Complexity Approximation Principle and Rissanen's Approach to Real-Valued Parameters

Yuri Kalnishkan*

Department of Computer Science, Royal Holloway, University of London
Egham, Surrey, TW20 0EX, United Kingdom
yura@dcs.rhbnc.ac.uk

Abstract. In this paper an application of the Complexity Approximation Principle to the non-linear regression is suggested. We combine this principle with the approximation of the complexity of a real-valued vector parameter proposed by Rissanen and thus derive a method for the choice of parameters in the non-linear regression.

1 Introduction

The Complexity Approximation Principle (CAP) was proposed in the paper [9] and it deals with the hypothesis selection problem. CAP is one of the implementations of the idea to trade-off the 'goodness-of-fit' of a hypothesis against its complexity. This idea goes back to the the celebrated Occam's razor and the scope of its implementations includes MDL and MML principles.

In this paper, we make an attempt to apply CAP to the choice of coefficients in the non-linear regression. The problem of evaluating the complexity of a real-valued vector emerges and we overcome it by adapting the approach proposed by Rissanen in [3]. We infer a formula that suggests a new estimate of the regression coefficients and it turns out to be a normalisation of the Least Squares (LS) estimate.

In Sect. 2 we formulate CAP in the form we need and describe the non-linear regression problem. In Sect. 3 we apply Rissanen's approach and obtain the minimisation problem; in Sect. 4 and 5 we discuss possible solutions. Sect. 6 contains the description and the results of our computational experiments. We compare our results with other regression techniques.

2 Preliminaries

2.1 CAP

In this paper, the special case of CAP relevant to the batch settings and the square-loss measure of discrepancy is considered. We will now formulate CAP in this weak form.

* Supported partially by EPSRC through the grant GR/M14937 ("Predictive complexity: recursion-theoretic variants") and by ORS Awards Scheme.

R. López de Mántaras, E. Plaza (Eds.): ECML 2000, LNAI 1810, pp. 203–210, 2000.

Suppose we are given a data sequences $z = ((x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)) \in (X \times \Omega)^*$, where $\Omega = [a, b]$ is the set of *outcomes* and X is the set of *signals*. Our goal is to find the decision rule $\mathcal{R} : X \to \mathbb{R}$ that suits the data best in a given class of decision rules $\mathfrak{R}$. The performance of $\mathcal{R}$ is assessed by some measure of loss or discrepancy $\lambda(\mathcal{R}(x), y)$, where y is the actual outcome which corresponds to the signal x. We assume that $\lambda(\mathcal{R}(x), y) = (\mathcal{R}(x) - y)^2$. We want $\mathcal{R}$ to perform well i.e. to suffer small loss on pairs (*signal*, *outcome*) $\in X \times \Omega$ that may arrive in the future.

The classical Least Squares (LS) approach suggests minimising the total square loss of $\mathcal{R}$ on the sequence z, i.e. $\mathrm{Loss}^{\mathrm{sq}}_{\mathcal{R}}(z) = \sum_{i=1}^{l} (\mathcal{R}(x_i) - y_i)^2$. A decision rule $\widehat{\mathcal{R}} \in \mathfrak{R}$ is called a *LS estimate* if the minimum

$$\min_{\mathcal{R} \in \mathfrak{R}} \mathrm{Loss}^{\mathrm{sq}}_{\mathcal{R}}(z) \tag{1}$$

is attained at $\widehat{\mathcal{R}}$.

LS works perfectly well in many applications unless the problem of overfitting occurs. The given data z may be influenced by noise or round-off error so following carefully all the peculiarities of our data we may end up with an estimate which makes no sense. That is why the idea to penalise the growth of complexity of $\mathcal{R}$ emerges. Instead of finding the minimum (1), one may search for $\mathcal{R}$ minimising

$$\min_{\mathcal{R} \in \mathfrak{R}} \left(\mathrm{Loss}_{\mathcal{R}}(z) + \mathcal{K}(\mathcal{R}) \right) \quad , \tag{2}$$

where $\mathcal{K}$ is some measure of complexity of $\mathcal{R}$. In the papers of Rissanen (see e.g. [3,4], or [5], various formulae analogous to (2) were investigated. These papers deal with the problem of choice of a probabilistic model and thus the measures of loss $\mathrm{Loss}_{\mathcal{R}}(z)$ different from the square loss $\mathrm{Loss}^{\mathrm{sq}}_{\mathcal{R}}(z)$ are considered there.

