Optimal Bayesian 2D-discretization for variable ranking in regression

Marc Boullé and Carine Hue

France Télécom R&D Lannion,
Firstname.Name@orange-ft.com

Abstract. In supervised machine learning, variable ranking aims at sorting the input variables according to their relevance w.r.t. an output variable. In this paper, we propose a new relevance criterion for variable ranking in a regression problem with a large number of variables. This criterion comes from a discretization of both input and output variables, derived as an extension of a Bayesian non parametric discretization method for the classification case. For that, we introduce a family of discretization grid models and a prior distribution defined on this model space. For this prior, we then derive the exact Bayesian model selection criterion. The obtained most probable grid-partition of the data emphasizes the relation (or the absence of relation) between inputs and output and provides a ranking criterion for the input variables. Preliminary experiments both on synthetic and real data demonstrate the criterion capacity to select the most relevant variables and to improve a regression tree.

1 Introduction

In a data mining project, the data preparation step aims at providing a dataset for the modeling step [CCK+00]. Variable (or feature) selection consists in selecting a subset of the variables which is useful for a given problem. This selection process is an essential part of data preparation, which becomes critical in case of databases having large numbers of variables (order of thousands of). Indeed, the risk of overfitting the data quickly increases with the number of input variables, which is known as the curse of dimensionality. The objective of variable selection is three-fold: to improve the performance of predictors, to provide faster and more cost-effective predictors and to allow an easier interpretation of the prediction [GE03]. Variable selection methods generally use three ingredients [LG99]: a criterion to evaluate the relevance of a variable subset and compare variable subsets, a search algorithm to explore the space of all possible variable subsets and a stopping criterion. Variable selection is often linked to variable ranking which aims at sorting the variables according to their relevance. These two problems clearly differ as a subset of useful variables may exclude redundant but relevant variables. Conversely, the subset of the most relevant variables can be suboptimal among the subsets of equal size. Compared to variable selection, variable ranking is much more simple as it does not need any search algorithm

but only the evaluation of the relevance criterion for each variable. For linear dependencies, the classical relevance criterion is the correlation coefficient or its square. To capture non linear dependencies, the mutual information is more appropriate but it needs estimates of the marginal and joint densities which are hard to obtain for continuous variables.
In this paper, we introduce a new relevance criterion for variable ranking in a regression problem with a large number of input variables. This criterion is based on the discretization of both the input and the output variables. Discretization has been widely studied in the case of supervised classification [Cat91] [Hol93] [DKS95] [LHTD02]. Our discretization method for regression extends our discretization method for the classification case to deal with numeric output variables. We apply a non parametric Bayesian approach to find the most probable discretization given the data. Owing to a precise definition of the space of discretization models and to a prior distribution on this model space, we derive a Bayes optimal evaluation criterion. We then use this criterion to evaluate each input variable and rank them. Besides a new variable ranking criterion, our method provides a robust discretization-based interpretation of the dependence between each input variable and the output variable and an estimator of the conditional densities for the considered regression problem.
The remainder of the document is organized as follows. Section 2 presents our discretization method for regression, with its criterion and optimization algorithm. In Section 3, we show preliminary experimental results both on synthetic and real data.

2 The MODL Discretization Method for Regression

We begin by recalling the principles of the Bayesian approach and the MDL approach [Ris78] for the model selection problem. We then present our approach (called MODL) which results in a Bayesian evaluation criterion of discretizations and the greedy heuristic used to find a near Bayes optimal discretization. We first present the principle of our discretization method for classification, and then extend it to the case of regression.

