Problems of Machine Learning

Alexei Ya. Chervonenkis

Institute of Control Sciences, Moscow, Russia
chervnks@ipu.ru

The problem of reconstructing dependencies from empirical data became very important in a very large range of applications. Procedures used to solve this problem are known as "Methods of Machine Learning" [1,3]. These procedures include methods of regression reconstruction, inverse problems of mathematical physics and statistics, machine learning in pattern recognition (for visual and abstract patterns represented by sets of features) and many others. Many web network control problems also belong to this field. The task is to reconstruct the dependency between input and output data as precisely as possible using empirical data obtained from experiments or statistical observations.

Input data are composed of descriptions (curves, pictures, graphs, texts, messages) of input objects (we denote an input by x) and may be presented by vectors in Euclidian space or vectors of discrete values. In the latter case they may be sets of discrete features or even textual descriptions. An output value y may be given by a real value, vector or a discrete value. In the case of pattern recognition problem, output values may be names of classes (patterns), to which the input object belongs.

A training set is given by a sequence of pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$. One needs to find a dependency $y = F(x)$ such that forecast output values $y^* = F(x)$ for new input objects are most close to actual output values y, corresponding to the inputs x. Several schemes of training sequence generation are possible. From the theoretical point of view, it is most convenient to consider that the pairs are generated independently by some constant (but unknown) probability distribution $P(x, y)$, and the same distribution is used to generate new pairs. However, in practice the assumption of independency fails. Sometimes the distribution changes in time. In this case adaptive schemes of learning should be used, where the reconstructed function also changes in time. In some tasks there is no assumption about existing of any probability distribution on the set of pairs. Then the solution is to construct a function that properly approximates real dependency over its domain. If dependency between input and output variables is linear (or the best linear approximation is looked for), then well known Least Square Method is used to estimate the dependency coefficients. Still, if the training set is small (not large enough) in comparison with the number of arguments, then LSM does not work or works inefficiently. In this case some kinds of regularization are used. If dependency is sought in the class of polynomials of finite degree, then the problem may be reduced to the previous one by adding degrees of initial arguments. In the case of many arguments it is necessary to

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 21–23, 2011.

include degree products of initial arguments. In this case, of course, the number of unknown coefficients grows rapidly and the problem becomes intractable. Besides algebraic polynomials, trigonometric ones can be used, or, in general, expansions over preselected system of basis functions. If the dependency is sought in the form of piece-wise linear or piece-wise continuous function, then the standard least square method cannot be applied and other tools should be used, such as artificial neural networks. If the number of basis function necessary for proper approximation is too large, then kernel technique may be applied, which allows one to estimate the function value at a given point or at a set of given points. The so-called inverse problems of mathematical physics and statistics are also of this type.

Machine learning in pattern recognition is a particular case of dependency reconstruction. Here the output value is the name of a class. This class of problems covers a very large range of applications from image recognition to recognition of certain type of DNA sites, message classification, or recognition of unauthorized access attempts. In all cases of dependency reconstruction, one chooses a priori a certain class of models and then selects from it a model that in a sense is the best for describing dependency between input and output.

Three general questions arise in relation to any learning approach:

1. Is there a good model in the class that we have chosen?
2. May we hope that if the model behaves well on the training set, will it also behave well on new data?
3. Is there an efficient algorithm for selecting a proper model from a given class of models?

The smaller is the set of models, the closer is the point delivering minimum to empirical risk to the point delivering minimum to the true risk (within the class). On the other hand, the chance to find a good model within a small class is less than that for a large class. A solution may be as follows: Consider a set of expanding classes and choose the optimal size of a class, depending on the size of the training sample. For example, it is possible to increase (or decrease) the degree of approximating polynomial, or increase the number of terms in trigonometric expansion (or expansion over any other system of basic functions). In the case of linear function it is possible to define an order on the arguments and then sequentially increase the number of input variables in this order. One also can look over all combinations of arguments, increasing their number. In the case of piece-wise linear approximation one may change the number of pieces or number of neurons of the artificial neural network.

How one can find the best size of the class? The simplest way is to reserve a part of data set given for learning (validation set), using the rest part for finding the best model within expanding classes and then testing the result using the validation set, and at last selecting the model with the best score on the validation test. However, usually training data are lacking and it is costly to reserve some part of it. In this case one can use a means of control that does not need reservation, e.g. cross-validation. Another way is to estimate analytically

the difference between empirical and real risk values depending on training data and the size of a class from which we select the best model. The reason is that the empirical risk always deceases (not increases) with the class size, while the true risk passing its minimum starts increasing with the class size. Analytically it is possible to determine a corresponding correction. Other methods for choosing optimal complexity of a decision are possible. Note that the model complexity does not always directly correlates with the size of a class of models. For example, some very complicated procedures of feature preprocessing may be proposed, and the best decision rule is sought for a narrow class described in terms of secondary features resulting from preprocessing.

Let us consider in more detail some principles of choosing optimal model complexity within analytical approach. It is well known that maximum likelihood principle gives good results if the number of model parameters is rather low in comparison with the size of training data. On the other hand, Bayesian approach [2] gives the optimal result in the case where the number of parameters is very large or even infinite, but it requires a priori distribution over the set of their possible values. This approach results in a form of regularization, and it gives optimal degree of the selected model complexity. But how one can find this a priori distribution? Bayesian approach does not give any answer to this question. However, in many cases the a priori distribution depends on few parameters. Hence, it is promising to hybridize maximum likelihood and Bayesian approaches [4]: constructing likelihood function for the parameters of the a priori distribution, finding their optimal value and using them in Bayesian method. In some cases this approach gives very efficient algorithms.

Another idea is to use analytical estimates of the uniform closeness of empirical risk to the true risk (in absolute or relative form) and to use them as the estimates of the difference between empirical and real risk values. Then it is possible to use the estimates for choosing the best model complexity as it is mentioned above.

References

1. Vapnik, V.N., Chervonenkis, A.Y.: Theory of Pattern Recognition (in Russian). Moscow, Nauka (1974)
2. MacKay, J.D.C.: Bayesian interpolation. Neural Computation 4(3), 415–447 (1992)
3. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (2000)
4. Chervonenkis, A.Y.: A combined Bayes - Maximum Likelihood method for regression. In: Riccia, G.D., Lenz, H.-J., Kruse, R. (eds.) Data Fusion and Perception, Springer, Wien, New-York (2001)