The paper [9] provides both a motivation and a refinement to (2). In the paper [10], a value $\mathcal{K}^{\mathrm{sq}}(z)$ called the *predictive (square-loss) complexity* of z is introduced and [9] shows that the inequality

$$\mathcal{K}^{\mathrm{sq}}(z) \le \mathrm{Loss}^{\mathrm{sq}}_{\mathcal{R}}(z) + \frac{(b-a)^2 \ln 2}{2} KP(\mathcal{R}) + C \tag{3}$$

holds for any computable decision rule $\mathcal{R}$, where KP stands for the prefix complexity (for definitions see [2]) and the constant C does not depend upon z and $\mathcal{R}$. CAP suggests minimising the right-hand side of (3), i.e. $\widetilde{\mathcal{R}}$ is called a CAP estimate if the minimum

$$\min_{\mathcal{R} \in \mathfrak{R}} \left(\mathrm{Loss}^{\mathrm{sq}}_{\mathcal{R}}(z) + \frac{(b-a)^2 \ln 2}{2} KP(\mathcal{R}) \right) \tag{4}$$

is attained at $\widetilde{\mathcal{R}}$.

2.2 Non-linear Regression

To construct the set of decision rules for the non-linear regression, we start with a sequence of functions $f_1, f_2, \ldots$ which map X into $\mathbb{R}$. The set $\mathfrak{R}$ consists of all finite linear combinations $\theta_1 f_1 + \theta_2 f_2 + \ldots + \theta_k f_k$, where $\theta = (\theta_1, \theta_1, \ldots, \theta_k)^T$ is a finite-dimensional real-valued column vector. If we fix a dimension $k \in \mathbb{N}$, we will obtain the set $\mathfrak{R}_k$. Clearly, $\mathfrak{R} = \bigcup_{i=1}^{\infty} \mathfrak{R}_k$. A decision rule $\mathcal{R}$ may be identified with the corresponding θ and therefore we may identify $\mathfrak{R}$ with $\mathbb{R}^*$ and $\mathfrak{R}_k$ with $\mathbb{R}^k$. Let us introduce k–dimensional string vectors $F_i^{(k)} = (f_1(x_i), f_2(x_i), \ldots, f_k(x_i))^T$, where $1 \le i \le l$ and $1 \le k < +\infty$, and $(l \times k)$–matrixes $F^{(k)}$ such that the element $F_{i,j}^{(k)}$ equals the j–th coordinate of $F_i^{(k)}$, where $1 \le i \le k$, $1 \le j \le l$, and $1 \le k < +\infty$. Let us also introduce a column vector $Y = (y_1, y_2, \ldots, y_l)^T$.

In the k-dimensional case, the LS formula (1) reads as

$$\min_{\theta \in \mathbb{R}^k} \sum_{i=1}^{l} (F_i^{(k)} \theta - y_i)^2 \ . \tag{5}$$

If $l \ge k$, then the k–dimensional LS estimate $\widehat{\theta}^{(k)}$ is given by the equation

$$\hat{\theta}^{(k)} = \left(\left(F^{(k)} \right)^T F^{(k)} \right)^{-1} \left(F^{(k)} \right)^T Y \tag{6}$$

(see e.g. [1]).

If we minimise (5) over k as well, we will probably either come to no solution or come to a solution corresponding to the exact fit. Another disadvantage of (5) is that it does not penalise the growth of coordinates of θ.

In the k–dimensional case, CAP formula (4) reads as

$$\min_{\theta \in \mathbb{R}^k} \left(\sum_{i=1}^{l} (F_i^{(k)} \theta - y_i)^2 + \frac{(b-a)^2 \ln 2}{2} KP(\theta \mid k) \right) \tag{7}$$

and in the case of the unbounded (*finite*) dimension we get

$$\min_{\theta \in \mathbb{R}^*} \left(\sum_{i=1}^{l} (F_i^{(d(\theta))} \theta - y_i)^2 + \frac{(b-a)^2 \ln 2}{2} KP(\theta) \right) \ , \tag{8}$$

where $d(\theta)$ denotes the dimension of θ. The problem is to find a natural approximation of $KP(\theta)$ and we are discussing this problem in the next section.

