Computational Machine Learning in Theory and Praxis*

Ming Li**
University of Waterloo

Paul Vitányi***
CWI and Universiteit van Amsterdam

Abstract. In the last few decades a computational approach to machine learning has emerged based on paradigms from recursion theory and the theory of computation. Such ideas include learning in the limit, learning by enumeration, and probably approximately correct (pac) learning. These models usually are not suitable in practical situations. In contrast, statistics based inference methods have enjoyed a long and distinguished career. Currently, Bayesian reasoning in various forms, minimum message length (MML) and minimum description length (MDL), are widely applied approaches. They are the tools to use with particular machine learning praxis such as simulated annealing, genetic algorithms, genetic programming, artificial neural networks, and the like. These statistical inference methods select the hypothesis which minimizes the sum of the length of the description of the hypothesis (also called 'model') and the length of the description of the data relative to the hypothesis. It appears to us that the future of computational machine learning will include combinations of the approaches above coupled with guaranties with respect to used time and memory resources. Computational learning theory will move closer to practice and the application of the principles such as MDL require further justification. Here, we survey some of the actors in this dichotomy between theory and praxis, we justify MDL via the Bayesian approach, and give a comparison between pac learning and MDL learning of decision trees.

1 Introduction

In the past three decades, the art of machine learning has evolved from two opposed ends: from the pure theoretical approach there have emerged theories

* The first author was supported in part by NSERC operating grant OGP-046506, ITRC, and a CGAT grant. The second author was by NSERC through International Scientific Exchange Award ISE0125663, and by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, and by NWO through NFI Project ALADDIN under Contract number NF 62-376.
** Computer Science Department, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Email: mli@math.uwaterloo.ca
*** CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl

that are getting closer to practice; from the applied side *ad hoc* theories have arisen that often lack theoretical justification. However, there is also a sound body of methods, often statistics based, which are widely applied in practical situations.

It is the purpose of this article to examine part of this situation. On the one hand we follow the thread of inductive inference and pac learning; from the other end we examine an approach related to statistics, Bayesian reasoning, and MDL. It appears to us that the future of computational machine learning will involve combinations of these approaches coupled with guaranties with respect to used time and memory resources. It is clear that computational learning theory will move closer to practice and the application of principles such as MDL require further justification. In this paper, we justify certain applications of MDL via the Bayesian approach.

It is not always realized that most of traditional statistics is about computational learning. It is always involved with algorithms to obtain the general from the particular. Not only is this approach to machine learning mathematical, but also very much applied. A novel approach in statistical learning is based on minimum description length of hypotheses and data together—one way to express the so-called MDL principle.

A general task of statistical learning is to select the most plausible hypothesis from among all possible hypotheses in the light of experimental evidence. The classic method to do so is Bayes' rule. The problem with applying Bayes' rule is that one requires the prior probabilities of the possible hypotheses first. Unfortunately, it is often impossible to obtain these.

One way out of this conundrum is to require inference of hypotheses to be completely data driven. The MDL approach, [28, 20], embodies this idea. MDL is usually presented as justified in and of itself by philosophical persuasion. We give a derivation for MDL from Bayes' rule, identifying similarities and differences, and give a comparison of pac learning criteria and MDL algorithms for the practical topic of decision tree learning. This report will be published in the celebration *Lecture Notes in Computer Science, Vol. 1000*, Springer Verlag, Heidelberg, 1995. The MDL discussion in this paper consists for the most part of preliminary results of ongoing work.

2 Learning In The Limit

Following E.M. Gold, [7], we are given an effective enumeration of partial recursive functions $f_1, f_2, \ldots$. Such an enumeration can be the functions computed by Turing machines, but also the functions computed by finite automata. We want to infer a particular function f. To do so, we are presented with a sequence of examples $D = e_1, e_2, \ldots, e_n$, containing elements (possibly with repetitions) of the form

$$e = \begin{cases} (x, y, 0) \text{ if } f(x) \neq y, \\ (x, y, 1) \text{ if } f(x) = y. \end{cases}$$

For $n \to \infty$ we assume that D contains all elements of the displayed form.

To infer a function, Gold proposed the method of *induction by enumeration.* We eliminate all functions which are inconsistent with D from left to right in the enumeration of functions, up to the index of the first consistent function. Each step we receive a new example e, set $D := D, e$, and repeat this process. Eventually, when we meet a function in the enumeration which is identical to f, such an enumeration process converges, and the function will not change any more. This deceptively simple idea has generated a large body of sophisticated literature, see [1] for a survey.

The Gold model apparently only partially captures human learning, even as an idealized model. It requires the learner to identify a function precisely while this clearly does not happen in human practice. It also requires the machine to identify the target function with certainty. This again is not generally (or mostly not) required in practice. That is, this model is less relevant to practice since its goal is too high. Moreover, for almost all nontrivial concepts it takes exponential time to find the *first* concept that is consistent with the data at each new example (as required by the identification in the limit procedure) and it takes forever to converge.

3 Learning In Polynomial Time and Examples

The Gold paradigm ignores the time requirement. Clearly, adding significant (like polynomial) time restrictions directly to the inductive inference model make almost all functions, except for a few trivial ones not learnable. The next step towards a 'feasible' learning theory was taken by L. Valiant in 1984, [26].

Modeling practical learning by putting some time bound on the learning process implies that not all examples can turn up. This means it is impossible to infer a concept precisely, and therefore we can only hope to learn the concept approximately. But we also have to account for where the examples originate. We could be presented with unrepresentative examples for the concept to be inferred. Valiant assumes that the examples are drawn randomly from a sample space according to a probability distribution. The approximation between the target concept and the learned concept can be expressed in the probability of the set of examples on which the two concepts disagree.

We give a short description of the pac-model in the simplest case. The sample space S contains *examples.* A *concept* c is a subset of S. A *concept class* $\mathcal{C}$ is a set of concepts.

