Workshop the Minimum Description Length Principle

Exercises

Paul van Leeuwen (paul2.vanleeuwen@devolksbank.nl)
Eelko Ubels (eelko.ubels@devolksbank.nl)

13 June 2023

Introduction

Exercises are meant as an introduction to the Minimum Description Length (MDL) principle and are by no means meant to stop here. They are merely for illustrative purposes; in (academic) applications they are more advanced. For a decent introduction into MDL, see [1]. Presumed knowledge: basic calculus, presentation of the MDL principle. Results of exercises could be used in subsequent exercises. Challenging exercises are preceded by the word Hint in parentheses. The hints themselves are given on a separate page after the exercises.

1 Morse Code

In 1840 Samuel Morse faced a similar problem as we do: how to code the Western alphabet $\mathbb{B} = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}, 0, 1, \ldots, 9\}$ such that on expectation the smallest amount of bits is needed for whatever message. Every symbol is a concatenation of *dits* (one hit with the sending device with a fixed duration represented by a dot $\cdot$) and *dahs* (three times the duration of a *dit* represented by a bar $-$). For example, the letter A is a dit followed by a dah ($\cdot -$). See Figure 1 for the complete mapping of the $\mathbb{B}$ to the Morse symbols.[1]

1. You can assign a single dit to just one symbol. What letter would you recommend?
2. A simple alternative is to have a fixed code length for every symbol. How many bits would we need to represent all elements of $\mathbb{B}$ in a binary way?
3. Samuel did not choose for this simple alternative but aimed to minimise the *expected* code length. Why?

2 Design Your Own Code

Suppose you are given the alphabet $\mathbb{B} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}\}$ with corresponding observed frequencies $(15, 7, 6, 6, 5)$.[2] So in total 39 symbols are communicated. In this exercise you will design a coding scheme $C\colon \mathbb{B} \mapsto \{0,1\}^m$ with at most m symbols and $m \in \mathbb{N}_+$ chosen by you. A possible mapping of the symbol B could be $C(\mathrm{B}) = 101$. Note that in the exercises below you do not need to stick with $C(\mathrm{B}) = 101$.

[1] Retrieved from [2].
[2] Example taken from [3].

