Regression classifier in adaptive technical vision system

To cite this article: A Shmyrin et al 2019 J. Phys.: Conf. Ser. 1278 012001

View the article online for updates and enhancements.
Regression classifier in adaptive technical vision system

A Shmyrin$^1$, N Mishachev$^2$, V Kavygin$^3$ and A Kuznetsov$^4$

$^1$Department of Higher Mathematics, Lipetsk State Technical University, 30, Moskovskaya Str., Lipetsk 398600, Russia
$^2$Department of Applied Mathematics, Informatics and Mechanics, Voronezh State University, 1 University sq., Voronezh, 394018, Russia

$^1$E-mail: amsh@lipetsk.ru
$^2$E-mail: nmish@lipetsk.ru
$^3$E-mail: kabugun@mail.ru
$^4$E-mail: tema94lipetsk@yandex.ru

Abstract. The article deals with the problem of constructing a classifier for a set of $n$ physical objects encoded by vectors of the space $\mathbb{R}^N$. In this well-known problem of pattern recognition, we are interested in the case when the number of factors $N$ is much larger than the number of training vectors. As a classifier we use a linear homogeneous function $f: \mathbb{R}^N \to \mathbb{R}^1$. It is assumed that for each object with the number $i$ there is a cluster set $\tilde{K}_i$ consisting of $K_i$ training vectors corresponding to the observations of this object, for example, a set of its digitized photographs. In the case when $a$ linear homogeneous classifier, as a rule, exists. We describe the algorithm for constructing the classifier and discuss the properties of this algorithm.

1. Introduction

A huge number of publications have been devoted to the classical task of pattern recognition. Currently, the most popular are neural network classifier, the construction of which requires a large number of training samples [1, 2]. Note that the problem of identification for the $N$ coefficients of the linear regression equation can be considered as a learning problem for the simplest neural network of direct propagation with $N$ inputs, one output and no hidden layers. In this regression problem, it is usually assumed that the number of training samples is significantly larger than the number of identifiable coefficients $N$, and therefore the linear system for finding coefficients is overdetermined. The required coefficients can be found as a pseudo-solution of the system, that is, as a solution minimizing the square error (the least squares solution). In [3] and also in earlier works on the adaptive system of technical vision, a simple algorithm for pattern recognition was proposed, designed for a situation where the number of factors $N$ characterizing recognizable objects (patterns), is much bigger than the number of training samples $K$. In this algorithm, a linear homogeneous function was used as a classifier for a set of $n$ objects, while the system for finding its coefficients, that is, finding the coefficients of the linear homogeneous regression equation, was strongly underdetermined because of the condition $K \ll N$. As a method of solving the system, the Kaczmarz algorithm (of sequential projection) [4, 5] was used. This, in addition to the good rate of convergence, makes it easy to correct the results while adding new data (see [6, 7]). At the first glance, this type of approach to the problem of pattern recognition looks rather naive and lightweight, but experiments with the program that
implements the algorithm have proved its effectiveness in the described class of situations. In addition, the problem of optimizing the algorithm turned out to be meaningful and non-trivial.

2. Description of the algorithm

The algorithm is very simple and informally described in the introduction, but for a further discussion (in the next section) of some details and possible modifications, we give a formal description.

We shall assume that the vectors in $\mathbb{R}^n$ are, for definiteness, rows. Let the recognizable objects be encoded by vectors in the space $\mathbb{R}^N$ and let the number of classes of objects is $n$, while the total amount $K$ of training samples $\hat{K} \subset \mathbb{R}^N$ is equal to $K_1 + \cdots + K_n$, where $K_i$ is the number of samples in the corresponding class of objects (i.e. in the cluster) $\hat{K} \subset \mathbb{R}^N, i = 1, \ldots, n, \hat{K} = \hat{K}_1 \cup \cdots \cup \hat{K}_n$. Next, let

$$X_j^i = [x^i_{j1}, \ldots, x^i_{jN}] \in \mathbb{R}^N$$

be the vector of coordinates of the $j$-th sample in $K_i$, i.e. $i = 1, \ldots, n$ and $j = 1, \ldots, K_i$. The system of linear equations for finding the coefficients $[c_1, \ldots, c_N] \in \mathbb{R}^N$ of the linear homogeneous classifier

$$f_c: \mathbb{R}^N \to \mathbb{R}, f_c(X) = C \cdot X^T = c_1 x_1 + \cdots + c_N x_N,$$

has the form

$$\begin{cases}
X^1 \cdot C^T = B_1^T \\
\cdots \quad \cdots \\
X^n \cdot C^T = B_n^T
\end{cases} \quad (1)$$