Consider a concept class $\mathcal{C}$. For each concept $f \in \mathcal{C}$ and example $v \in S$,

$$f(v) = \begin{cases} 1 \text{ if } v \text{ is a positive example of } f \\ 0 \text{ if } v \text{ is a negative example of } f. \end{cases}$$

The learning algorithm draws an examples from the sample space S according to a fixed but unknown probability distribution P. Each example in the sample comes with a label 'Positive' or 'Negative'.

A concept class $\mathcal{C}$ is *pac-learnable* (probably approximately correct learnable) iff there exists a (possibly randomized) learning algorithm A such that, for each

$f \in \mathcal{C}$ and ϵ $(0 < \epsilon < 1)$, algorithm A halts in a finite number of steps and examples, and outputs a concept $h \in \mathcal{C}$ which satisfies the following. With probability at least $1 - \epsilon$,

$$\sum_{f(v) \neq h(v)} P(v) < \epsilon.$$

A concept class is *polynomially pac-learnable* iff it is pac-learnable and the learning algorithm always halts within time and number of examples $p(l(f), 1/\epsilon)$, for some polynomial p, where $l(f)$ is the length of a given binary program to compute f.

Subsets of natural numbers, the set of finite automata, and the set of Boolean formulas are all examples of concept classes. It turns out that none of the above classes is known to be polynomial time pac-learnable. The classes that are known to be polynomial time pac learnable include for example k-DNF which is the set of DNF formulas such that each term contains at most k variables.

In the past decade, following the introduction of the pac model, more negative (nonlearnability) results were proved than positive (learnability) results. Although the pac model has relaxed the requirements from the Gold model by allowing learning approximately and only with high probability instead of with certainty, it seems that we are still asking for too much. In practical learning situations, the class of probability distributions is often restricted; the learner usually is allowed to ask questions; there are may even be helpful teachers. Great efforts have been made to improve the pac-model to be more close to practical situations. New models are proposed to allow the learner to ask various kinds of questions, to have good teachers [1], to learn under a restricted class of distributions [11].

4 Learning With MDL

The previous models were mainly about formal 'criteria' for learning such as learning in the limit and pac learning. The notion of learning by enumeration is an algorithm for learning, albeit a very unpractical one. Similarly, the MDL principle is an algorithmic paradigm that is widely applied. That is, it is widely applied at least in spirit; to apply it literally may run in computational difficulties since it involves finding an optimum in a exponentially large set of candidates as noted for example in [12]. Yet in some cases one can approximate this optimum, [27, 30].

A practice oriented theory like MDL, although often lacking in justification, apparently works and is used by practitioners. The MDL principle is very easy to use in some loose sense, but it is hard to justify. A user of the MDL principle does not need to *prove* that the concept class concerned is learnable, rather he needs to choose a concept that can be described shortly and without causing too many errors (and he needs to balance these two things). In this section, we will describe the MDL principle and establish its justification from first principles.

4.1 Bayesian Reasoning

Consider a situation in which one has a set of observations of some phenomenon, and also a (possibly countably infinite) set of hypotheses which are candidates to explain the phenomenon. For example, we are given a coin and we flip it 100 times. We want to identify the probability that the coin has outcome 'head' in a single coin flip. That is, we want to find the bias of the coin. The set of possible hypotheses is uncountably infinite if we allow each real bias in $[0,1]$, and countably infinite if we allow each rational bias in $[0,1]$.

For each hypothesis H we would like to assess the probability that H is the 'true' hypothesis, given the observation of D. This quantity, $\Pr(H|D)$, can be described and manipulated formally in the following way.

Consider a sample space Ω. Let D denote a sample of outcomes, say experimental data concerning a phenomenon under investigation. Let $H_1, H_2, \ldots$ be an enumeration of countably many hypotheses concerning this phenomenon, say each H_i is a probability distribution over Ω. The list $\mathcal{H} = \{H_1, H_2, \ldots\}$ is called the *hypothesis space*. The hypotheses H_i are exhaustive and mutually exclusive.

For example, say the hypotheses enumerate the possible rational (or computable) biases of the coin. As another possibility there may be only two possible hypotheses: hypothesis H_1 which says the coin has bias 0.2, and hypothesis H_2 which puts the bias at 0.8.

Suppose we have *a priori* a distribution of the probabilities $P(H)$ of the various possible hypotheses in $\mathcal{H}$ which means that $\sum_{H \in \mathcal{H}} P(H) = 1$. Assume further more that for all $H \in \mathcal{H}$ we can compute the probability $\Pr(D|H)$ that sample D arises if H is the case. Then we can also compute (or approximate in case the number of hypotheses with nonzero probability is infinite) the probability $\Pr(D)$ that sample D arises at all

$$\Pr(D) = \sum_{H \in \mathcal{H}} \Pr(D|H)P(H).$$

From the definition of conditional probability it is easy to derive **Bayes' formula**

$$\Pr(H|D) = \frac{\Pr(D|H)P(H)}{\Pr(D)}. \tag{1}$$

The prior probability $P(H)$ is often considered as the learner's *initial degree of belief* in hypothesis H. In essence Bayes' rule is a mapping from *a priori* probability $P(H)$ to *a posteriori* probability $\Pr(H|D)$ determined by data D.

Continuing to obtain more and more data, this way the total inferred probability will concentrate more and more on the 'true' hypothesis. We can draw the same conclusion of course, using more examples, by the law of large numbers. In general, the problem is not so much that in the limit the inferred probability would not concentrate on the true hypothesis, but that the inferred probability gives as much information as possible about the possible hypotheses from only a limited number of data. Given the prior probability of the hypotheses, it is easy to obtain the inferred probability, and therefore to make informed decisions. However, in general we don't know the prior probabilities. The following

MDL approach in some sense replaces an unknown prior probability by a fixed 'universal' probability.