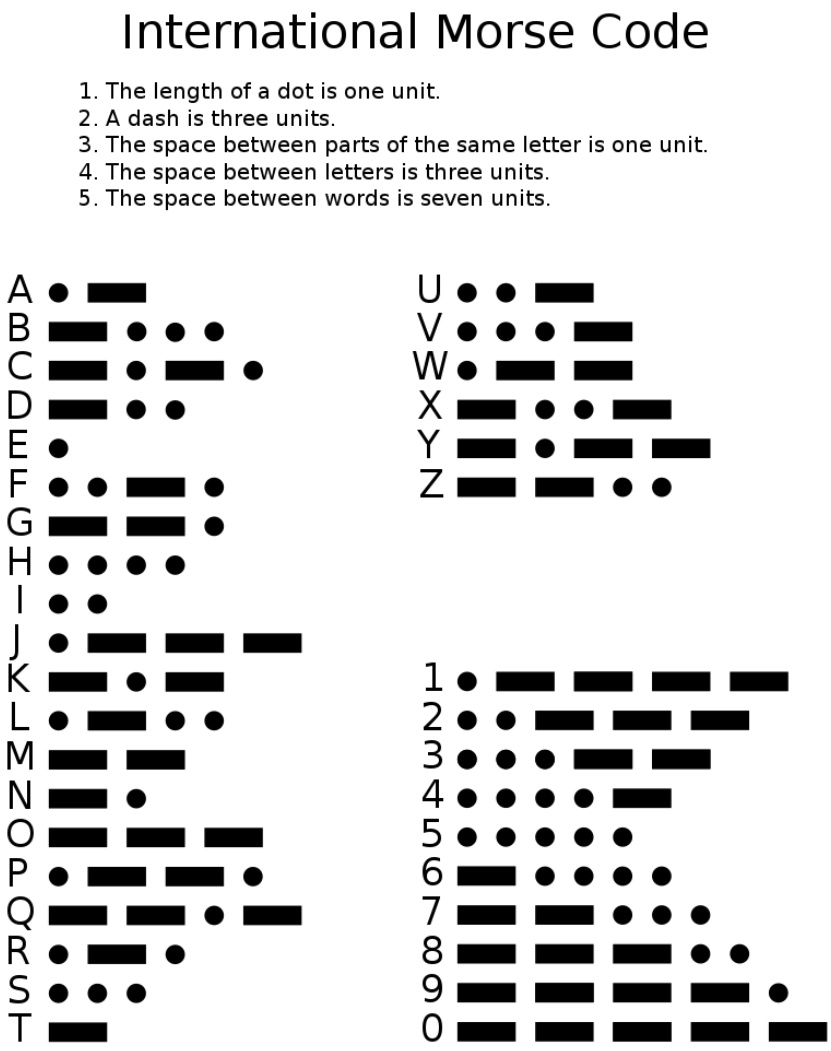


Figure 1: International Morse Code of the Western alphabet and the 10 digits.

1. Code the alphabet $\mathbb{B}$ using a uniform coding scheme: every symbol uses the same number of bits.
2. What is the expected code length using this coding scheme?

Whatever concatenation of zeros and ones (or dits and dahs in Morse code) you assign to every symbol, the resulting code should be *prefix*: reading from left to right always results in the same interpretation. In practice this means that, reading from left to right, no code should be a part of another code. Let $C\colon \mathbb{B} \mapsto \{0,1\}^m$ be the mapping from an element of the alphabet $\mathbb{B}$ to a concatenation of at most m zeros and ones. An example of bad coding is the following: $C(\mathrm{A}) = 10$ and $C(\mathrm{C}) = 100$ is not prefix since the 10 may symbolise A or C followed by a 0.

3. Is Morse-code a prefix code? See Figure 1 for the complete mapping of the $\mathbb{B}$ of Exercise 1 to the Morse symbols.
4. (Hint) We can do better when the probability a symbol shows up is different from another. Assume the observed frequencies are representative for all future messages to be coded. What would be a more efficient coding scheme? Make sure the code is prefix.

3 From Code Length to Probabilities

Suppose we have a message of length $k \in \mathbb{N}_+$ symbols from the binary alphabet $\{0,1\}$. Let p_i, for $i = 1, \ldots, k$ be the probability that symbol i will occur and let $\ell_i \in \mathbb{N}_+$ be the corresponding code length; $\sum_{i=1}^{k} p_i = 1$. Then the expected code length is

$$\mathbb{E}[L] = \sum_{i=1}^{k} p_i \ell_i$$

The expected code length, using the Shannon-Fano code $\ell_i = -\log_2 p_i$, is given by