2.1 Bayesian versus MDL model selection techniques

In the Bayesian approach, the searched model is the one which maximizes the probability $p(Model|Data)$ of the model given the data. Using Bayes rule and since the probability is constant while varying the model, this is equivalent to maximizing:

$$p(Model)p(Data|Model) \tag{1}$$

Given a *prior* distribution of the models, the searched model can be obtained provided that the calculation of the probabilities $p(Model)$ and $p(Data|Model)$ is feasible. For classical parametric model families, these probabilities are generally intractable and the Bayesian Information Criterion (BIC) [Sch78] is a

well-known penalized Bayesian selection model criterion. As detailed in the sequel, our approach, called MODL, conducts to an exact Bayesian model selection criterion.
To introduce the MDL approach, we can reuse the Bayes rule, replacing the probabilities by their negative logarithms. These negative logarithms of probabilities can be interpreted as Shannon code lengths, so that the problem of model selection becomes a coding problem. In the MDL approach, the problem of model selection is to find the model that minimizes:

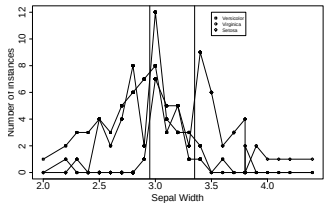
$$DescriptionLength(Model) + DescriptionLength(Data|Model) \quad (2)$$

The relationship between the Bayesian approach and the MDL approach has been examined by [VL00]. The Kolmogorov complexity of an object is the length of the shortest program encoding an effective description of this object. It is asymptotically equal to the negative log of a probability distribution called the *universal distribution*. Using these notions, the MDL approach turns into *ideal MDL*: it selects the model that minimizes the sum of the Kolmogorov complexity of the model and of the data given the model. It is asymptotically equivalent to the Bayesian approach with a universal prior for the model. The theoretical foundations of MDL allow focusing on the coding problem: it is not necessary to exhibit the prior distribution of the models. Unfortunately, the Kolmogorov complexity is not computable and can only be approximated.
To summarize, the Bayesian approach allows selecting the optimal model relative to the data, once a prior distribution of the models is fixed. The MDL approach does not need to define an explicit prior to find the optimal model, but the optimal description length can only be approximated and the approach is valid asymptotically.

2.2 The MODL approach in supervised classification

The objective of a supervised discretization method is to induce a list of intervals which splits the numerical domain of a continuous input variable, while keeping the information relative to the output variable. A compromise must be found between information quality (homogeneous intervals in regard to the output variable) and statistical quality (sufficient sample size in every interval to ensure generalization). For instance, we present on left of Figure 1 the number of instances of each class of the Iris dataset w.r.t the sepal width variable. The problem is to find the split of the domain $[2.0, 4.4]$ in intervals which gives us optimal information about the repartition of the data between the three classes.

In the MODL approach [Bou06], the discretization is turned into a model selection problem. First, a space of discretization models is defined. The parameters of a specific discretization are the number of intervals, the bounds of the intervals and the output frequencies in each interval. Then, a prior distribution is proposed on this model space. This prior exploits the hierarchy of the parameters: the number of intervals is first chosen, then the bounds of the intervals and finally the output frequencies. The choice is uniform at each stage of the



	]2.0,2.95[	[2.95, 3.35[	[3.35, 4.4[
Versicolor	34	15	1
Virginica	21	24	5
Setosa	2	18	30
Total	57	57	36

Fig. 1. MODL discretization of the Sepal Width variable for the classification of the Iris dataset in 3 classes

hierarchy. A Bayesian approach is applied to select the best discretization model, which is found by maximizing the probability $p(Model|Data)$ of the model given the data. Using the Bayes rule and since the probability $p(Data)$ is constant under varying the model, this is equivalent to maximize $p(Model)p(Data|Model)$. Let N be number of instances, J the number of output values, I the number of intervals for the input domain. $N_{i.}$ denotes the number of instances of input value in the interval i (total per column), $N_{.j}$ is the number of instances of class j (total per raw), and N_{ij} the number of instances of output value j in the interval i. In the context of supervised classification, the number of classes J and the number of instances per class $N_{.j}$ are supposed known. A discretization model is then defined by the parameter set $\left\{I, \{N_i\}_{1\leq i\leq I}, \{N_{ij}\}_{1\leq i\leq I, 1\leq j\leq J}\right\}$. We remark that the data partition obtained by applying such a discretization model is invariant by any monotonous variable transformation since it only depends on the variable ranks. Owing to the definition of the model space and its prior distribution, the Bayes formula is applicable to exactly calculate the prior probabilities of the models and the probability of the data given a model. Taking the negative log of the probabilities, this provides the evaluation criterion given in formula (3):

