REVISITING ENUMERATIVE TWO-PART CRUDE MDL FOR BERNOULLI AND MULTINOMIAL DISTRIBUTIONS

Marc Boullé, Fabrice Clérot, Carine Hue

Orange Labs - 22300 Lannion - France
{firstname.lastname}@orange.com

ABSTRACT

We leverage the Minimum Description Length (MDL) principle as a model selection technique for Bernoulli distributions and compare several types of MDL codes. We focus on the enumerative two-part crude MDL code, suggest a Bayesian interpretation for finite size data samples, and exhibit a strong connection with the NML approach. We obtain surprising impacts on the estimation of the model complexity together with superior compression performance. This is then generalized to the case of the multinomial distributions. Both the theoretical analysis and the experimental comparisons suggest that one might use the enumerative code rather than NML in practice, for Bernoulli and multinomial distributions.

This paper is an extended abstract of the research report [1], which contains all the detailed analysis, proofs and extensive experiments.

1. INTRODUCTION

Model selection is a key problem in statistics and data mining, and the MDL approaches [2] to model selection have been extensively studied in the literature [3], with successful applications in many practical problems. Simple models such as Bernoulli or mainly multinomial distributions are important because they are easier to analyze theoretically and useful in many applications. For example, the multinomial distribution has been used as a building block in more complex models, such as naive Bayes classifiers, Bayesian networks , decision trees or coclustering models. These models involve up to thousands of multinomials blocks, some of them with potentially very large numbers of occurrences and outcomes. For example in [4], half a billion call detail records (occurrences) are distributed on one million coclusters (outcomes). These various and numerous applications critically rely on the use of effective and efficient MDL code lengths to get a robust and accurate summary of the data.

The MDL approaches come with several flavors, ranging from theoretical but not computable to practical but sub-optimal. Ideal MDL [5] relies on the Kolmogorov complexity, that is the ability of compressing data using a computer program. However, it suffers from large constants depending on the description method used and cannot be computed, not even approximated in the case of two-part codes [6]. Practical MDL leverages description methods that are less expressive than general-purpose computer languages. It has been employed to retrieve the best model given the data in case of families of parametrized statistical distributions. Crude MDL is a basic MDL approach with appealing simplicity. In two-part crude MDL, you just have to encode the model parameters and the data given the parameter, with a focus on the code length only. However, crude MDL suffers from arbitrary coding choices. Modern MDL relies on universal coding resulting in Refined MDL [3], with much stronger foundations and interesting theoretical properties. In this paper, we investigate the enumerative two-part crude MDL code for the Bernoulli and multinomial models, exhibit a strong connection with the NML approach, with surprising impacts on the estimation of the model complexity and superior compression performance.

The rest of the paper is organized as follows. Section 2 describes a particular two-part crude MDL code based on enumerations and establishes the connection of its parameter coding length with its NML parametric complexity. Section 3 summarizes comparisons between this enumerative MDL code and the standard NML code for Bernoulli distributions. Section 4 presents an extension of the enumerative two-part crude MDL code to multinomial distributions. Finally, Section 5 summarizes this paper.

2. ENUMERATIVE TWO-PART CRUDE MDL

We present the enumerative two-part crude MDL code for Bernoulli distributions, suggest a finite data sample Bayesian interpretation and show a connection with the NML approach. Let us consider the Bernoulli model with $\theta \in [0, 1]$ in the case of binary sequences $x^n \in X^n$ of size n. Let $k(x^n)$ be the number of ones in x^n.

2.1. Enumerative interpretation

The enumerative two-part crude MDL code for Bernoulli distributions has already been proposed in the past literature, under the names of *index* or *enumerative* code (see for example [3] Example 10.1 *Coding by Giving an Index*). First, we enumerate all possible $\theta = \frac{i}{n}$ parameter values given the sample size n. We then use $\log(n + 1)$ bits to encode θ. Second, given $\widehat{\theta}(x^n) = \frac{k(x^n)}{n}$, we enumerate all the $\binom{n}{k}$ binary sequences with $k = k(x^n)$ ones and encode the data x^n using $\log \binom{n}{k}$ bits. This gives a total code length of

$$L(\widehat{\theta}(x^n), x^n) = \log(n+1) + \log \frac{n!}{k!(n-k)!}. \quad (1)$$

Interestingly, this crude MDL approach results in the same code length as that obtained using *Predictive Coding* or *Mixture Coding* with a uniform prior [3]. This code has also been studied by [3] (Chapter 10, Section 10.2) under the name *Conditional Two-Part Universal Code*, which suggests that at least for the Bernoulli model, this code is strictly preferable to the ordinary two-part code.