$$E = \sum_{i=1}^{k} -p_i \log_2 p_i \tag{1}$$

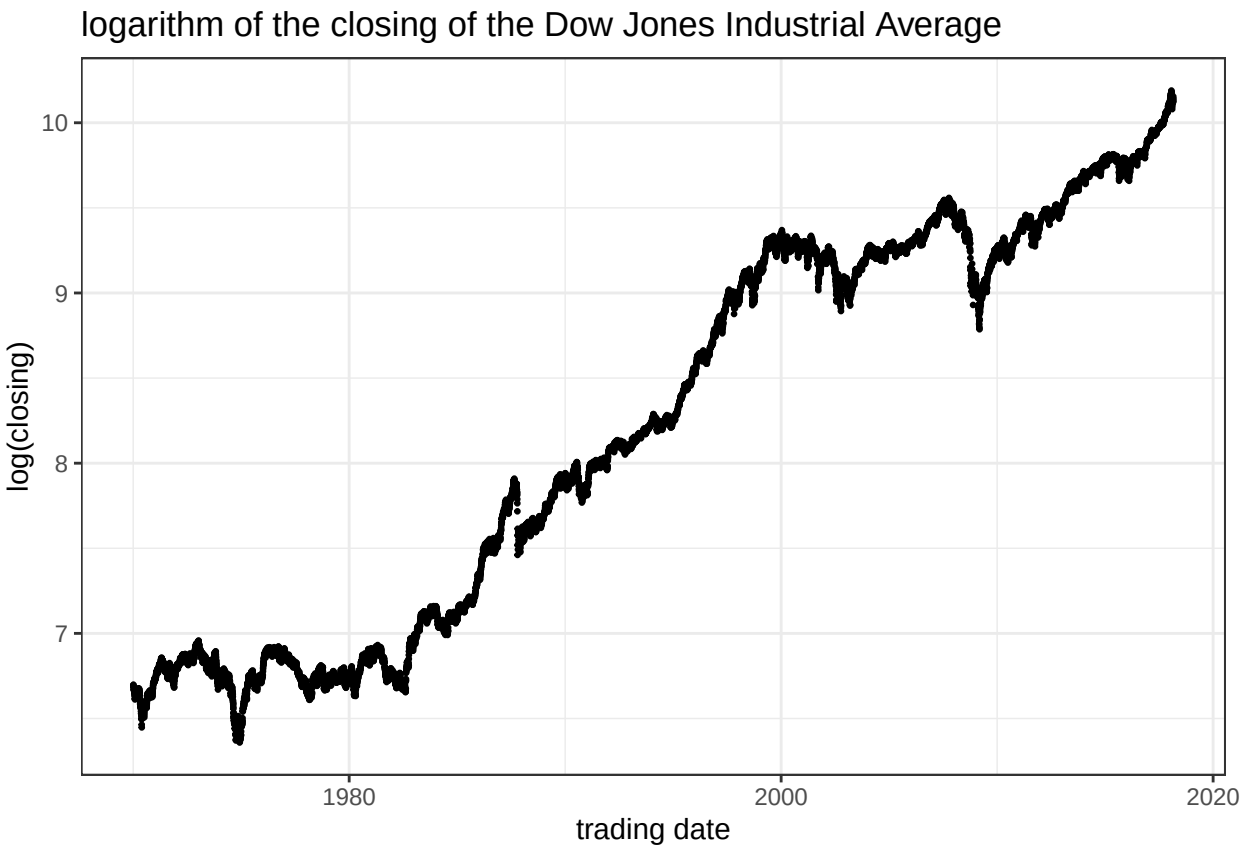
We could code the entire message them symbol by symbol, processing them from left to right, without taking the complete sequence of symbols into account. This is as if the complete message of length k can be broken down into k separate submessages. Alternatively, we could assign a symbol to every possible sequence of $\{0,1\}^k$.

1. Suppose that we code the total message symbol by symbol. Explain why the choice of $\ell_i = -\log_2 p_i$ is not always realistic to accurately represent the code length of all corresponding p_i.
2. (Hint) Choose $\ell_i = \lceil -\log_2 p_i \rceil$. Derive an upper bound for $\mathbb{E}[L]$ in terms of E.
3. Using $\ell_i = -\log_2 p_i$ we have a connection between the probability that event i occurs, where event i is represented by symbol i, and code length ℓ_i. Although this choice of ℓ_i does not give realistic code lengths, explain why it is useful after all.

4 From Data to Parameter Estimation

In this exercise we will use the same approach as in [4]. They model the Dow Jones Industrial Average (DJIA) using different models where they choose the winning model using MDL. For every trading day the

closing of the DJIA is denoted by $P_t \in \mathbb{R}_+$ where t runs from $t_0 =$ January 1970 until $T = 12$ June 1995, i.e. 6,430 trading days.[3] See the next figure for a visualisation of P_t.



Take, for $t \geq 1$, the log daily return $R_t = P_t - P_{t-1}$ and the volatility $V_t = 0.9V_{t-1} + 0.1R_t^2$ with V_0 the variance of the series P_t. R_t has a corresponding indicator: 1 (0) when $P_t > P_{t-1}$ ($P_t < P_{t-1}$), analogously for V_t. This results in two binary strings of length 6,430 - 1 = 6,429. R_t has 3,166 (49.25%) ups and V_t has 3,166 (30.55%) ups. As with Morse code, the sender and receiver need to agree upfront how the events are coded to a concatenation of zeros and ones. We do not have prior knowledge on the outcomes of the indicators of R_t and V_t.

1. As a baseline we consider the uniform coding scheme. How many bits do we require to code the ups and downs of R_t? And how many for the ups and downs of V_t?

2. As before, we can do better when the probability of one event is sufficiently different from $\frac{1}{2}$. For what time series R_t or V_t do you expect, *a priori*, to realise the highest compression?

Two-part MDL is, among the possible MDL methods, a simplistic method to measure the level of compression. The two refers to two stages: (i) describe the data description mechanism H with code length $L(H)$, and (ii) given H describe the data D with code length $L(D|H)$. The sender, or coder, agrees with the receiver, or decoder, to model the ups and downs of R_t and V_t with a Bernoulli model. So before the data can be coded, first the data description mechanism needs to be communicated from the sender to the receiver, followed by the data, i.e. the concatenation of zeros and ones.