3 Complexity of Real-Valued Parameters

In this section, we apply the estimate of the complexity of θ proposed in [3]. As we mentioned above, [3] deals with probabilistic models rather than with the problem of regression but the expression considered in [3] may be regarded

as a special case of general formula (2). The main result of [3] is the following approximation of the complexity of $\theta \in \mathbb{R}^k$:

$$\mathcal{K}(\theta \mid k) \approx \log^* \left[C(k) \|\theta\|^k_{\mathrm{Loss}} \right] \quad . \tag{9}$$

Here $\log^* \alpha$ stands for $\log_2 \alpha + \log_2 \log_2 \alpha + \ldots$, where only positive terms are included, $C(k)$ stands for the volume of the k–dimensional unit ball, and $\|\theta\|_{\mathrm{Loss}}$ denotes the norm of θ induced by the second derivative of $\mathrm{Loss}_\theta(z)$ taken at the point corresponding to the 'maximum likelihood' estimate, i.e.

$$\|\theta\|_{\mathrm{Loss}} = \sqrt{\theta^T \left(D^2_\theta \mathrm{Loss}_\theta(z)|_{\theta=\widehat{\theta}} \right) \theta} \ , \tag{10}$$

where the minimum $\min_{\varphi \in \mathbb{R}^k} \mathrm{Loss}_\varphi(z)$ is achieved at $\varphi = \widehat{\theta}$.

As one can see, the formula is not independent of the minimisation problem (2) we are going to solve. The derivation of (10) may be outlined as follows. The estimate given by (2) is supposed to be close to the 'maximum likelihood' estimate $\widehat{\theta}$ so $\mathrm{Loss}_\theta(z)$ may be replaced by its second order approximation in the neighbourhood of $\widehat{\theta}$. Then $\mathbb{R}^k$ is split into small rectangles such that inside each rectangle the approximation of $\mathrm{Loss}_\theta(z)$ takes values which are sufficiently close to each other and then the rectangles are enumerated according to the 'spiral fashion'.

We will now apply (10) to our problems (7) and (8). One may easily see that

$$D^2_\theta \sum_{i=1}^{l} (F_i^{(k)}\theta - y_i)^2 = \sum_{i=1}^{l} \left(F_i^{(k)} \right)^T F_i^{(k)} \ , \tag{11}$$

i.e. we obtain the sum of outer (Kronecker) products. Hence

$$\|\theta\|^2_{\mathrm{Loss}} = \sum_{i=1}^{l} \left[\theta^T \left(F_i^{(k)} \right)^T F_i^{(k)} \theta \right] = \sum_{i=1}^{l} \left(F_i^{(k)} \theta \right)^2 \ . \tag{12}$$

Caring out the substitution, we obtain the following k–dimensional minimisation problem:

$$\min_{\theta \in \mathbb{R}^k} \left(\sum_{i=1}^{l} (F_i^{(k)}\theta - y_i)^2 + \frac{(b-a)^2 \ln 2}{2} \log^* C(k) \left[\sum_{i=1}^{l} \left(F_i^{(k)} \theta \right)^2 \right]^{k/2} \right) \ . \tag{13}$$

If we approximate $KP(\theta)$ by $KP(\theta \mid k) + KP(k)$ and $KP(k)$ by $\log^*(k)$ (see [3] and [2]), where $k = d(\theta)$ is the dimension of θ, we will obtain the general formula

$$\min_{\theta \in \mathbb{R}^*} \left(\sum_{i=1}^{l} (F_i^{(d(\theta))}\theta - y_i)^2 + \right.$$
$$\left. \frac{(b-a)^2 \ln 2}{2} \log^* C(k) \left[\sum_{i=1}^{l} \left(F_i^{(k)} \theta \right)^2 \right]^{k/2} + \frac{(b-a)^2 \ln 2}{2} \log^* d(\theta) \right) \ . \tag{14}$$

The last term $\frac{(b-a)^2 \ln 2}{2} \log^* d(\theta)$ in (14) guarantees the existence of a minimum as long as long as minimums in (13) exist.

4 Minimisation

One can easily that the parameter θ appears in (13) only in the products $F_i^{(k)}\theta$, where $1 \leq i \leq l$, so one may introduce the new vector parameter $a = F^{(k)}\theta$ ranging over the subspace $\operatorname{Im} F^{(k)} \subseteq \mathbb{R}^l$. Therefore (13) has the form

$$\min_{a \in \operatorname{Im} F^{(k)}} \left(\|a - Y\|^2 + f(\|a\|) \right) \ , \tag{15}$$

where $\|a\|$ stands for the standard Euclidean norm $\|a\| = \sqrt{a_1^2 + a_2^2 + \ldots + a_l^2}$ and f is a real-valued function of a real-valued parameter. It follows easily that the minimum is attained at $\tilde{a}$ collinear to the projection $\hat{Y}$ of Y on $\operatorname{Im} F^{(k)}$. Namely if x_0 is the solution of