where

$$X^i = \begin{bmatrix}
x^i_{11} & \cdots & x^i_{1N} \\
\vdots & \ddots & \vdots \\
x^i_{k_i1} & \cdots & x^i_{k_iN}
\end{bmatrix}$$

and $B_i = [b^i, \ldots, b^i] = b^i[1, \ldots, 1] \in \mathbb{R}^{K_i}$. This system consists of $K = K_1 + \cdots + K_n$ equations

$$X_j^i \cdot C^T = b^i; i = 1, \ldots, n, j = 1, \ldots, K_i.$$

Vector $\hat{B} = [b^1, \ldots, b^n] \in \mathbb{R}^N$ with different coordinates $b^i$ that are subsequently identifiers of clusters (or objects) is chosen arbitrarily; usually $\hat{B} = [1, \ldots, n]$. In real applications, the efficiency of the algorithm is related to the condition $K \ll N$, but for further theoretical considerations it suffices to assume that $K \leq N$. Under this condition, the set of all classifiers with given values of $b^i$ on each of the clusters of $K_i$ is, in the general case, an $(N - K)$-dimensional affine subspace of $\mathbb{R}^N$. To simplify the notation, we write the system (1) in the form

$$A \cdot C^T = B,$$  \quad (2)

where $A_X = [X^1, \ldots, X^n]^T$ and $B = [B_1, \ldots, B_n]^T$, then the coefficients of the classifier $f_c$ with minimal norm $\|C\|$ will be given by the formula $C_{min}^T = A_X^{-T} \cdot B$, where $A_X^{-T}$ is the pseudoinverse matrix for $A_X$.

The direct application of this formula, with the calculation of the pseudo-inverse matrix, requires approximately $O(K^2N^2)$ operations. Instead of the pseudoinversion it is convenient to use the iterative
Kaczmarz algorithm [4], which requires approximately $O(NK)$ operations. This algorithm has the following simple geometric meaning. Projection of the vector $x_0 \in \mathbb{R}^N$ onto the hyperplane $P$ given by the equation $a \cdot x^T = b$ is defined by the formula

$$x_1 = x_0 + \frac{(b - a \cdot x_0^T)}{||a||} \frac{a}{||a||}$$

(descent from the point $x_0$ along the unit normal $a/||a||$ to the hyperplane $P$ by the distance $(b - a \cdot x_0^T)/||a||$ from $x_0$ to $P$). Iterating this formula for an arbitrary initial vector $x_0 = c_0 = [c_0^1, ..., c_0^K]$ and the cyclic sequence of equations of system (1), we obtain a sequence of vectors convergent to some solution of the system (1). For convergence to a solution with minimal norm it suffices to take $c_0 = [0, ..., 0]$ as the initial vector.

3. Discussion of the algorithm and its possible modifications

Let us recall that we assume the condition $K \leq N$.

3.1. The existence of a classifier

In general, for most sets of raw data $\hat{K} \subset \mathbb{R}^N$, an undetermined (or defined) system (1) is compatible, and therefore a linear classifier exists. This trivial condition for the compatibility of the system (the Kronecker-Capelli theorem) in this case can be supplemented by more meaningful geometric observations associated with the structure of the vector $\hat{B}$ on the right-hand side of the system. A linear homogeneous classifier $f_c$ must take a constant value $b^i$ on vectors from the cluster $\hat{K}_i$. Since $b^{i_1} \neq b^{i_2}$ for $i_1 \neq i_2$, then the necessary condition for existence classifier is the condition that there are no pairwise intersections of the affine subspaces $A(\hat{K}_i)$ generated by the clusters $\hat{K}_i$. Note that the intersections for the linear shells of $\hat{K}_i$ are not forbidden. The simplest examples with $K_1 = \cdots = K_n = 1$ and points $\hat{K}_1, ..., \hat{K}_n$, located on one affine line in an order different from $1, ..., n$ show that the condition for the absence of intersections is not sufficient if the vector $\hat{B}$ is fixed. Question: is this condition sufficient for the existence of the classifier, if we also allow the choice of the vector of identifiers $\hat{B}$? This question is most relevant for $K$ close to $N$, when the relative measure of the set of systems that are $\varepsilon$-close to inconsistent, can be quite large. In our practical tasks, as a rule, $N \gg K$ and therefore the relative measure of systems that are $\varepsilon$-close to inconsistent, is small. But even in this case the question is quite interesting, since it can be reformulated as a question of the dependence of the relative measure of the set of «bad» systems on the choice of the vector $\hat{B}$.