3. Derive the maximum likelihood estimator $\hat{p}$ for p.

[3] Note that [4] start at June 1962 but that data is unavailable via standard retrieval methods. Therefore the end date 12 June 1995 is chosen such that the total same number of trading days over this period is the same as in [4].

4. (Hint) How many bits do you need to communicate H, the maximum likelihood estimate $\hat{p}$, using a uniform coding scheme?

5. (Hint) Given that $\hat{p}$ is communicated, how many bits do you need to communicate the ups and downs of R_t and V_t, i.e. $L(D|H)$?

6. Does that result in an improvement relative to the number of bits required for the uniform coding scheme as derived in (1)?

5 Markov Modelling

An improvement over the Bernoulli model could be a Markov model; more precise: a time-discrete Markov chain. In this way we are able to capture, if present, time-dependency in the ups and downs of R_t and V_t. Let $C_t \in \{0, 1\}$ denote the event that $R_t > R_{t-1}$ with realisations $c_t \in \{0, 1\}$. The Markov property states that the probability of an up on the next day only depends on the realisation on the previous k values: $\mathbb{P}[C_{t+1} = 1 | C_t = c_t, \ldots, C_{t-k} = c_{t-k}]$ for $t = 1, \ldots, T$ and $k \geq 0$. The probability transition matrix $\mathbf{P}$ for k lags is a matrix of size $\left(2^k \times 2\right)$ where $\mathbf{P}_{ij}$ is the probability at row i and column j with $i = 1, \ldots, 2^k$ and $j = 1, 2$. In this case, i ranges over all possible routes the lagged target values can take and j indicates whether the target value on the next day is a 0 ($j = 1$) or a 1 ($j = 2$). By definition, the probabilities of one row over all possible columns should be 1:

$$\sum_{j=1}^{2} \mathbf{P}_{ij} = 1$$

For example, for $k = 2$ we have

$$\mathbf{P} = \begin{pmatrix} p_{00\to0} & p_{00\to1} \\ p_{01\to0} & p_{01\to1} \\ p_{10\to0} & p_{10\to1} \\ p_{11\to0} & p_{11\to1} \end{pmatrix}$$

1. Argue that the Bernoulli model is a nested model of the k^{th}-order Markov model family for $k \in \mathbb{N}$. In other words, what is k for the Bernoulli model?

2. In what case would a higher-order Markov model be superior over a Bernoulli model?

3. Study the autocorrelation function of $R_t - R_{t-1}$ and $V_t - V_{t-1}$. In Matlab you may want to use the function `autocorr`. The realisations of R_t and V_t can be extracted from the Excel-file that was sent to you by mail by making use of the Matlab-function `readtable`.

4. (Hint) Derive the maximum likelihood vector of model parameters for the first-order Markov model.

5. To code a first-order Markov model, we need to first code the data description mechanism H with code length $L(H)$. Then, given H, give the code length of the data D, which is $L(D|H)$.

6. This two-stage approach is rather simplistic. Give suggestions on how to improve this approach.

6 Hints

(2.4)
Hint 1: First calculate the probability a certain symbol occurs by relating the frequency of occurrence to the total number of symbols.

Hint 2: The higher the probability a symbol occurs, the smaller the code length can be. So assign the smallest code, which is thus a concatenation of zeros and ones, to A and the largest code to E.

(3.2) $\lceil -\log_2 p_i \rceil \leq -\log_2 p_i + 1$

(4.4) What possible values could $\hat{p}$ take?

(4.5) We are interested in the expected value of L. $\mathbb{E}[L]$ can be approximated by E, see Equation (1), with $\hat{p} = k/n$ and summing over all n events that end in an up or a down.

(5.4) The Markov model has two model parameters: $p_{0\to0}$ and $p_{1\to0}$ since $p_{0\to1} = 1 - p_{0\to0}$ and $p_{1\to1} = 1 - p_{1\to0}$.

Bibliography

[1] P. D. Grünwald, *The minimum description length principle*. MIT press, 2007.

[2] Wikipedia, "International morse code," 2023-06-13. https://en.wikipedia.org/wiki/Morse_code#/media/File:International_Morse_Code.svg (accessed Jun. 09, 2023).

[3] Wikipedia, "Shannon-fano coding," 2023-06-09. https://en.wikipedia.org/wiki/Shannon%E2%80%93Fano_coding (accessed Jun. 09, 2023).

[4] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.