$$\min_{x \in \mathbb{R}} \left((x - \|\hat{Y}\|)^2 + f(x) \right) \ , \tag{16}$$

then the minimum in (15) is achieved at $\tilde{a} = \frac{x_0}{\|\hat{Y}\|}\hat{Y}$. It follows from the definition of the LS estimate $\hat{\theta}^{(k)}$, that $\hat{Y} = F^{(k)}\hat{\theta}^{(k)}$ hence, by linearity, the minimum in (13) is achieved at

$$\tilde{\theta}^{(k)} = \frac{x_0}{\|F^{(k)}\hat{\theta}^{(k)}\|}\hat{\theta}^{(k)} \ . \tag{17}$$

In statistics, there exists a qualitative analogy to this formula. Stein's paradox suggest normalising the Maximum Likelihood estimate in the case of the normal distribution and the square loss. See [8,7] for details.

5 Dual Variables

Suppose that the number of parameters k exceeds l the number of given examples. In this case, there is no unique LS estimate is not unique. We have many vectors θ which correspond to the exact fit. Formula (13) does not include θ unless it is multiplied by $F_i^{(k)}$ and therefore (13) does not allow us to distinguish between different sets of parameters that still give equal predictions on the training set. Hence, $\tilde{\theta}^{(k)}$ from (17) provides a solution for (13) if $\hat{\theta}^{(k)}$ suffers zero loss on z.

It is natural to choose $\theta \in \mathbb{R}^k$ with the smallest Euclidean norm $\|\theta\|$. Such θ is given by the method of dual variables (see, e.g. [6]). According to this method, the value $\widehat{\mathcal{R}}(x) = \sum_{i=1}^k \theta_i^{(k)} f_i(x)$ of the decision rule $\widehat{\mathcal{R}}$ corresponding to the LS estimate with the smallest value of $\|\theta\|$ on a signal x is given by the formula

$$\widehat{\mathcal{R}}(x) = Y^T \left(\mathbf{K}^{(k)} \right)^{-1} \mathbf{k}^{(k)}(x) \ , \tag{18}$$

where $\mathbf{K}^{(k)}$ is an $(l \times l)$–matrix such that $\mathbf{K}_{i,j}^{(k)} = K^{(k)}(x_i, x_j)$ for $1 \leq i, j \leq l$, $\mathbf{k}^{(k)}(x)$ is an l–dimensional vector such that $\mathbf{k}_i^{(k)}(x) = K^{(k)}(x_i, x)$ for $1 \leq i \leq l$,

and $K^{(k)} : X \times X \to \mathbb{R}$ is the *kernel* associated with the non-linear regression problem under consideration, i.e.

$$K^{(k)}(x', x'') = \sum_{i=1}^{k} f_i(x') f_i(x'') \ . \tag{19}$$

Note that the size of $\mathbf{K}^{(k)}$ does not increase with the increase of k.

6 Experiments and Discussion

6.1 Toy Examples

We consider the following one-dimensional toy problem. Consider the function $y = \sin(x)$ on the interval $[-A, A]$ and the Gaussian noise $\xi \sim \mathcal{N}(0, \sigma^2)$. Our approach requires tight bounds so we must bound the range of the noise. If $\sin(x) + \xi$ falls outside the interval $[-1, 1]$, we replace it by the nearest number, either 1 or -1. Both training and test examples are taken according to the uniform distribution. We try to approximate the data by k-dimensional polynomials.

We calculate the LS estimate by (1), normalise it according to (17), and compare the difference. The main empirical result here may be formulated in the following way. Formula (17) overperforms the simple LS estimate on very 'complicated' and 'noisy' problems, i.e. cases with large values of A and σ^2 and small numbers of training examples. Otherwise our correction can only spoil the LS estimate.

Fig. 6.1 shows the squared loss on the training set of size 100 averaged over 1000 independent trials. The results correspond to the case $A = 6$, $\sigma^2 = 0.5$, the size of training sets equals 25. You may see that this case is very difficult and the best LS estimates of degree 3 perform only slightly better than those of degree 0, i.e. constant predictions.

Unfortunately, experiments with (14) failed. The graph of complexity with respect to the degree exhibits an increasing pattern and does not allow to locate the optimal degree.