2.2. Bayesian interpretation

Let $\mathcal{M} = \{P_\theta \,|\, \theta \in [0,1]\}$ be the class of all Bernoulli distributions. We propose to focus on the family of models $\mathcal{M}^{(n)} = \{P_\theta \,|\, \theta = \frac{i}{n},\ 0 \leq i \leq n\}$ that are models of description for finite size data samples. $\mathcal{M}^{(n)}$ is related to the set of all the possible maximum likelihood estimates of θ (from $\mathcal{M}$) for binary strings of size n. The interest of using $\mathcal{M}^{(n)}$ is that the number of model parameters is now finite instead of uncountable infinite. Using a uniform prior on the model parameters in $\mathcal{M}^{(n)}$, we get $P(\theta = \frac{i}{n}) = 1/|\mathcal{M}^{(n)}|$, leading to $L(\theta) = \log(n+1)$.

Given $\theta = \frac{i}{n} \in \mathcal{M}^{(n)}$, we now have to encode the data x^n. If $k(x^n)/n \neq \theta$, we cannot encode the data and $P(x^n|\theta) = 0$. If $k(x^n)/n = \theta$, the observed data is consistent with the model parameter, and we assume that all the possible observable data are uniformly distributed. The number of binary strings with k ones is the binomial coefficient $\binom{n}{k}$. Thus the probability of observing one of them is $P(x^n|\widehat{\theta}(x^n)) = 1/\binom{n}{k}$. We have a discrete likelihood that concentrates the probability mass on binary strings that can be observed given the model parameter. As a result, coding lengths are defined only for strings that are consistent with the model parameter. This gives a total code length of

$$L(\widehat{\theta}(x^n), x^n) = \log(n+1) + \log \frac{n!}{k!(n-k)!}. \quad (2)$$

2.2.1. Generative model for the enumerative Bernoulli distribution

Given a sequence length n and $\theta = \frac{i}{n} \in \mathcal{M}^{(n)}$, we can formulate these models as generative models of sequences with exactly i ones and $n-i$ zeros. For example, from a sequence of n zeros, we randomly choose i times without replacement a zero in the sequence and replace it with a one. For this generative model, we have the following likelihood, as seen previously:

$$P(x^n|\theta = \frac{i}{n}) = \mathbb{1}_{\{\frac{i}{n} = \frac{k(x^n)}{n}\}} 1/\binom{n}{k(x^n)}. \quad (3)$$

For the case of the Bayes mixture model with uniform prior $w(\theta) = \frac{1}{n+1}, \theta = \frac{i}{n},\ 0 \leq i \leq n$, we have

$$P_{Bayes}(x^n) = \sum_{i=0}^{n} w(\frac{i}{n}) P(x^n|\theta = \frac{i}{n}), \quad (4)$$

$$= \frac{1}{n+1} \frac{k(x^n)!(n-k(x^n))!}{n!}. \quad (5)$$

The negative log of this probability actually corresponds to the code length of the enumerative code. Interestingly, the standard Bernoulli model and the enumerative one are related to slightly different generative models, but their Bayes mixture under the uniform prior leads to the same distribution. In Section 2.3, we will see that on the opposite, their normalized maximum likelihood distribution is not the same.

2.2.2. Cardinality of models spaces

Let us consider the union of the $\mathcal{M}^{(n)}$ models for all the sample sizes:

$$\mathcal{M}^{(\mathbb{N})} = \cup_{n \in \mathbb{N}} \mathcal{M}^{(n)}. \quad (6)$$

Interestingly, $\mathcal{M}^{(\mathbb{N})}$ is very close to $\mathcal{M}$, with $\theta \in \mathbb{Q}$ rather than $\theta \in \mathbb{R}$. Thus, the number of model parameters in $\mathcal{M}^{(\mathbb{N})}$ is countable infinite rather than uncountable infinite, which provides a significant simplification.

