M

Minimum Description Length Principle

Teemu Roos
Department of Computer Science, Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland

Abstract

The minimum description length (MDL) principle states that one should prefer the model that yields the shortest description of the data when the complexity of the model itself is also accounted for. MDL provides a versatile approach to statistical modeling. It is applicable to model selection and regularization. Modern versions of MDL lead to robust methods that are well suited for choosing an appropriate model complexity based on the data, thus extracting the maximum amount of information from the data without over-fitting. The modern versions of MDL go well beyond the familiar $\frac{k}{2} \log n$ formula.

Philosophy

The MDL principle is a formal version of Occam's razor. While the Occam's razor only suggests that between hypotheses that are compatible with the evidence, one should choose the simplest one, the MDL principle also quantifies the compatibility of the hypotheses with the evidence. This leads to a trade-off between the complexity of the hypothesis and its compatibility with the evidence ("goodness of fit").

The philosophy of the MDL principle emphasizes that the evaluation of the merits of a model should not be based on its closeness to a "true" model, whose existence is often impossible to verify, but instead on the data. Inspired by Solomonoff's theory of universal induction, Rissanen postulated that a yardstick of the performance of a statistical model is the probability it assigns to the data. Since the probability is intimately related to code length (see below), the code length provides an equivalent way to measure performance. The key idea made possible by the coding interpretation is that the length of the description of the model itself can be quantified in the same units as the code length of the data, namely, bits. Earlier, Wallace and Boulton had made a similar proposal under the title minimum message length (MML) (Wallace and Boulton 1968). A fundamental difference between the two principles is that MML is a Bayesian approach while MDL is not.

The central tenet in MDL is that the better one is able to discover the regular features in the data, the shorter the code length. Showing that this is indeed the case often requires that we assume, for the sake of argument, that the data are generated by a true distribution and verify the statistical behavior of MDL-based methods under this assumption. Hence, the emphasis on

© Springer Science+Business Media New York 2016
C. Sammut, G.I. Webb (eds.), *Encyclopedia of Machine Learning and Data Mining*,
DOI 10.1007/978-1-4899-7502-7_894-1

the freedom from the assumption of a true model is more pertinent in the philosophy of MDL than in the technical analysis carried out in its theory.

Theory

The theory of MDL addresses two kinds of questions: (*i*) the first kind asks what is the shortest description achievable using a given model class, i.e., universal data compression; (*ii*) the second kind asks what can be said about the behavior of MDL methods when applied to model selection and other machine learning and data mining tasks. The latter kind of questions are closely related to the theory of statistical estimation and statistical learning theory. We review the theory related to these two kinds of questions separately.

Universal Data Compression

As is well known in information theory, the shortest expected code length achievable by a uniquely decodable code under a known data source, p^*, is given by the entropy of the source, $H(p^*)$. The lower bound is achieved by using a code word of length $\ell^*(x) = -\log p^*(x)$ bits for each source symbol x. (Here and in the following, log denotes base-2 logarithm.) Correspondingly, a code-length function ℓ is optimal under a source distribution defined by $q(x) = 2^{-\ell(x)}$. (For the sake of notational simplicity, we omit a normalizing factor $C = \sum_x 2^{-\ell(x)}$ which is necessary in case the code is not complete. Likewise, as is customary in MDL, we ignore the requirement that code lengths be integers.) These results can be extended to data sequences whereupon we write $x^n = x_1 \ldots x_n$ to denote a sequence of length n.

While the case where the source distribution p^* is known can be considered solved in the sense that the average-case optimal code-length function ℓ^* is easily established as described above, the case where p^* is unknown is more intricate. Universal data compression studies similar lower bounds when the source distribution is not known or when the goal is not to minimize the expected code length. For example, when the source distribution is only known to be in a given *model class* (a set of distributions), $\mathcal{M}$, the goal may be to find a code that minimizes the *worst-case* expected code length under any source distribution $p^* \in \mathcal{M}$. A uniquely decodable code that achieves near-optimal code lengths with respect to a given model class is said to be *universal*.