4.2 Minimum Description Length

A relatively recent method in statistics was developed by C. Wallace [28, 29], who called the principle involved *Minimum Message Length* and J. Rissanen [18, 20] who named it *Minimum Description Length* leading to *stochastic complexity* [22]. In accordance with Occam's dictum, it tells us to go for the explanation that compresses the data the most.

Definition 1. Given a sample of data, and an effective enumeration of theories, the MDL principle is to select the theory which minimizes the sum of

- the length, in bits, of the description of the theory; and
- the length, in bits, of data when encoded with the help of the theory.

Intuitively, with a more complex description of the hypothesis H, it may fit the data better and therefore decreases the misclassified data. If H describes all the data, then it does not allow for measuring errors. A simpler description of H may be penalized by increasing the number of misclassified data. If H is a trivial hypothesis that contains nothing, then all data are described literally and there is no generalization. The rationale of the method is that a balance in between seems required. However, straightforward application of the idea may lead to obviously false results in pathological cases.

Example 1. Suppose we want to determine the probability of success p of a Bernoulli process $(p, 1-p)$ (a coin with probability p of coming up heads) through a data sample consisting of a sequence of independent coin flips. Now assume the contrived example that we set hypothesis H (bias p expressed in binary expansion) and data D (a sequence of h(ead)s and t(ail)s) as

$$\begin{aligned} H_p &:= p = 0.r_1 \dots r_n \\ D &:= s_1 \dots s_n. \end{aligned}$$

Let the hypothesis $H = 0.1$ (in binary, so the bias is $1/2$) be both the true one and the most likely one to lead to outcome D. Let $D = s_1 \dots s_n$ be a random typical outcome of fair coin tosses. Such an outcome will have about the same number of heads and tails. But straightforward application of the MDL principle selects H_p with $p = 0.s_1 \dots s_n$.

4.3 Derivation of MDL from Bayes' Rule

Here, we present a rigorous analysis of a pure version of MDL. While this does not rigorously prove anything about applied MDL, the relations with Bayesian inference we establish for pure MDL are corroborated by empirical evidence for applied MDL as in [14, 24, 3].

How to derive MDL from first principles, *in casu* Bayes' rule as given by Equation 1? In the latter equation we are only concerned with maximizing the term $\Pr(H|D)$ over H. Taking the negative logarithm at both sides of the equation, this is equivalent to *minimizing* the expression $-\log \Pr(H|D)$ over H:

$$-\log \Pr(H|D) = -\log \Pr(D|H) - \log P(H) + \log \Pr(D).$$

Since the probability $\Pr(D)$ is constant under varying H, this means we want to find H which minimizes:

$$-\log \Pr(D|H) - \log P(H). \tag{2}$$

Because of space limitations, we refer the reader to [12] for all definitions and analysis of auxiliary notions. This way we also do not deviate from the main argument, do not obstruct the knowledgeable reader, and do not confuse or discourage the reader who is as yet unfamiliar with Kolmogorov complexity theory.

To obtain the MDL principle it suffices to replace the probabilities involved in Equation 2 by the so-called *universal probability* $\mathbf{m}(\cdot)$. The analysis of the *conditions* under which this substitution is *justified*, or, conversely, how application of MDL is equivalent to Bayesian inference using *admissible* probabilities, is deferred to the next section.

$$\begin{aligned} \log P(H) &= \log \mathbf{m}(H) + O(1), \\ \log \Pr(D|H) &= \log \mathbf{m}(D|H) + O(1). \end{aligned} \tag{3}$$

According to [12],

$$\begin{aligned} -\log \mathbf{m}(H) &= K(H) \pm O(1), \\ -\log \mathbf{m}(D|H) &= K(D|H) \pm O(1), \end{aligned}$$

where $K(\cdot)$ is the prefix complexity. Therefore, using the substitution of Equation 3 we can replace the sum of Equation 2 by the sum of the minimum lengths of effective self-delimiting programs which compute descriptions of H and $D|H$. That is, we look for the H that minimizes

$$K(D|H) + K(H), \tag{4}$$

which is the code-independent, recursively invariant, absolute form of the MDL principle.

We mention that the term $-\log \Pr(D|H)$ is also known as the *self-information* in information theory and the negative log-likelihood in statistics. It can now be regarded as the number of bits it takes to redescribe or encode D with an ideal code relative to H.

In [5] (see also [12], Theorem 3.5 on p. 195) it is proved that symmetry of information holds for individual objects in the following sense

$$\begin{aligned} K(H,D) &= K(H) + K(D|H, K(H)) + O(1) \\ &= K(D) + K(H|D, K(D)) + O(1). \end{aligned} \tag{5}$$

Since it is not more difficult to describe some object if we get more conditional information, we have $K(D|H, K(H)) \leq K(D|H) + O(1)$. Thus, by Equation 5 the quantity in Equation 4 satisfies

$$K(H) + K(D|H) \geq K(H, D) + O(1) \geq K(D) + O(1).$$

For the trivial hypothesis $H_0 = D$ or $H_0 = \emptyset$ we find

$$K(H_0) + K(D|H_0) = K(D) + O(1).$$

At first glance this would mean that the hypothesis H_{mdl} which minimizes the sum of Equation 4 could be set to D or $\emptyset$, which is absurd in general. However, we have only derived the validity of Equation 4 under the condition that Equation 3 holds. The crucial part of our justification of MDL is to establish precisely when Equation 3 is valid, which is the main thrust of this paper.