2.3. NML interpretation

Let us compute the NML parametric complexity of this enumerative code, on the basis of the discrete likelihood presented in Section 2.2. We have

$$COMP^{(n)}(\theta) = \log \sum_{y^n \in X^n} P_{\widehat{\theta}(y^n)}(y^n), \quad (7)$$

$$= \log \sum_{k=0}^{n} \binom{n}{k} \left(1/\binom{n}{k}\right), \quad (8)$$

$$= \log(n+1). \quad (9)$$

Interestingly, we find exactly the same complexity term $\log(n+1)$ as the coding length of the best hypothesis in the enumerative two-part crude MDL code presented in Section 2.1. This shows that the enumerative code is both a two-part and a one-part code. It is parametrization invariant and optimal w.r.t. the NML approach, with minimax regret guarantee. Surprisingly, its parametric complexity is asymptotically twice that of the NML code or the standard BIC regularization term. We further investigate on the comparison between the enumerative and NML codes in next section

3. CODE COMPARISON FOR THE BERNOULLI DISTRIBUTION

Table 1. Parametric and stochastic complexity per code.

Code name	$COMP^{(n)}_{name}$	$L_{name}\left(x^n \middle| \widehat{\theta}(x^n)\right)$
enumerative	$\log(n+1)$	$\log \frac{n!}{k!(n-k)!}$
NML	$\frac{1}{2}\log\frac{n\pi}{2} + o(1)$	$\log \frac{n^n}{k^k(n-k)^{n-k}}$

In this section, we compare the standard NML code [3] and enumerative two-part crude MDL code (Section 2) for the Bernoulli distribution. Table 1 reminds the parametric and stochastic complexity of each considered code.

The theoretical and empirical comparison results presented below are a summary of the extended report [1].

3.1. Stochastic complexity term

Let $\delta L\left(x^n|\widehat{\theta}(x^n)\right) = L_{nml}\left(x^n|\widehat{\theta}(x^n)\right) - L_{enum}\left(x^n|\widehat{\theta}(x^n)\right)$.

The stochastic complexity term of the enumerative code is always smaller than that of the NML code for non-degenerated binary strings:

$$\forall n, \forall x^n \in X^n,\ 0 < k(x^n) < n, \delta L\left(x^n|\widehat{\theta}(x^n)\right) > 0. \quad (10)$$

Using the approximation given in [3] (formula 4.36) with the Bernoulli parameter $\theta = \widehat{\theta}(x^n)$, we have

$$\delta L\left(x^n|\widehat{\theta}(x^n)\right) = \frac{1}{2}\log(2\pi n \mathrm{var}(\theta)) + O(1/n). \quad (11)$$

The difference of coding length is always positive but not uniform. For $k(x^n) = 0$, $\delta L\left(x^n|\widehat{\theta}(x^n)\right) = 0$. For $k(x^n) \approx n/2$, $\delta L\left(x^n|\widehat{\theta}(x^n)\right) \approx \frac{1}{2}\log(\frac{n\pi}{2})$.

These results demonstrate that the enumerative code provides a better encoding of the data with the help of the model for any binary strings, all the more for strings with equidistributed zeros and ones.

3.2. Parametric complexity term

The parametric complexity term of the enumerative code is always strictly greater than that of the NML code and asymptotically twice it.

$$\forall n > 1, COMP^{(n)}_{enum} > COMP^{(n)}_{NML}. \quad (12)$$

$$\lim_{n\to\infty} \frac{COMP^{(n)}_{enum}}{COMP^{(n)}_{NML}} = 2. \quad (13)$$

3.3. Overall code length

Both codes have the same length for two parameter values $\{\theta_{inf}, \theta_{sup}\}$, with $\theta_{inf} \approx 0.114$ and $\theta_{sup} = 1 - \theta_{inf}$.

For heavily unbalanced Bernoulli distributions ($\theta \in [0, \theta_{inf}[\ \cup\ ]\theta_{sup}, 1]$), the NML code is shorter and for $\theta \in \{0, 1\}$, $|\delta| \approx \frac{1}{2}(\log n - \log \frac{\pi}{2})$.