Rissanen's groundbreaking 1978 paper (Rissanen 1978) gives a general construction for universal codes based on *two-part codes*. A two-part code first includes a code for encoding a distribution, q, over source sequences. The second part encodes the data using a code based on q. The length of the second part is thus $-\log q(x^n)$ bits. The length of the first part, $\ell(q)$, depends on the complexity of the distribution q, which leads to a trade-off between complexity measured by $\ell(q)$ and goodness of fit measured by $\log q(x)$:

$$\min_q(\ell(q) - \log q(x^n)). \tag{1}$$

For parametric models that are defined by a continuous parameter vector θ, a two-part coding approach requires that the parameters be quantized so that their code length is finite. Rissanen showed that given a k-dimensional parametric model class, $\mathcal{M} = \{p_\theta ; \theta \in \Theta \subset \mathbb{R}^k\}$, the optimal quantization of the parameter space Θ is achieved by using accuracy of order $1/\sqrt{n}$ for each coordinate, where n is the sample size. The resulting total code length behaves as $-\log \hat{p}(x^n) + \frac{k}{2} \log n + \mathcal{O}(1)$, where $\hat{p}(x^n) = \max\{p_\theta(x^n) : \theta \in \Theta\}$ is the maximum probability under model class $\mathcal{M}$. Note that the leading terms of the formula are equivalent to the Bayesian information criterion (BIC) by Schwarz (Schwarz 1978). Later, Rissanen also showed that this is a *lower bound* on the code length of any universal code that holds for all but a measure-zero subset of sources in the given model class (Rissanen 1986).

The above results have subsequently been refined by studying the asymptotic and finite-sample values of the $\mathcal{O}(1)$ residual term for specific model classes. The resulting formulas lead to a more accurate characterization of

model complexity, often involving the Fisher information (Rissanen 1996).

Subsequently, Rissanen and others have proposed other kinds of universal codes that are superior to two-part codes. These include Bayes-type mixture codes that involve a prior distribution for the unknown parameters (Rissanen 1986), predictive forms of MDL (Rissanen 1984; Wei 1992), and, most importantly, normalized maximum likelihood (NML) codes (Yuri 1987; Rissanen 1996). The latter have the important point-wise minimax property that they achieve the minimum worst-case point-wise redundancy:

$$\min_{q} \max_{x^n} -\log q(x^n) + \log \hat{p}(x^n),$$

where the maximum is over all possible data sequences of length n and the minimum is over all distributions.

Behavior of MDL-Based Learning Methods

The philosophy of MDL suggests that data compression is a measure of the success in discovering regularities in the data, and hence, better compression implies better modeling. Showing that this is indeed the case is the second kind of theory related to MDL.

Barron and Cover proposed the *index of resolvability* as a measure of the hardness of estimating a probabilistic source in a two-part coding setting (see above) (Barron and Cover 1991). It is defined as

$$R_n(p^*) = \min_{q} \left(\frac{\ell(q)}{n} + D(p^* \,||\, q) \right),$$

where p^* is the source distribution and $D(p^* \,||\, q)$ denotes the Kullback-Leibler divergence between p^* and q. Intuitively, a source is easily estimable if there exists a simple distribution that is close to the source. The result by Barron and Cover bounds the Hellinger distance between the true source distribution and the distribution $\hat{q}$ minimizing the two-part code length, Eq. (1), as

$$d_H^2(p^*, \hat{q}) \leq \mathcal{O}(R_n(p^*)) \quad \text{in } p^*\text{-probability.}$$

For model selection problems, consistency is often defined in relation to a fixed set of alternative model classes and a criterion that selects one of them given the data. If the criterion leads to the simplest model class that contains the true source distribution, the criterion is said to be consistent. (Note that the additional requirement that the selected model class is the simplest one is needed in order to circumvent a trivial solution in nested model classes where simpler models are subsets of more complex model classes.) There are a large number of results showing that various MDL-based model selection criteria are consistent; for examples, see the next section.

Applications

MDL has been applied in a wide range of applications. It is well suited for model selection problems where one needs not only to estimate continuous parameters but also their number and, more generally, the *model structure*, based on statistical data. Other approaches applicable in many such scenarios include Bayesian methods (including minimum message length), cross validation, and structural risk minimization (see Cross-References below).