4.4 The Rôle of Universal Probability

It is well known, see for example [12], that the so-called Shannon-Fano code for an ensemble of source words distributed according to probability Q is a prefix code E_Q with $l(E_Q(x)) = -\log Q(x)$ satisfying

$$\sum_x Q(x) l(E_Q(x)) = \min_E \{\sum_x Q(x) l(E(x)) : E \text{ is a prefix code}\} + O(1).$$

Therefore, without any restrictions Equation 2 is equivalent to

$$E_{\Pr(\cdot|H)}(D) + E_P(H).$$

To satisfy restriction Equation 3 we *can simply assume* that the prior probability $P()$ in Bayes' rule Equation 1 is $\mathbf{m}()$. Whether this can be justified or not is a question which we address later.

However, we *cannot assume* that the probability $\Pr(\cdot|H)$ equals $\mathbf{m}(\cdot|H)$. Namely, as explained at length in Section 4.1, probability $\Pr(\cdot|H)$ may be totally determined by the hypothesis H. Depending on H therefore, $l(E_{\Pr(\cdot|H)}(D))$ may be *vastly* different from $K(D|H)$ indeed. This holds especially for 'simple' data D which have low probability under assumption of hypothesis H.

Example 2. Let us look at a simple example evidencing this discrepancy. Suppose we flip a coin of unknown bias n times. Let hypothesis H and data D be defined by:

$$\begin{aligned} H &:= \text{ Probability 'head' is } 1/2 \\ D &:= \underbrace{hh\ldots h}_{n \text{ times '}h\text{'(ead)s}} \end{aligned}$$

Then we have $\Pr(D|H) = 1/2^n$ and

$$l(E_{\Pr(\cdot|H)}(D)) = -\log \Pr(D|H) = n.$$

In contrast,

$$K(D|H) \leq \log n + 2 \log\log n + O(1).$$

However, the theory dealing with randomness of individual objects states that under certain conditions $-\log \Pr(D|H)$ and $K(D|H)$ are close.

First, assume that $\Pr(\cdot|\cdot)$ is a recursive function. That is, it can be computed to any required precision for each argument D and conditional H. Then we can appeal to the following known facts, [12]. Firstly, it is known that $\mathbf{m}(D|H) \geq 2^{-K(\Pr(\cdot|H))} \Pr(D|H)$. Therefore,

$$\log(\mathbf{m}(D|H)/\Pr(D|H)) \geq -K(\Pr(\cdot|H)) \geq -K(H) + O(1). \tag{6}$$

The last inequality arises since from H we can compute $\Pr(\cdot|H)$ by assumption on $\Pr(\cdot|\cdot)$. Secondly, for D which are sufficiently *random* with respect to the distribution $\Pr(\cdot|H)$ (with respect to H therefore) in the sense of Martin-Löf one can show, [12], pp. 228-229, that

$$\log(\mathbf{m}(D|H)/\Pr(D|H)) \leq 0. \tag{7}$$

The overwhelming majority of D's is random in this sense because for each H we have

$$\sum_D \Pr(D|H) 2^{\log \mathbf{m}(D|H)/\Pr(D|H))} = \sum_D \log \mathbf{m}(D|H) \leq 1,$$

since $\mathbf{m}(\cdot|H)$ is a probability distribution. For D's which are random in the appropriate sense, Equations 6, 7 mean by Equation 3 that

$$K(D|H) - K(\Pr(\cdot|H)) \leq -\log \Pr(D|H) \leq K(D|H).$$

Above, by assuming that the *a priori* probability $P(H)$ of hypothesis H is in fact the universal probability we obtained $-\log P(H) = \mathbf{m}(x)$. However, we do not need to make this assumption. For recursive $P(\cdot)$, we can analyze the situation when H is random in the required sense with respect to $P(\cdot)$. The first inequality below holds since $\mathbf{m}()$ majorizes $P()$ this way; the second inequality expresses the assumption of randomness of H,

$$\mathbf{m}(H) \geq 2^{-K(P)} P(H)$$
$$\log(\mathbf{m}(H)/P(H)) \leq 0.$$

Then,

$$K(H) - K(P) \leq -\log P(H) \leq K(H). \tag{8}$$

Altogether we find

$$K(D|H) + K(H) - \alpha(P,H) \leq -\log \Pr(D|H) - \log P(H) \leq K(D|H) + K(H), \tag{9}$$

with $\alpha(P,H) = K(\Pr(\cdot|H)) + K(P)$ and we note that $K(\Pr(\cdot|H)) \leq K(H) + O(1)$. We call this formula the **Fundamental Intequality (FI)** because it describes the fundamental relation between Bayes' Rule and MDL, and it is left for us to interpret it.

4.5 Validity Range of FI

We begin by stressing that our arguments validate the bounds in Equation 9 only in case simultaneously H is P-random and D is $\Pr(\cdot|H)$-random. What is the meaning of this?

H is P-random means that the true hypothesis must be 'typical' for the prior distribution P in the sense that it must belong to all effective majorities (sets on which the majority of P-probability is concentrated). In the [12] it is shown that this is the set of H's such that with high P-probability $K(H) \approx -\log P(H)$. In case $P(H) = \mathbf{m}(H)$, that is, the prior distribution equals the universal distribution, then for *all* H we have $K(H) = -\log P(H)$, that is, all hypotheses are random with respect to the universal distribution..

Let us look at an example of a distribution where some hypotheses are random, and some other hypotheses are nonrandom. Let the possible hypotheses correspond to the binary strings of length n, and P is the uniform distribution which assigns probability $P(H) = 1/2^n$ to each hypothesis $H \in \{0,1\}^n$. Then $H = 00\ldots0$ has low complexity $K(H) \leq \log n + 2\log\log n$. However, $-\log P(H) = n$. Therefore, by Equation 8, H is not P-random. If we obtain H by n flips of a fair coin however, then with overwhelming probability we will have that $K(H) = n + O(\log n)$ and therefore $-\log P(H) \approx K(H)$ and H is P-random.