$$\log N + \log \binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log \binom{N_i+J-1}{J-1} + \sum_{i=1}^{I} \log \frac{N_i!}{N_{i,1}!N_{i,2}!\ldots N_{i,J}!} \tag{3}$$

The first term of the criterion stands for the choice of the number of intervals and the second term for the choice of the bounds of the intervals. The third term corresponds to the choice of the output distribution in each interval and the last term encodes the probability of the data given the model. The complete proof can be found in [Bou06].

Once the optimality of the evaluation criterion is established, the problem is to design a search algorithm in order to find a discretization model that minimizes the criterion. In [Bou06], a standard greedy bottom-up heuristic is used to find a good discretization. The method starts with initial single value intervals and then searches for the best merge between adjacent intervals. The best merge is performed if the MODL value of the discretization decreases after the merge and the process is reiterated until no further merge can decrease the criterion.

In order to further improve the quality of the solution, the MODL algorithm performs post-optimizations based on hill-climbing search in the neighborhood of a discretization. The neighbors of a discretization are defined with combinations of interval splits and interval merges. Overall, the time complexity of the algorithm is $O(JNlog(N))$. The MODL discretization method for classification provides the most probable discretization given the data sample. Extensive comparative experiments report high quality performance. For the example given, the three obtained intervals are shown on left of Figure 1. The contingency table on the right gives us comprehensible rules such as "for a sepal width in $[2.0, 2.95]$, the probability of occurence of the Versicolor class is $34/57 = 0,60$".

2.3 Extending the approach to regression

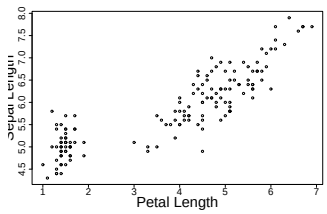


Fig. 2. Scatter-plot of the Petal Length and Sepal Length variables of the Iris dataset

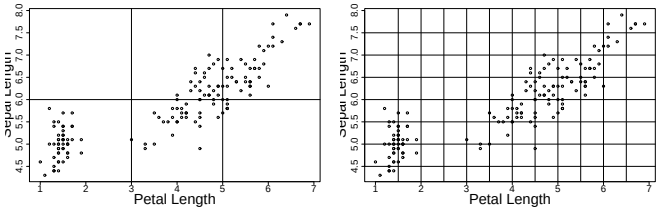


Fig. 3. Two discretization grids with 6 or 96 cells, describing the correlation between the Petal Length and Sepal Length variables of the Iris dataset.

In order to illustrate the regression problem, we present in figure 2 the scatter-plot of the Petal Length and Sepal Length variables of the Iris dataset [Fis36]. The figure shows that Iris plants with petal length below 2 cm always have a sepal length below 6 cm. If we divide the sepal length values into two output intervals of values (below or beyond 6 cm), we can provide rules to describe the correlation between the input and the output variable. The regression problem is now turned into a classification problem. In this context, the objective of the