For balanced Bernoulli distributions ($\theta \in]\theta_{inf}, \theta_{sup}[$), the enumerative code is shorter and for $\theta \approx \frac{1}{2}$, $|\delta| \approx \log \frac{\pi}{2}$.

3.4. Empirical comparisons

Extensive comparisons are reported in [1].

Under the uniform distribution, most binary strings are better compressed with the enumerative code and the average compression is slightly better than using the NML code, with a margin that is asymptotically about $\log \frac{\pi}{2}$.

In a biased versus fair coin classification experiment, both the NML and enumerative codes are used as classifiers by predicting a bias if they can encode a sequence with a coding length shorter than that of the random code ($n \log 2$), and predicting fair otherwise. Overall, both codes exhibit a similar behavior w.r.t. the coin classification problem, with accuracy increasing from 0.5 for small n to 1 for large n, and a slow increase rate for small bias and a fast one for large bias. Except in the tiny samples with $n \leq 20$, the difference of accuracy between the two codes never exceeds around 15%. However, there are some interesting differences. The enumerative code is better at detecting bias while the NML code is better at detecting fair, and the overall accuracy of prediction exhibits a variety of behaviors, with tiny differences. When the bias is small (θ_{bias} close from $\frac{1}{2}$), the enumerative code is slightly more accurate in the non-asymptotic case, needing less data to achieve a correct accuracy. When the bias is large (θ_{bias} close from 0 or 1), the advantage is this time in favor of the NML code. In all cases, the differences between both codes get tiny for large n, in the asymptotic case.

4. THE CASE OF MULTINOMIAL DISTRIBUTION

Let us consider the multinomial model with parameter $\theta = (\theta_1, \ldots, \theta_m)$, $\sum_{j=1}^m \theta_j = 1, \forall j, \theta_j > 0$, such that $P_\theta(X = j) = \theta_j$, in the case of m-ary sequences $x^n \in X^n$ of size n. For a given sequence x_n, $P_\theta(x_n) = \prod_{j=1}^m \theta_j^{n_j}$, where n_j is the number of occurrences of outcome j in sequence x^n.

4.1. Enumerative two-part crude MDL

Like in the Bernoulli case, the enumerative code for multinomial can be obtained using a two-part crude MDL approach, a Bayesian interpretation or a NML interpretation. We present below the Bayesian interpretation.

Given a sample size n, the number of tuples $(n_1, \ldots, n_m)$ such that $\sum_{j=1}^m n_j = n$ is $\binom{n+m-1}{m-1}$. We then encode the multinomial model parameter using a uniform prior

$$P\left(\theta = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_m}{n}\right)\right) = 1/\binom{n+m-1}{m-1},$$

leading to $L(\theta) = \log \binom{n+m-1}{m-1}$.

Second, we have to encode the data x^n at best given the θ parameter. We suggest using a probability distribution for encoding the finite size data sample x^n, with the following likelihood. For $\theta \neq \left(\frac{n_1(x^n)}{n}, \frac{n_2(x^n)}{n}, \ldots, \frac{n_m(x^n)}{n}\right)$, we cannot encode the data and $P(x^n|\theta) = 0$. For $\theta = \widehat{\theta}(x^n) = \left(\frac{n_1(x^n)}{n}, \frac{n_2(x^n)}{n}, \ldots, \frac{n_m(x^n)}{n}\right)$, the observed data is consistent with the model parameter and we assume that all the possible observable data are uniformly distributed. The number of m-ary strings where the number of occurrences of outcome j is n_j is given by the multinomial coefficient $\frac{n!}{n_1!n_2!\ldots n_m!}$. Thus the probability of observing one particular m-ary string is $P(x^n|\widehat{\theta}(x^n)) = 1/\frac{n!}{n_1!n_2!\ldots n_m!}$. This gives a total code length of

$$L(\widehat{\theta}(x^n), x^n) = \log \binom{n+m-1}{m-1} + \log \frac{n!}{n_1!n_2!\ldots n_m!}. \quad (14)$$

4.2. Theoretical and empirical comparisons

The NML code has a parametric complexity that can be either approximated [3, 7] with errors that are hard to quantify in the non-asymptotic case or calculated exactly (e.g. algorithm in $o(n+m)$ [8]) at the expense of computation time. Table 2 presents the parametric and stochastic complexity of each considered code.