D is $\Pr(\cdot|H)$-random means that the data are random with respect to the probability distribution $\Pr(\cdot|H)$ induced by the hypothesis H. Therefore, we require that the sample data D are 'typical' that is 'randomly distributed' with respect to $\Pr(\cdot|H)$. If, for example, $H = (\mu, \sigma)$ induces the Gaussian distribution $\Pr(\cdot|(\mu,\sigma)) = N(\mu,\sigma)$ and the data D are concentrated in a tail of this distribution, like $D = 00\ldots0$, then D is atypical with respect to $\Pr(\cdot|H)$ in the sense of being non-random because it violates the $\Pr(\cdot|H)$-randomness test Equation 7.

4.6 If Optimal Hypothesis Violates FI

The only way to violate the Fundamental Inequality is that either D is not $\Pr(\cdot|H)$-random and by Equation 7, $-\log \Pr(D|H) > K(D|H)$, or that H is not P-random and by Equation 8, $-\log P(H) > K(H)$. We give an example of the first case.

We sample a polynomial $H_2 = ax^2 + bx + c$ at n arguments choosen uniformly at random from the interval $[0,1]$. The sampling process introduces Gaussian errors in the function values obtained. The set of possible hypotheses is the set of polynomials. Assume that all numbers involved are of fixed bounded precision.

Because of the Gaussian error in the measuring process, with overwhelming probability the only polynomials H_{n-1} which fit the sample precisely are of degree $n-1$. Denote the data by D. Now this hypothesis H_{n-1} is likely to minimize $K(D|H)$ since we have just to describe the n lengths of the intervals between the sample points along the graph of H_{n-1}. However, for this H_{n-1} data D is certainly not $\Pr(\cdot|H_{n-1})$-random, since it is extremely unlikely, and hence untypical, that D arises when sampling H_{n-1} with Gaussian error. Therefore, Equation 7 is violated which means that $-\log \Pr(D|H_{n-1}) > K(D|H_{n-1})$ contrary to what we used in deriving the Fundamental Inequality. With prior probability $P(\cdot) := \mathbf{m}(\cdot)$, which means $-\log P(\cdot) = K(\cdot) + O(1)$, this moreover violates the Fundamental Inequality.

In contrast, H_2 will show the data sample D random to it in which case the Fundamental Inequality holds. Now what happens if H_{n-1} is the true hypothesis and the data sample D by chance is as above? In that case the Fundamental Inequality is violated and Bayes' Rule and MDL may select very different hypotheses as the most likely ones, respectively.

4.7 If Optimal Hypothesis Satisfies FI

Suppose the H that minimizes either $-\log \Pr(D|H) - \log P(H)$ (selection according to Bayes' Rule) or $K(D|H) + K(H)$ (selection according to MDL) satisfies the Fundamental Inequality. This condition restricts the set of *admissible* hypotheses with the result that not necessarily *a priori* holds $K(D|H) + K(H) \approx K(D)$ for some minimizing trivial admissible hypothesis (like $H = \emptyset$ or $H = D$ which are not admissible). When moreover $\alpha(P, H)$ as in Equation 9 is small, then Bayes' Rule and MDL select the same optimizing hypothesis. The smallness of $\alpha(P, H)$ means that both the prior distribution P is simple, and that the probability distribution $\Pr(\cdot|H)$ over the data samples induced by hypothesis H simple. In contrast, if $\alpha(P, H)$ is large, which means that either of the mentioned distributions is not simple, for example when $K(\Pr(\cdot|H)) = K(H)$ for complex H, then there may be some discrepancy. Namely, in Bayes' Rule our purpose is to maximize $\Pr(H|D)$, and the hypothesis H which minimizes $K(D|H) + K(H)$ also maximizes $\Pr(H|D)$ up to a $2^{-\alpha(P,H)}$ multiplicative factor. Conversely, the H that maximizes $\Pr(H|D)$ also minimizes $K(D|H) + K(H)$ up to an additive term $\alpha(P, H)$. That is, with

$$\begin{aligned} H_{\text{mdl}} &:= \text{minarg}_H \{K(D|H) + K(H)\} \\ H_{\text{bayes}} &:= \text{maxarg}_H \{\Pr(H|D)\} \end{aligned} \tag{10}$$

we have

$$\begin{aligned} 2^{-\alpha(P,H)} \le \frac{\Pr(H_{\text{mdl}}|D)}{\Pr(H_{\text{bayes}}|D)} \le 1 \\ \alpha(P, H) \ge K(D|H_{\text{mdl}}) + K(H_{\text{mdl}}) - K(D|H_{\text{bayes}}) - K(H_{\text{bayes}}) \ge 0. \end{aligned} \tag{11}$$

4.8 What MDL Does

We can now assess what prior distributions and assumptions about the relation between the data sample and selected hypothesis MDL assumes. That is, how we can translate MDL in terms of Bayes' Rule. Identifying application of MDL with application of Bayes' rule on some prior distribution P we must assume that the Fundamental Inequality is satisfied for H_{mdl} as defined in Equation 10. This means that H_{mdl} is P-random for the used prior distribution P. One choice to guarantee this is to choose

$$P(\cdot) := \mathbf{m}(\cdot)(= 2^{-K(\cdot)}).$$

This is a valid choice even though $\mathbf{m}$ is not recursive, since the latter requirement arose from the requirement that $\mathbf{m}(\cdot)/P(\cdot)$ be enumerable, which is certainly guaranteed by choice of $P(\cdot) := \mathbf{m}(\cdot)$. This choice of prior distribution over the hypotheses is an objective and recursively invariant quantified form of Occam's razor: simple hypothesis H (with $K(H) \ll l(H)$ have high probability, and complex or random hypothesis H (with $K(H) \approx l(H)$) have low probability, see [12]. This choice of prior distribution is most convenient, since the randomness test $\log \mathbf{m}(H)/P(H) = 0$ for *each* hypothesis H. This means that all hypotheses H are random with respect to distribution $\mathbf{m}(\cdot)$, see [12].