MODL 2D-discretization method is to describe the distribution of the output intervals given the rank of the input value. This is extented to the regression case, where the issue is now to describe the rank of the output value given the rank of the input value. Discretizing both the input and output variable allows such a description, as shown in figure 3. The problem is still a model selection problem. Compared to the classification case, one additional parameter has to be optimized: the number of output intervals. A compromise has to be found between the quality of the correlation information and the generalization ability, on the basis of the grain level of the discretization grid. Let us now formalize this approach using a Bayesian model selection approach. A regression discretization model is defined by the parameter set $\left\{I,\ J,\ \{N_i\}_{1\leq i\leq I}\,,\ \{N_{ij}\}_{1\leq i\leq I,\ 1\leq j\leq J}\right\}$. Unlike the supervised classification case, the number J of intervals in the output domain is now unknown but the number of instances $N_{.j}$ can be deduced by adding the N_{ij} for each interval. We adopt the following prior for the parameters of regression discretization models:

1. the numbers of intervals I and J are independent from each other, and uniformly distributed between 1 and N,
2. for a given number of input intervals I, every set of I interval bounds are equiprobable,
3. for a given input interval, every distribution of the instances on the output intervals are equiprobable,
4. the distributions of the output intervals on each input interval are independent from each other,
5. for a given output interval, every distribution of the rank of the output values are equiprobable.

The definition of the regression discretization model space and its prior distribution leads to the evaluation criterion given in formula (4) for a discretization model M:

$$\begin{aligned} c_{reg}(M) = 2\log(N) + \log\binom{N+I-1}{I-1} + \sum_{i=1}^{I}\log\binom{N_i+J-1}{J-1} \\ + \sum_{i=1}^{I}\log\frac{N_i!}{N_{i,1}!N_{i,2}!\ldots N_{i,J}!} + \sum_{j=1}^{J}\log N_j! \end{aligned} \tag{4}$$

Compared with the classification case, there is an additional $log(N)$ term which encodes the choice of the number of output intervals, and a new last term (sum of $\log(N_j!)$) which encodes the distribution of the output ranks in each output interval. To give a first intuition, we can compute that for $I = J = 1$ the criterion value is $2\log(N) + \log(N!)$ (about 615 for $N = 150$) and for $I = J = N$ it gives $2\log(N) + \log\binom{2N-1}{N-1} + N\log(N)$ (about 224 for $N = 150$).
We adopt a simple heuristic to optimize this criterion. We start with an initial random model and alternate the optimization on the input and output variables.

For a given output distribution with fixed J and N_j, we optimize the discretization of the input variable to determine the values of I, N_i and N_{ij}. Then, for this input discretization, we optimize the discretization of the output variable to determine new values of J, N_j and N_{ij}. The process is iterated until convergence, which usually takes between two and three steps in practice. The univariate discretization optimizations are performed using the MODL discretization algorithm. This process is repeated several times, starting from different random initial solutions. The best solution is returned by the algorithm. The evaluation criterion $c_{reg}(M)$ given in formula (4) is related to the probability that a regression discretization model M explains the output variable. We then propose to use it to build a relevance criterion for the input variables in a regression problem. The input variables can be sorted by decreasing probability of explaining the output variable. In order to provide a normalized indicator, we consider the following transformation of c_{reg}:

$$g\left(M\right) = 1 - \frac{c_{reg}\left(M\right)}{c_{reg}\left(M_{\emptyset}\right)},$$

where $M_{\emptyset}$ is the null model with only one interval for the input and output variables. This can be interpreted as a compression gain, as negative log of probabilities are no other than coding lengths [Sha48]. The compression gain $g(M)$ hold its values between 0 and 1, since the null model is always considered in our optimization algorithm. It has value 0 for the null model and is maximal when the best possible explanation of the output ranks conditionally to the input ranks is achieved.
Our method is non parametric both in the statistical and algorithmic sense : any statistical hypothesis needs to be done on the data distribution (like Gaussianity for instance) and, as the criterion is regularized, there is no parameter to tune before minimizing it. This strong point enables to consider large datasets.

