcome x of a random experiment is defined as:

$$h(x) = log_2 \frac{1}{p_x}$$

, where p_x represents the probability of observing the outcome x. The value $h(x)$ means, in a general sense of the word, the amount of information we gain about the state of the world after running the experiment and obtaining the result x. We can also say it measures the amount of uncertainty of that outcome. If it's probability p_x is big, there is no "surprise" when the result x is obtained, but if p_x is small, you get very "surprised" with that information.

(b) The entropy $H(X)$ of an ensemble X is defined as:

$$H(X) = \sum_{x \in X} p_x log_2 \frac{1}{p_x}$$

The value $H(X)$ is a measure of the average information content of the ensemble X. As the formula states, it is a weighted average of the Shannon information content of each outcome x of the emsemble, where the weight of each outcome is it's probability.

(c) A convex function is a function where, for any pair of points in the graph of the function, the line between those points lies above the graph between the points. In more formal therms, as stated by MacKay in [1], a function f is convex if for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

For example:

i. $f(x) = x^2$ is convex over $(-\infty, \infty)$;

ii. $f(x) = e^x$ is convex over $(-\infty, \infty)$;

iii. $f(x) = sin(x)$ is not convex over $(-\infty, \infty)$;

iv. $f(x) = x^3$ is not convex over $(-\infty, \infty)$;

(d) Jensen's inequality states that if f is a convex function and x is a random variable then:

$$\mathcal{E}[f(x)] \geq f(\mathcal{E}[x])$$

, where $\mathcal{E}$ denotes expectation.

(e) The formula for the raw bit content of an ensemble X is

$$H_0(X) = log_2 \mid \mathcal{A}_X \mid$$

, where $\mathcal{A}_X$ is the set of possible outcomes of the ensemble X. It represents, in general, the number of binary questions that are needed to identify an outcome x from X for sure. It can also be seen as the smallest length necessary to map each outcome of X to a binary string.

(f) The smallest δ-suficcient subset S_δ of an ensemble X is the smallest subset of outcomes of X such that the probability of any outcome x from X not belong do S_δ is less than or equal to δ, i.e., $P(x \notin S_\delta) \leq \delta$ for all $x \in \mathcal{A}_X$.

(g) The essential bit content of X is $H_\delta(X) = log_2 \mid S_\delta \mid$. It means that if we are willing to accept a probability of error δ, than we can compress the source into $H_\delta(X)$ bits per symbol. As we can see, if $\delta = 0$, i.e., no error is tolerated, the value is exactly the raw bit content of the ensemble.

(h) Without getting to much involved in the mathematical details, the Shannon's source coding theorem states that the entropy of the ensemble X is the theorical limit of the amount of bits per symbol you can use to compress the values of your source. So, when compressing a random variable with entropy $H(X)$, a compression using no less than $H(X)$ bits can be done with negligible risk of information loss. When less than $H(X)$ bits are used, information will be lost.

2. As stated before, the entropy $H(X)$ of an ensemble can be defined as:

$$H(X) = \sum_{x \in X} p_x log_2 \frac{1}{p_x}$$

, where p_x represents the probability of observing the outcome x. Note that, for each outcome x, we have $p_x \geq 0$, which implies that $p_x log_2 \frac{1}{p_x} \geq 0$ (For convention, we assume $0 log_2 \frac{1}{0}$ to be 0, since it's where the limit goes). Therefore, for each outcome, there is a contribution to $H(X)$ that is greater or equal to zero. At last, we have the sum over the contribution of every outcome, values that are greater or equal to zero, which is also greater or equal to zero, so $H(X) \geq 0$.