With prior $P(\cdot)$ set to $\mathbf{m}(\cdot)$, and the Fundamental Inequality satisfied, it follows that D must be $\Pr(\cdot|H_{\mathrm{mdl}})$-random. With the choosen prior, and given data sample D, this restriction is therefore the only constraint on the range of hypothesis from which we can choose H_{mdl}. Hence we can interprete MDL as an application of Bayes' Rule with as prior distribution the universal distribution $\mathbf{m}(\cdot)$ and selection of a hypothesis H_{mdl} which shows the given data sample random to it in the precise sense of $\Pr(\cdot|H_{\mathrm{mdl}})$-randomness of individual objects as developed in [12].

Since the notion of individual randomness incorporates all effectively testable properties of randomness, practical application of MDL will select the simplest hypothesis which balances the $K(D|H)$ and $K(H)$ and also shows the data sample D random (as far as we are able to tell) with respect to the selected hypothesis H_{mdl}.

This is the 'reason' why the hypothesis selected by MDL is not simply the one which perfectly fits the data. With some amount of overstatement on can say that if one obtains perfect data for a true hypothesis, then MDL interprets these data as data obtained from a simpler hypothesis subject to measuring errors. Consequently, in this case MDL is going to give you the *false simple* hypothesis and *not* the *complex true* hypothesis.

- That is, MDL only gives us the true hypothesis if the data satisfies certain conditions relative to the true hypothesis. Stated differently: there are only data and no true hypothesis for MDL. The principle simply obtains the hypothesis which is suggested by the data and it is assumed that the data are random with respect to the hypothesis. For proper application of MDL assuming a true underlying hypothesis one must also assume that the data are not completely perfect.

5 Applying MDL

Unfortunately, the function K is not computable, [12]. In this paper we will not address the question which encoding one uses in practice, but refer to references [20, 17, 28, 30, 27]. For practical applications one must settle for easily computable approximations.

In statistical applications, H is some statistical distribution (or model) $H = P(\theta)$ with a list of parameters $\theta = (\theta_1, \ldots, \theta_k)$, where the number k may vary and influence the (descriptional) complexity of θ. (For example, H can be a normal distribution $N(\mu, \sigma)$ described by $\theta = (\mu, \sigma)$.) Each parameter θ_i is truncated to fixed finite precision. The data sample consists of n outcomes $\mathbf{y} = (y_1, \ldots, x_n)$ of n trials $\mathbf{x} = (x_1, \ldots, x_n)$ for distribution $P(\theta)$. The data D in the above formulas is given as $D = (\mathbf{x}, \mathbf{y})$. By expansion of conditional probabilities we have therefore

$$\Pr(D|H) = \Pr(\mathbf{x}, \mathbf{y}|H) = \Pr(\mathbf{x}|H) \cdot \Pr(\mathbf{y}|H, \mathbf{x}).$$

In the argument above we take the negative logarithm of $\Pr(D|H)$, that is,

$$-\log \Pr(D|H) = -\log \Pr(\mathbf{x}|H) - \log \Pr(\mathbf{y}|H, \mathbf{x}).$$

Taking the negative logarithm in Bayes' rule and the analysis of the previous section now yields that MDL selects the hypothesis with highest inferred probability satisfying $\mathbf{x}$ is $\Pr(\cdot|H)$-random and $\mathbf{y}$ is $\Pr(\cdot|H, \mathbf{x})$-random.

Learning Polynomials. In biological modeling, we often wish to fit a polynomial f of unknown degree to a set of data points D such that it can predict future data y given x. Even if the data did come from a polynomial curve of degree, say two, because of measurement errors and noise, we still cannot find a polynomial of degree two fitting all n points exactly. In general, the higher the degree of fitting polynomial, the greater the precision of the fit. For n data points, a polynomial of degree $n-1$ can be made to fit exactly, but probably has no predictive value. In general, we are either given the sampling positition $\mathbf{x}$, or we restrict the viable models to those for which $\mathbf{x}$ fairly covers the range of arguments with typical behavior. That is, we do not want $\mathbf{x}$ to be bunched together in one corner of the representative domain. This means that we can treat the term $-\log \Pr(\mathbf{x}|H)$ as a constant. Hence, applying MDL we look for $H_{\mathrm{mdl}} := \mathrm{minarg}_H \{K(\mathbf{y}|H, \mathbf{x}) + K(H)\}$.

Let us apply the MDL principle where we describe all $(k-1)$-degree polynomials by a vector of k entries, each entry with a precision of d bits. Then, the entire polynomial is described by

$$kd + O(\log kd) \text{ bits.} \tag{12}$$

(We have to describe k, d, and account for self-delimiting encoding of the separate items.) For example, $ax^2 + bx + c$ is described by (a, b, c) and can be encoded by about $3d$ bits. Each datapoint (x_i, y_i) which needs to be encoded separately with precision of d bits per coordinate costs about $2d$ bits.

To apply the MDL principle we must trade the cost of hypothesis H (Equation 12) against the cost of describing $\mathbf{y}$ with help of H and $\mathbf{x}$. As a trivial example, suppose the $n-1$ out of n datapoints fit a polynomial of degree 2 exactly, but only 2 points lie on any polynomial of degree 1 (a straight line). Of course, there is a polynomial of degree $n-1$ which fits the data precisely (up to precision). Then the MDL cost is $3d+2d$ for the 2nd degree polynomial, $2d+(n-2)d$ for the 1st degree polynomial, and nd for the $(n-1)$th degree polynomial. Given the choice among those three options, we select the 2nd degree polynomial for all $n>5$. For a more sophisticated approach accounting for the average encoding cost of exceptions assuming they are Gaussian distributed, see [20].