Some example applications include the following:

1. Autoregressive models, Markov chains, and their generalizations such as *tree machines* were among the first model classes studied in the MDL literature, see Rissanen (1978, 1984) and Weinberger et al. (1995).
2. Linear regression. Selecting a subset of relevant covariates is a classical example of a situation involving models of variable complexity, see Speed and Yu (1993), Wei (1992), and Rissanen (2000).
3. Discretization of continuous covariates enables the use of learning methods that use discrete data. The granularity of the discretization can be determined by applying MDL, see Fayyad and Irani (1993).
4. The structure of probabilistic graphical models encodes conditional independencies

and determines the complexity of the model. Their structure can be learned by MDL, see, e.g., Lam and Bacchus (1994) and Silander et al. (2010)

Future Directions

The development of efficient and computationally tractable codes for practically relevant model classes is required in order to apply MDL more commonly in modern statistical applications. The following are among the most important future directions:

- While the original $\frac{k}{2}\log n$ formula is still regularly referred to as "the MDL principle," future work should focus on modern formulations involving more advanced codes such as the NML and its variations.
- There is strong empirical evidence suggesting that coding strategies with strong minimax properties lead to robust model selection methods, see, e.g., Silander et al. (2010). Tools akin to the index of resolvability are needed to gain better theoretical understanding of the properties of modern MDL methods.
- Scaling up to modern big data applications, where model complexity regularization is crucial, requires approximate versions of MDL with sublinear computational and storage requirements. Predictive MDL is a promising approach in handling high-throughput streaming data scenarios.

Cross-References

- Complete Minimum Description Length
- Cross Validation
- Inductive Inference
- Learning Graphical Models
- Minimum Message Length
- Model Evaluation
- Occam's Razor
- Overfitting
- Regularization
- Structural Risk Minimization
- Universal Learning Theory

Recommended Reading

Good review articles on MDL include Barron et al. (1998); Hansen and Yu (2001). The textbook by Grünwald (2007) is a comprehensive and detailed reference covering developments until 2007 Grünwald (2007).

Barron A, Cover T (1991) Minimum complexity density estimation. IEEE Trans Inf Theory 37(4):1034–1054

Barron A, Rissanen J, Yu B (1998) The minimum description length principle in coding and modeling. IEEE Trans Inf Theory 44:2734–2760

Fayyad U, Irani K (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajczy R (ed) Proceedings of the 13th International Joint Conference on Artificial Intelligence and Minimum Description Length Principle, Chambery. Morgan Kauffman

Grünwald P (2007) The Minimum Description Length Principle. MIT Press, Cambridge

Hansen M, Yu B (2001) Model selection and the principle of minimum description length. J Am Stat Assoc 96(454):746–774

Lam W, Bacchus F (1994) Learning Bayesian belief networks: an approach based on the MDL principle. Comput Intell 10:269–293

Rissanen J (1978) Modeling by shortest data description. Automatica 14(5):465–658

Rissanen J (1984) Universal coding, information, prediction, and estimation. IEEE Trans Inf Theory 30:629–636

Rissanen J (1986) Stochastic complexity and modeling. Ann Stat 14(3):1080–1100

Rissanen J (1996) Fisher information and stochasic complexity. IEEE Trans Inf Theory 42(1):40–47

Rissanen J (2000) MDL denoising. IEEE Trans Inf Theory 46(7):2537–2543

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Silander T, Roos T, Myllymäki P (2010) Learning locally minimax optimal Bayesian networks. Int J Approx Reason 51(5):544–557

Speed T, Yu B (1993) Model selection and prediction: normal regression. Ann Inst Stat Math 45(1):35–54

Wallace C, Boulton D (1968) An information measure for classification. Comput J 11(2):185–194

Wei C (1992) On predictive least squares principles. Ann Stat 20(1):1–42

Weinberger M, Rissanen J, Feder M (1995) A universal finite memory source. IEEE Trans Inf Theory 41(3):643–652

Yuri Shtarkov (1987) Universal sequential coding of single messages. Probl Inf Transm 23(3):3–17