3. (a)

$$\mathcal{E}(f(x)) = p_a f_a + p_b f_b + p_c f_c = 0.1 * 10 + 0.2 * 5 + 0.7 * \frac{10}{7}$$

$$\mathcal{E}(f(x)) = 1 + 1 + 1 = 3$$

$$\mathcal{E}(\frac{1}{P(x)}) = p_a \frac{1}{p_a} + p_b \frac{1}{p_b} + p_c \frac{1}{p_c} = 1 + 1 + 1 = 3$$

(b) For an arbitrary ensemble X, $\mathcal{E}(\frac{1}{P(x)})$ is exactly the number of possible outcomes of X.

$$\mathcal{E}(\frac{1}{P(x)}) = \sum_{x \in \mathcal{A}_X} P(x) \frac{1}{P(x)} = \sum_{x \in \mathcal{A}_X} 1 = |\mathcal{A}_X|$$

(c) Let $f(x) = log \frac{1}{x}$ and $x = \frac{1}{P(x)}$. This gives the following, using Jensen's inequality:

$$H(X) = -\mathcal{E}[f(\frac{1}{P(x)})] \leq -f(\mathcal{E}[\frac{1}{P(x)}])$$

Since $\mathcal{E}[\frac{1}{P(x)}] = |\mathcal{A}_X|$, we have:

$$H(X) \leq -f(|\mathcal{A}_X|) = log|\mathcal{A}_X|$$

4. The problem clearly states a geometric distribution, i. e., the probability distirbution of the number of Bernoulli trials needed to get one success. In this scenario, getting heads is considered success. Let's consider $P_{heads} = P_{tails} = \frac{1}{2}$. For a geometric distribution, we have that the probability of the number of flips required until the first head occurs to be:

$$P(X = k) = (1 - P_{tails})^{k-1} P_{heads} = (\frac{1}{2})^k$$

Therefore, we have:

$$H(X) = \sum_{k=0}^{\infty} P(X = k) log_2 \frac{1}{P(X = k)} = \sum_{k=0}^{\infty} (\frac{1}{2})^k log_2 \frac{1}{(\frac{1}{2})^k}$$

$$H(X) = \sum_{k=0}^{\infty} (\frac{1}{2})^k k = \frac{\frac{1}{2}}{(1 - \frac{1}{2})^2} = \frac{\frac{1}{2}}{\frac{1}{4}} = 2$$

5. If X and Y are independent, we have $P(x, y) = P(x)P(y)$. From there, we can show:

$$h(x, y) = log \frac{1}{P(x, y)}$$

$$h(x,y) = log\frac{1}{P(x)P(y)}$$
$$h(x,y) = log\frac{1}{P(x)} + log\frac{1}{P(y)}$$
$$h(x,y) = h(x) + h(y)$$

Now, with this result, we can show:

$$H(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x,y)h(x,y)$$
$$H(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x,y)(h(x) + h(y))$$
$$H(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x,y)h(x) + \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x,y)h(y)$$
$$H(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x)P(y)h(x) + \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x)P(y)h(y)$$
$$H(X,Y) = \sum_{x \in \mathcal{A}_X} P(x)h(x) \sum_{y \in \mathcal{A}_Y} P(y) + \sum_{y \in \mathcal{A}_Y} P(y)h(y) \sum_{x \in \mathcal{A}_x} P(x)$$
$$H(X,Y) = \sum_{x \in \mathcal{A}_X} P(x)h(x) + \sum_{y \in \mathcal{A}_Y} P(y)h(y)$$
$$H(X,Y) = H(X) + H(Y)$$

6. No, it couldn't. It's easy to see that based on the pidgeonhle principle. If every possible outcome were compressed into a binary code of length shorter than $H_0(X)$, then some outcomes would be mapped to the same binary code, which configures a type of lossy compressor. Therefore, it wouldn't be possible to map c back to x with 100% reliability, data would be lost during the process of encoding.

7. They are right because the probability of each ball being the odd one stills the same, $\frac{1}{12}$, but they are also wrong because we know, for sure, that either one of the six balls on the lighter set is the odd one, an it's lighter than the other, or one of the six balls in the heavier side is the odd one, and it's heavier than the others. Since after the experiment the probability os each ball being the odd one remains the same, the information gained about *which is the odd ball* is zero. On the other hand, the information about *which is the odd ball and whether it is geavy or light* is not zero, as we can see below.

References

[1] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms.* 7th edition, 2005.