Learning Decision Trees. Let us compare the application of MDL and of pac-learning for a widely used and practical problem like learning decision trees. Before proceeding, recall that pac-learning is not an algorithmic method, but rather a formal criterium to determine whether a given algorithm has performed satisfactory within that criterium. In contrast, the MDL principle is a method for designing algorithms. Theoretical approaches worry whether certain formal requirements like those for pac-learning of certain problems can be achieved or not for any algorithm within certain time bounds. As a consequence pac-learning certain concept classes often results in essentially tautological statements where any algorithm that selects an hypothesis consistent with the data satisfies the pac-learning criterium. Algorithmic approaches like MDL worry whether the algorithms satisfy certain performance guaranties or give practically acceptable performance.

A *decision tree* defined on $\{0,1\}^n$ is a binary tree T, each of whose internal nodes is labeled with one of n *attributes*, and whose leaves are labeled by the *class value* 0 or 1. The value of T on an input vector is obtained by traversing the tree from the root to the appropriate leaf as follows: at each internal node choose the left or right descendant depending on whether the input bit corresponding the attribute of that node is 0 or 1, respectively. See [12], page 313 for a concrete example.

Decision trees are the key concept in learning systems such as ID3 [16] and CART [2], and have found use in many applications. There is a large amount of machine learning literature on learning decision trees (see, for example, [15, 16, 17]). However, the learnability of decision trees remains at this time of writing one of the most prominent open questions in learning theory. Other than heuristic learning algorithms using principles such as minimum description length and maximum entropy, the theoretically best learning algorithm [4] pac-learns a decision tree using $O(n^{\log n})$ time and samples.

Learning decision trees in the pac sense from labeled examples usually requires finding the approximately smallest decision tree with high probability. For some time, it has been open that whether or not the decision version of computing the smallest decision tree is NP-complete. Recently, the following results were obtained in [8].

1. The decision version of computing the minimum decision tree is NP-complete.
2. In general a target decision tree cannot be approximated in polynomial time by decision trees using data consisting of labeled examples to within a factor

of $2^{\log^\delta n}$ for $\delta < 1$, unless NP is contained in DTIME[$2^{\mathrm{polylog} n}$].

This gives evidence that it may be not possible to learn decision trees by any algorithm in polynomial time satisfying the pac model requirements.

Due to its practical importance, heuristic algorithms to learn decision trees are rife. M. Wax and J. Rissanen [23] and Quinlan and Rivest [17] applied the MDL principle to learn decision trees as follows.

We are given a set of data, possibly with noise, representing a collection of examples. Each example is represented by a data item in the data set, which consists of a tuple of *attributes* followed by a binary *class value* indicating whether the example with these attributes is a positive or negative example. A consistent decision tree can be viewed as a classification procedure of the data. The internal nodes in the tree are *decision nodes.* Each such node specifies a test of a particular attribute; the possible answers are labeled on the arcs leaving the decision node to the two descendant nodes. A leaf of the tree specifies a class to which the object that passed the attribute tests and arrived at this leaf belongs. Given a set of data, we can construct many different decision trees of various sizes which agree with the data. Some data, for example noisy data, may not obey the predicting rule defined by the decision tree. We usually have a choice between using a small *imperfect* tree which classifies some data falsely or a big *perfect* tree which correctly classifies all given data.

The aim is in general to construct a decision tree that has the smallest error rate in classifying future data. Is the *smallest perfect decision tree* really a good predictor? It turns out that in *practice* this is not the case. Due to the presence of noise or inadequacy of the given attributes, selecting a perfect decision tree 'overfits' the data and gives generally poor predictions. Many *ad hoc* rules have been suggested and used for overcoming this problem.

The MDL principle appears to provide a solution and generally works well in practice. Essentially, given the data sample *without* the class values, we look for a reasonably small tree such that most data are correctly classified by the tree. We encode the inconsistent data as exceptions. We minimize the sum of

- the number of bits to encode a (not necessarily perfect) decision tree T, that is, the model which gives the $-\log P(H)$ term; and
- the number of bits to encode (describe) the examples (x_i, y_i) in the sample $(\mathbf{x}, \mathbf{y})$ that are inconsistent with T, given the entire data sample $\mathbf{x}$ without the class values $\mathbf{y}$. This gives the $-\log P(\mathbf{y}|H, \mathbf{x})$ term.

Now a coding method must be provided. This is important in applications, since it determines where the optimum is. If the encoding of trees is not efficient, then we may end up with a very small tree (with relatively large depth), and too many examples become exceptions. An inefficient encoding of the exceptions would result in overly large trees. In both cases, the prediction of unseen data is affected. Quinlan and Rivest encode a tree by making a depth-first traversal. At each internal node, write down the attribute name in some self-delimiting form followed by its edge label. At a leaf write down the class value. If the tree is not perfect, then the data that do not fit in the tree are encoded separately as exceptions (in some economical way using the provided total data sample without the class values).

Apparently, there is a problem of 'subjectivity' of encoding. More efficient ways of encoding the model (tree) may result in a larger decision tree, and more efficient methods of encoding the exceptions may result in a smaller decision tree. Quinlan and Rivest have observed that the decision trees their algorithm produced were slightly smaller than they should be. This suggests that the *model cost* is too high, that is, the model is encoded using too many bits per parameter. These trees were found by direct comparison. Nohre, [13], has described a tree pruning algorithm for binary trees to model Markov processes for data compression. This method finds the MDL subtree starting from an arbitrarily large tree. Rissanen in [22] has demonstrated how to use a sharper formula for *stochastic complexity* in applying Nohre's algorithm to overcome the tendency of producing too small decision trees. All these MDL style algorithms are reported to be quite competitive with other learning algorithms for decision trees.