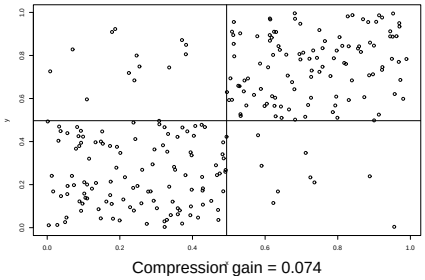
3 Experimental evaluation

In this section we first present the performance of the MODL 2D-discretization method on artificial datasets. Then, we apply it to rank the input variables of the Housing dataset from U.C. Irvine repository [DNM98] and show the interest of such a ranking criterion to improve regression tree performance.

3.1 Synthetic data experiments

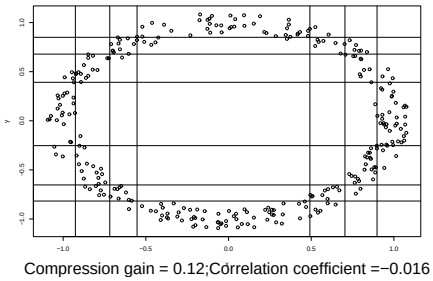
We first test our method on a *noise pattern* dataset of size 100 where the input and output variables are independent and uniformely distributed on $[0; 1]$. As expected, the absence of relevant information in X to predict Y produces a 1 by 1 partition, i.e., a null compression gain.

Secondly, we test the ability of the MODL 2D-discretization to partition a noisy XOR pattern. Our dataset contains one hundred instances uniformely distributed in the square $[0; 0.5] \times [0; 0.5]$ and one other hundred in the square



	I1	I2	Total
O2	13	112	125
O1	116	9	125
Total	129	121	250

Fig. 4. Correlation diagram with optimal MODL grid for noisy XOR dataset.



	I1	I2	I3	I4	I5	I6	I7	Total
S7	0	0	0	41	2	0	0	43
S6	0	0	12	5	7	6	0	30
S5	3	15	1	0	0	10	3	32
S4	21	1	0	0	0	1	36	59
S3	3	14	0	0	0	22	3	42
S2	0	6	9	0	10	2	0	27
S1	0	0	1	60	6	0	0	67
Total	27	36	23	106	25	41	42	300

Fig. 5. Correlation diagram with optimal MODL grid for noisy circle synthetic data.

$[0.5; 1] \times [0.5; 1]$. Fifty instances have been added uniformely in the square $[0; 1] \times [0; 1]$. The optimal MODL partition with compression gain of 0.074 precisely detects the noisy XOR pattern as shown in Fig. 4. The associated contingency table gives the number of instances in each cellule of the partition. It enables to construct conditional density estimators as follows : from the first column we can say that if x is in $[0; 0.5]$, then y is in $0; 0.5]$ with probability $\frac{116}{116+13} = 0.90$. The last synthetic experiment shows how the proposed method detects the presence of relevant information when two variables are not linearly correlated. We generate for this purpose a *circle* data set: three hundred instances have been generated on the circle of radius 1 with an additional noise such that their module is uniformely distributed in [0.9 ; 1.1]. As the empirical correlation is equal to -0.0169, any method based on the search of linear dependence fails. In contrast, the MODL 2D-discretization method underlines the relation between the two variables since the obtained compression gain is not zero. The optimal grid clearly identifies interesting regions as shown in Fig. 5.

3.2 Housing data

In this section, we study the regression problem of the Housing MEDV variable which describes housing values in suburbs of Boston. The Housing dataset con-

tains 506 instances, 13 numeric variables (including output variable MEDV) and 1 binary-valued variable which are described in Table 1.

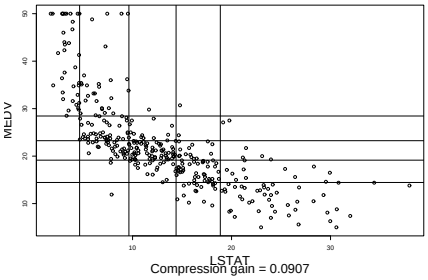
Table 1. Description of the 13 variables of the Housing dataset.