6.2 Boston Housing

The Boston Housing database (available at `ftp://ftp.ics.uci.com/pub/machine-learning-databases/housing`) is often used to test different non-linear regression techniques (see e.g. [6]). The entries of this database are strings of 14 parameters which describe houses in different neighbourhoods of Boston. The last elements of these strings are prices of houses in thousands of dollars, ranging from 5 to 50. We use prices as outcomes in our experiments.

We use the polynomial kernel

$$K^{(k)}(x', x'') = \left(x' \left(x'' \right)^T + 1 \right)^d \tag{20}$$

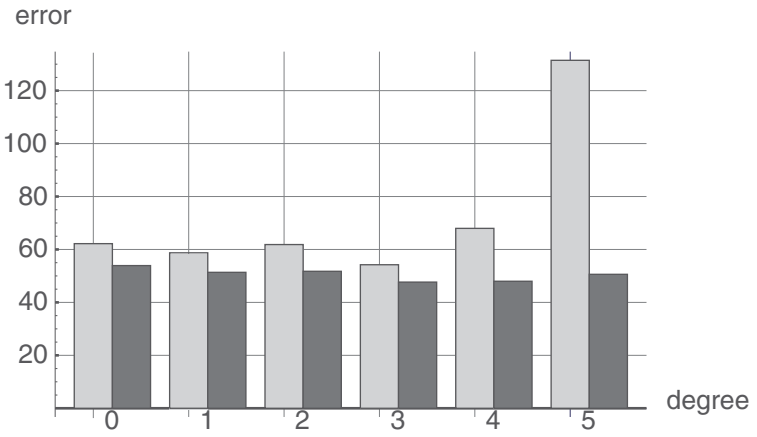

Fig. 1. The results on a toy example. The error of LS estimates is represented by light gray bars and the error of our method is represented by black ones

which correspond to the approximating of data by sums of normalised monomials of degree smaller then or equal to d. Following [6], we concentrate on $d = 5$. Our methodology is also similar to [6], we pick test sets of size 480 and training sets of size 25 randomly and repeat the procedure for 100 times.

The dimension k of the set of monomials equals $\binom{13+d}{13}$ but this natural assignment turns out to be meaningless. The correction coefficient we get is very close to 1 and it improves the performance of the algorithm by around a ten thousandth of a percent.

We may also consider using a kernel K as approximating the data by a linear combination of $\mathbf{k}_i^{(k)}(x) = K^{(k)}(x_i, x)$ (see Sect. 5). In this case, the dimension equal the size of the training set, which is much smaller.

If we make this assumption about the degree, the results become much more reasonable. Our method improves the performance by 14.7% (we obtain the average square loss over 100 trials equal to 69.1 against 81.0).

We must admit that our method is still no match to the ridge regression. The idea of the ridge regression (see, e.g. [6]) is to introduce an extra term to (18), i.e. to consider

$$\widehat{\mathcal{R}}(x) = Y^T \left(\mathbf{K}^{(k)} + aI\right)^{-1} \mathbf{k}^{(k)}(x) \ , \tag{21}$$

where $a > 0$ and I is the unit matrix. The paper [6] show that under the same settings the ridge regression is able to decrease the mistake down to 10.4.

Despite its theoretical justification, our method turns out to be much more rough than the ridge regression. In fact, ridge regression performs the same

task of penalising the growth of coefficients. The solution, given by the ridge regression minimises the expression $a\|\theta\| + \mathrm{Loss}_\theta^{\mathrm{sq}}$. This approach, motivated by empirical considerations, proves to be very sound.

7 Acknowledgements

I would like to thank Prof. V. Vovk and Prof. A. Gammerman for providing guidance to this work.

References

1. Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. The MIT Press, 1990. 205
2. M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 2nd edition, 1997. 204, 206
3. J Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983. 203, 204, 205, 206
4. J Rissanen. Stochastic complexity. *Journal of Royal Statistical Society*, 49(3):223–239, 1987. 204
5. J Rissanen. Stochastic complexity in learning. *Journal of Computer and System Sciences*, 55:89–95, 1997. 204
6. C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521, 1998. 207, 208, 209
7. Mark J. Schervish. *Theory of Statistics*. Springer–Verlag, 1995. 207
8. Gábor J. Székely. *Paradoxes in Probability Theory and Mathematical Statistics*. Akadémiai Kiadó Budapest, 1986. 207
9. V. Vovk and A. Gammerman. Complexity approximation principle. *The Computer Journal*, 42(4):318–322, 1999. 203, 204
10. V. Vovk and C. J. H. C. Watkins. Universal portfolio selection. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 12–23, 1998. 204