TD 1 – Shannon theory

Mathematics of data

18/09/24

Exercise 1. Consider an alphabet (s_1, s_2, s_3, s_4). The probabilities of appearance of the symbols s_k are $p_1 = 1/2$, $p_2 = 1/8$, $p_3 = 1/4$, $p_4 = 1/8$.

1. Compute the entropy $H(p)$ of the distribution of the considered alphabet.
2. If one were to define a fixed length binary code for this alphabet, how many bits would be needed to encode each symbol ?
3. Draw a Huffmann tree for the considered alphabet, explaining each step. Write the associated Huffmann code. What is its average number of bits per symbol ? Is it optimal ?

Exercise 2. Assume X is a discrete random variable with values in $\{1, \ldots, k\}$ and with probability distribution $p = (p_1, \ldots, p_k)$.

1. What is the probability distribution $q = (q_{i_1,\ldots,i_n})_{i_1,\ldots,i_n}$ of the random vector $(X_1, \ldots, X_n)$ on $\{1, \ldots, k\}^n$, where the X_i are independent copies of X ?
2. Compute the entropy $H(q)$ of q as a function of $H(p)$.
3. Consider an infinite sequence of independent symbols of distribution p. Show that by using a Huffman code on blocks of n consecutive symbols, the average number of bits *per symbol* tends to $H(p)$ as $n \to \infty$.

Exercise 3. Let $H \colon p \in \mathbb{R}_+^n \mapsto -\sum_{k=1}^n p_k \log_2 p_k$ be the entropy function. Denote $\Sigma_n \coloneqq \{(p_1, \ldots, p_n) \in [0,1]^n : \sum_{k=1}^n p_k = 1\}$ the simplex.

1. Show that H is strictly concave. (You may want to compute the Hessian of H.)
2. For which value of $p \in \mathbb{R}_+^n$ is H maximal ?
3. For which value of $p \in \Sigma_n$ is H maximal ?
4. For $(p, q) \in \Sigma_n \times \Sigma_m$, compute $H(p \otimes q)$, where $p \otimes q \coloneqq (p_i q_j)_{1 \le i \le n, 1 \le j \le m}$.

Exercise 4. Huffmann coding is natural to encode sequences of symbols that are independent of each other. Let us now consider the case where the distribution of each symbol depends on the previous one. Let X be a random process on $\{0,1\}^N$ such that $P(X_0 = 0) = 1/10$, $P(X_0 = 1) = 9/10$, the conditional law $X_{n+1}|X_n$ does not depend on $(X_0, \ldots, X_{n-1})$ and

$$
\begin{aligned}
P(X_{n+1} = 0|X_n = 0) &= 1/2 \qquad & P(X_{n+1} = 0|X_n = 1) &= 5/90 \\
P(X_{n+1} = 1|X_n = 0) &= 1/2 \qquad & P(X_{n+1} = 1|X_n = 1) &= 85/90.
\end{aligned}
$$

1. Show that for all $n \in \{0, \ldots, N-1\}$ it holds $P(X_n = 0) = 1/10$ and $P(X_n = 1) = 9/10$. What is the best average number of bits per symbol you can hope for with plain Huffmann coding ?
2. How could you extend Huffmann coding to take into account the correlations between symbols ?

Exercise 5. *Recall Shannon's reconstruction formula :*

$$
f(x) = \sum_n f(ns)\,\mathrm{sinc}(x/s - n) \tag{1}
$$

under some regularity conditions on f (nice decay at infinity and compact support of the Fourier transform), where $\mathrm{sinc}(x) \coloneqq \frac{\sin(\pi u)}{\pi u}$. *One issue with this formula is that the kernel* sinc *oscillates a lot and has slow decay, so that lots of terms in the sum (1) are non negligible. In practice, people use smoother and more localized kernels, such as splines (piecewise polynomials). This exercise is a short introduction to spline interpolation.*

Let $\varphi_0 \coloneqq \mathbf{1}_{[-1/2,1/2]}$ and $\varphi_k(x) \coloneqq \varphi_{k-1} * \varphi_0$, where $*$ denotes the convolution operator.

1. Show that $\mathrm{Supp}\,\varphi_k \subset [-\frac{k+1}{2}, \frac{k+1}{2}]$, that φ_k is a piecewise polynomial of degree k and that φ_k is $\mathcal{C}^{k-1}$.
2. Draw φ_0, φ_1 and φ_2 (approximately, for φ_2).
3. Let φ be one of the splines φ_k. The associated spline interpolation of a function f is of the form

$$
f \approx \tilde{f} \coloneqq \sum_n a_n \varphi(\cdot - n)
$$

where the coefficients a_n are set to have $f(n) = \tilde{f}(n)$ for all $n \in \mathbb{Z}$. What are the a_n if $\varphi = \varphi_0$? If $\varphi = \varphi_1$?