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

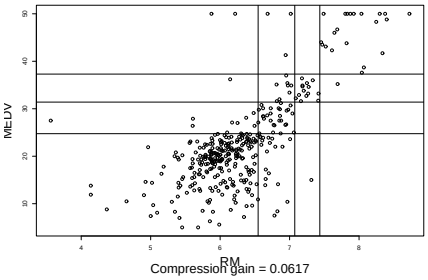
Table 2. Sorted compression gains and empirical correlation coefficients for the 12 numerical variables of the Housing regression dataset

Input variable	Compression gain	Correlation coefficient
LSTAT	0.092	-0.748
RM	0.0617	0.715
NOX	0.0444	-0.414
CRIM	0.0397	-0.377
INDUS	0.0395	-0.462
PTRATIO	0.0365	-0.523
AGE	0.0346	-0.384
DIS	0.0280	0.239
TAX	0.0252	-0.435
RAD	0.017	-0.36
B	0.0115	0.3
ZN	0.0109	0.358

We have split the Housing dataset in a 70% learning set and a 30% test set. Using the learning set, we have computed the optimal MODL 2D-discretizations for all of the twelve numeric variables. Considered as a relevance criterion, the associated compression gains are used to sort the variables according to the predictive information they contained w.r.t. the MEDV variable. To illustrate the interest of the MODL ranking criterion for variable selection, we then use this



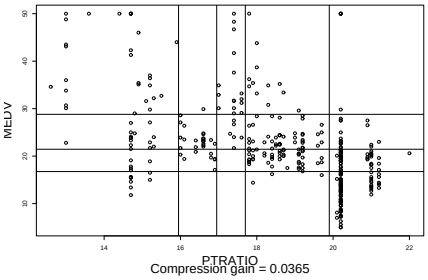
	I1	I2	I3	I4	I5	Total
O5	36	27	1	1	0	65
O4	0	39	12	1	2	54
O3	0	26	62	14	4	106
O2	0	1	17	38	11	67
O1	0	1	0	16	45	62
Total	36	94	92	70	62	354



	I1	I2	I3	I4	Total
O4	2	3	0	21	26
O3	1	8	14	1	24
O2	10	29	0	0	39
O1	244	20	1	0	265
Total	257	60	15	22	354

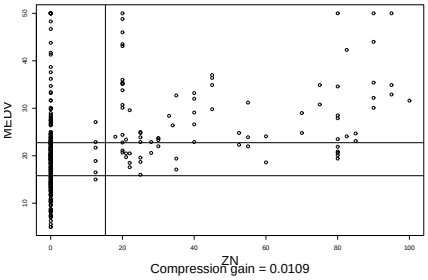
Fig. 6. Correlation diagram with optimal MODL grid for (LSTAT,MEDV) and (RM,MEDV) Housing variables

relevance criterion to improve CART regression trees [BFOS84]: we estimate such a tree with only the best MODL variable LSTAT, then with the two best variables and so on until the tree obtained with all the variables.
The sorted compression gains with the corresponding empirical correlation coefficients are shown in Table 2. For lack of space, the correlation diagram with the optimal MODL grids and the associated contingency table are shown in Fig 6 for the two better variables LSTAT and RM and in Fig 8 for the worst variable ZN. The PTRATIO variable seems also an interesting variable as its empirical correlation coefficient ranks it at the third position whereas the MODL criterion places it at the sixth (cf Fig 7). All these examples show the capacity of our MODL 2D-discretization algorithm to deal with complex datasets : the optimal grids present coarse grain when there is no predictive information or when the data are too noisy and capture fine details as soon as there is enough instances.
We then estimate the twelve trees with the 70% learning set. The first is estimated using only the best LSTAT variable, the second with the two best variables LSTAT and RM and so on until the twelfth tree estimated with all the variables. The obtained trees are used to predict the MEDV variable for both the learning and the test set. The resulting root mean squared errors are plotted in Fig. 9 for learning and test datasets. For both sets, we notice that:
- considering all the variables to estimate the regression tree is less efficient than considering only the more relevant ones according to the MODL criterion.