Table 2. Parametric and stochastic complexity per code.

Code name	$COMP_{name}^{(n)}$	$L_{name}\left(x^n \vert \widehat{\theta}(x^n)\right)$
enumerative	$\log \binom{n+m-1}{m-1}$	$\log \frac{n!}{n_1!\ldots n_m!}$
NML	$\approx \frac{m-1}{2} \log \frac{n}{2\pi}$	$\log \frac{n^n}{n_1^{n_1}\ldots n_m^{n_m}}$

In [1], extensive comparisons are reported, regarding the stochastic complexity terms, the parametric complexity terms, the overall code length, the expectation of the code length of all the m-ary sequences under the uniform distribution and the detection of biased dices. Overall, the results are similar to the case of Bernoulli distributions, with differences that increase linearly with the number of parameters.

5. SUMMARY

In this paper, we have revisited the enumerative two-part crude MDL code for the Bernoulli model, which compares favorably with the alternative standard NML code. We have suggested a Bayesian interpretation of the enumerative code, that relies on models for finite size samples and results in a discrete definition of the likelihood of the data given the model parameter. We have shown that the coding length of the model parameter is exactly the same as the model complexity computed by applying the NML formula using the definition of the enumerative maximum likelihood. This means that the enumerative code is both a one-part and two part code, which brings parametrization independence, optimality and simplicity. Surprisingly, the obtained parametric complexity is twice that of the alternative classical NML code or the standard BIC regularization term. The enumerative code has a direct interpretation in terms of two part codes for finite sample data. The model parameter is encoded using a uniform prior w.r.t. the sample size and the data are also encoded using a uniform prior among all the binary strings of given size that can be generated using the model parameter. Experimental comparisons between the enumerative and NML codes show that they are very similar, with small differences only. Under the uniform distribution, the enumerative code compresses most individual sequences slightly better, resulting in a slightly better compression on average. An application to the detection of biased coins demonstrates that the enumerative code has a better sensitivity to biased coins at the expense of more false detections in case of fair coins, but the differences are small and vanish asymptotically.

Extension to the multinomial model is also presented. Using the same approach, we obtain a very simple and interpretable analytic formula for the parametric complexity term, that once again is approximately twice that of the alternative classical NML code or the standard BIC regularization term. The resulting code, both one-part and two-part, is optimal w.r.t. NML approach and parameterization invariant, with a much simpler parametric complexity term. It compresses most strings better than the "classical" NML code with a constant margin and extremely few heavily unbalanced strings with a margin logarithmic in the sample size. Experimental comparisons extend the results obtained with Bernoulli distributions. Both codes are very similar, with small differences that roughly increase linearly with the number of model parameters.

Altogether, the theoretical and experimental results suggest that one might use the enumerative code rather than NML in practice, for Bernoulli and multinomial distributions.

6. REFERENCES

[1] M. Boullé, F. Clérot, and C. Hue, "Revisiting enumerative two-part crude MDL for Bernoulli and multinomial distributions (extended version)," Tech. Rep., arXiv, abs/1608.05522, 2016.

[2] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[3] P.D. Grünwald, *The minimum description length principle*, Adaptive computation and machine learning. MIT Press, 2007.

[4] R. Guigourès, D. Gay, M. Boullé, F. Clérot, and F. Rossi, "Country-scale exploratory analysis of call detail records through the lens of data grid models," in *ECML/PKDD*, 2015, pp. 37–52.

[5] P.M.B. Vitányi and M. Li, "Minimum description length induction, Bayesianism, and Kolmogorov complexity," *IEEE Transactions on information theory*, vol. 46, pp. 446–464, 2000.

[6] P. Adriaans and P. Vitányi, "The power and perils of MDL," in *IEEE International Symposium on Information Theory*, 2007, pp. 2216–2220.

[7] P. Kontkanen, *Computationally efficient methods for MDL-optimal density estimation and data clustering*, Department of Computer Science, series of publications A, report, 2009-11. University of Helsinki, 2009.

[8] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.