3.2. Condition of the minimum of the classifier norm

This condition ensures the best separation of clusters for the fixed vector $\hat{B}$, since for $\hat{B} = [b^1, ..., b^K]$ the distance between adjacent hyperplanes

$$P_i: f_c(X) = b^i \text{ and } P_{i+1}: f_c(X) = b^{i+1},$$

containing clusters $\hat{K}_i$ and $\hat{K}_{i+1}$, is $|b^i - b^{i+1}|/||c||$. Thus, the choice of the vector $c_0 = [0, ..., 0]$ as the initial point for the Kaczmarz algorithm is optimal.

3.3. Choice of the sequence of equations

It is well known that the choice of a sequence of equations in the Kaczmarz algorithm affects the rate of convergence (see, for example, the discussion in [8-10]). The cyclic order is not optimal in the general case. For consistent systems, in the absence of additional information, good results are obtained by generating a random sequence of equations [8]. In our case, in the presence of data
clusters, it makes sense to use double randomization of a sequence of equations: the next step, a cluster is randomly chosen and then the equation within the cluster; one can also exclude the selected cluster from the set of clusters selected in the next step [7].

3.4. Choice of the vector of identifiers
\[ \hat{B} = [b^1, ... , b^n]. \]
We can assume that the condition \( b^1 < ... < b^n \) always holds, and the order of enumeration of clusters in system (1) can be changed arbitrarily. It is easy to show, already on simple examples, that the fixed and non-permutable order of enumeration of clusters, generally speaking, can be very far from optimal. If, for example, the clusters are approximated well enough by the points of some affine line \( L \subset \mathbb{R}^N \), then the order of their succession relative to \( L \) is the optimal order of enumeration of clusters, and this order may not coincide with the order given by default. Further, in the same example, after determining the optimal sequence of clusters the optimal choice of the numbers \( b^1 < ... < b^n \) should take into account the average distances \( d^i \) between neighboring clusters, namely, the vector \( \hat{B} \) must be proportional to the vector \( D = [d^1, ... , d^n] \). In particular, the choice of the vector \( \hat{B} \) with the same distances between successive coordinates (as in the vector \( [1,...,n] \)) can be very poor. Of course, this example cannot be called typical, but it shows quite well the nontriviality of the problem of choosing the vector of identifiers \( \hat{B} \).

4. Conclusion
The proposed simple method of pattern recognition based on the linear classifier, as our experience of practical application shows, works quite successfully with the corresponding class of problems in which \( K \ll N \). To further improve the algorithm, we need to solve the problem of optimization of the classifier on the set of vector-identifiers \( \hat{B} \in \mathbb{R}^N \). This problem is mathematically meaningful and nontrivial.

Acknowledgments
The work is supported by the Russian Foundation for Basic Research and the Lipetsk region within the framework of the scientific project 19-48-480007 p_a.

References
[1] Ossowski S 2002 Sieci neuronowe do przetwarzania informacji [In Polish] (Warsaw: Of. Ed. Pol.)
[2] Rumelhart D E, Hinton G E, Williams R J, 1986 Learning Internal Representations by Error Propagation Parallel Distributed Processin (Cambridge, MA, MIT Press) I, pp 318–362.
[3] Kozlov A M, Shatskikh I I, Kavygin V V (2017) The intellectual skill system for automatically choice routing technology machining parts [In Russian] Vestnik of MSTU STANKIN 3(42) 90-93
[4] Kaczmarz S 1993 Approximate solution of systems of linear equations Internat. J. Control 57 1269–1271
[5] Byrne C 1997 Convergent Block-Iterative Algorithms for Image Reconstruction from Inconsistent Data IEEE Transactions on Image Processing IP-6 pp 1296–1304
[6] Shmyrin A M and Mishachev N M 2016 Neighborhood systems and Kaczmarz algorithm [In Russian] Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhнические nauki 21 pp. 2113-2120 DOI: 10.20310/1810-0198-2016-21-6-2113-2120
[7] Shmyrin A M, Mishachev N M, Trofimov E P 2018 Priority identification of neighborhood systems and randomized Kaczmarz algorithm Int. Russian
Automation Conf. (RusAutoCon) (Sochi Russia) pp 1-4 DOI: 10.1109/RUSAUTOCON.2018.85017722018

[8] Strohmer T and Vershynin R 2009 A randomized Kaczmarz algorithm with exponential convergence Journal of Fourier Analysis and Applications 15(1) 262–278

[9] Strohmer T and Vershynin R 2009 Comments on the Randomized Kaczmarz Method Journal of Fourier Analysis and Applications 15(4) 437–440

[10] Censor Y, Herman G, Jiang M A 2009 Note on the Behavior of the Randomized Kaczmarz Algorithm of Strohmer and Vershynin Journal of Fourier Analysis and Applications 15(4) 431–436