	I1	I2	I3	I4	I5	Total
O4	29	1	16	12	5	63
O3	15	21	6	40	16	98
O2	5	8	0	45	45	103
O1	8	0	0	3	79	90
Total	57	30	22	100	145	354

Fig. 7. Correlation diagram with optimal MODL grid for (PTRATIO,MEDV) Housing variables



	I1	I2	Total
O3	71	63	134
O2	119	22	141
O1	79	0	79
Total	269	85	354

Fig. 8. Correlation diagram with optimal MODL grid for (ZN,MEDV) Housing variable

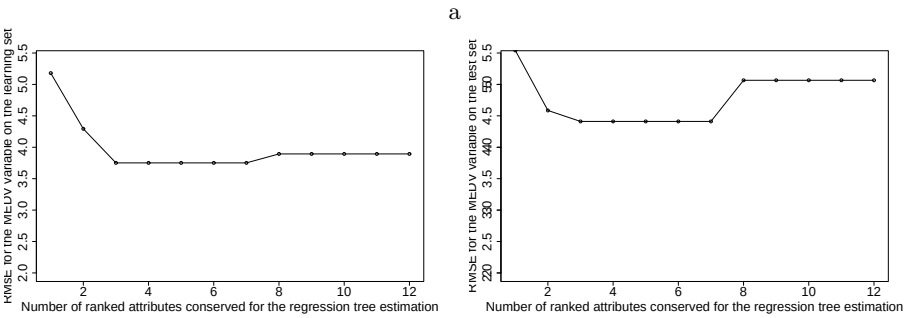


Fig. 9. Root Mean Squared Errors for the MEDV variable predicted from the regression trees estimated on a 70% learning dataset using an increasing number of ranked MODL variables (a) on the learning set (b) on the test set.

- the optimal tree is obtained with the three best variables and degrades after the incorporation of the eighth.

We can then conclude that, for this dataset, choosing the third tree during training step conducts to the best choice for the test set.

4 Conclusion and future work

The MODL 2D-discretization method proposed in this paper is a Bayesian model selection method for discretization grid models. The exact MODL criterion obtained enables to find the most probable discretization-based explanation of the data. Using a heuristic iterative algorithm which alternatively performs the discretization of the input and of the output variable, the obtained partitions accurately show linear and non linear relation and their compression gain can be used as a relevance criterion for the input variable ranking problem. It seems a very promising method to efficiently detect the relevant variables in large datasets during the data preparation step of regression problems.
In a future work, we plan to pursue the validation of our approach on larger numerous datasets and to use it to build a multivariate naive Bayes regressor exploiting the MODL discretized grids.

References

[BFOS84] L. Breiman, J. H. Friedman, A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[Bou06] M. Boullé. Modl :a Bayes optimal discretization method for continuous attributes. *Machine Learning*, accepted for publication 2006.

[Cat91] J. Catlett. On changing continuous variables into ordered discrete variables. In *In Proceedings of the European Working Session on Learning*, pages 87–102, 1991.

[CCK+00] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm 1.0: step-by-step data mining guide. *Applied Statistics Algorithms*, 2000.

[DKS95] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *In Proceedings of the 12th International Conference on Machine Learning*, pages 194–202, 1995.

[DNM98] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998.

[Fis36] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7, 1936.

[GE03] I. Guyon and André Elissef. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003.

[Hol93] R.C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 1993.

[LG99] Ph. Leray and P. Gallinari. Feature selection with neural networks. *Behaviormetrika*, 1999.

[LHTD02] H. Liu, F. Hussain, C.L. Tan, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 2002.

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 1978.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 1978.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell systems technical journal*, 1948.

[VL00] P.M.B. Vitanyi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. Inform. Theory*, 2000.