6 Alternative MDL-Like Principle

In the above interpretation of MDL we essentially look for a hypothesis H minimizing $K(D|H) + K(H)$. This always satisfies

$$K(D|H) + K(H) \geq K(D).$$

Another interpretation of MDL is sometimes useful, and in fact used in [6] to recognize on-line handwritten characters. In the new approach the idea is that we we define $E := D - D_H$, where D_H is the data set classified according to H. We want to minimize

$$K(H, E) \approx K(H) + K(E|H)$$

over H. That is, E denotes the subset of the data sample D which are *exceptions* to H in the sense of being 'not covered' by H. We want to find H such that the description of H and the exception data E *not covered* by H is minimized. Note that in this case always

$$\min_H \{K(H) + K(E|H)\} \leq K(\emptyset) + K(D) = K(D),$$

in contrast to the standard interpretation of MDL above. This incarnation of MDL is not straightforwardly derived by our approach above. We may interpret it that we look for the shortest description of an accepting program for the data consisting of a classification rule H and an exception list E. While this principle sometimes gives good results, application may lead to absurdity as the following shows.

In many problems we our data sample D consists of only positive examples. Suppose we want to learn (a grammar for) English language, given a corpus of data D. Here, data D can be the Bible or an even larger sample like the so-called 'Brown Corpus'. Then, according to our new MDL rule the best hypothesis is the trivial grammar H generating *all* sentences over the alphabet. Namely, this grammar gives $K(H) = O(1)$ independent of D and alsp $E := \emptyset$. Consequently,

$$\min_H \{K(H) + K(E|H)\} = K(H) = O(1),$$

which is absurd.

References

1. D. Angluin, Computational learning theory: survey and selected bibliography, *Proc. 24th ACM Symp. Theory Comp.*, 1992, 319-342.
2. L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Wadsworth International Group, Belmont, CA (1984).
3. W. Buntine, Personal communication, September 1994.
4. A. Ehrenfeucht and D. Haussler, Learning decision trees from random examples. *Proc. 1988 Workshop on Comp. Learning Theory*, Morgan-Kaufmann, 1988, 182-195.
5. P. Gács, On the symmetry of algorithmic information, *Soviet Math. Dokl.*, 15(1974), 1477-1480. Correction, Ibid., 15(1974), 1480.
6. Q. Gao and M. Li, An application of minimum description length principle to on-line recognition of handprinted alphanumerals, *11th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1989, 843-848.
7. E.M. Gold, Language identification in the limit, *Inform. Contr.*, 10(1967), 447-474.
8. T. Hancock, T. Jiang, M. Li, and J. Tromp, Lower bounds for learning decision lists and trees. *Proc. 12th Annual Symp. Theoret. Aspects of Comput. Sci.* (STACS'95), Lecture Notes in Computer Science, Springer-Verlag, Heidelberg, 1995, 527-538.
9. A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission*, 1:1(1965), 1-7.
10. M. Li and P.M.B. Vitányi, Inductive reasoning and Kolmogorov complexity, *J. Comput. Syst. Sci.*, 44(1992), 343-384.
11. M. Li and P. Vitányi, A theory of learning simple concepts under simple distributions, *SIAM J. Computing*, 20:5(1991), 915-935.
12. M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, 1993.
13. R. Nohre, Some Topics in Descriptive Complexity, Ph.D. Thesis, Linkoping University, Linkoping, Sweden, 1994.
14. D. MacKay, pp 59 in *Maximum Entropy and Bayesian Methods*, Kluwer, 1992. (Personal communication W. Buntine.)
15. J. Mingers, An Empirical Comparison of Selection Measures for Decision-Tree Induction. *Machine Learning*, 3(1989), 319-342.
16. J.R. Quinlan, Induction of decision trees, *Machine Learning*, 1(1986), 81-106.
17. J.R. Quinlan and R. Rivest, Inferring decision trees using the minimum description length principle, *Inform. Computation*, 80(1989), 227-248.
18. J. Rissanen, Modeling by the shortest data description, *Automatica-J.IFAC*, 14(1978), 465-471.
19. J. Rissanen, Universal coding, information, prediction and estimation, *IEEE Transactions on Information Theory*, IT-30(1984), 629-636.
20. J. Rissanen, *Stochastic Complexity and Statistical Inquiry*, World Scientific Publishers, 1989.

21. J. Rissanen, Stochastic complexity, J. Royal Stat. Soc., series B, 49(1987), 223-239. Discussion: pp. 252-265.
22. J. Rissanen, Stochastic complexity in learning, *Proc. 2nd European Conf. on Computational Learning Theory*, Lecture Notes in Artificial Intelligence, Vol. 904, P. Vitányi, Ed., Springer Verlag, Heidelberg, 1995, 196-210.
23. J. Rissanen and M. Wax, Algorithm for constructing tree strucutured classifiers, US Patent No. 4719571, 1988.
24. J. Segen, *Pattern-Directed Signal Analysis*, PhD Thesis, Carnegie-Mellon University, 1980.
25. R.J. Solomonoff, Complexity-based induction systems: comparisons and convergence theorems, *IEEE Trans. Inform. Theory*, IT-24(1978), 422-432.
26. L.G. Valiant, A Theory of the Learnable, *Comm. ACM*, 27(1984), 1134-1142.
27. V. Vovk, Minimum description length estimators under the universal coding scheme, *Proc. 2nd European Conference on Computational Learning Theory*, Lecture Notes in Artificial Intelligence, Vol. 904, P. Vitányi, Ed., Springer Verlag, Heidelberg, 1995, 237-251.
28. C.S. Wallace and D.M. Boulton, An information measure for classification, *Computing Journal*, 11(1968), 185-195.
29. C.S. Wallace and P.R. Freeman, Estimation and inference by compact coding, *J. Royal Stat. Soc., Series B*, 49(1987), 240-251, Discussion: 252-265.
30. K. Yamanishi, Approximating the minimum description length and its applications to learning, Manuscript NEC Research Labs, New Jersey, 1995.
31. A.K. Zvonkin and L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, *Russian Math. Surveys*, 25:6(1970), 83-124.