Теорія Ймовір. та Матем. Статист.
Вип. 78, 2008

Theor. Probability and Math. Statist.
No. 78, 2009, Pages 23–35
S 0094-9000(09)00759-5
Article electronically published on August 4, 2009

A BAYESIAN CLASSIFIER

UDC 519.21

B. A. ZALESSKY AND P. V. LUKASHEVICH

Abstract. We consider a new Bayesian classifier for the classification of multidimensional observations $X_1, \dots, X_n$ of $\mathbb{R}^k$ if the learning sample is known. We assume that the data are generated by two disjoint bounded sets $\Omega_0, \Omega_1 \subset \mathbb{R}^k$ and each vector X_i of the sample is a result of the observation after one of the sets Ω_ℓ, $\ell = 0, 1$, with a random error. In other words, we assume that a priori the Bayesian probability μ is given on the set $\Omega = \Omega_0 \cup \Omega_1$ and that every vector of observations X_i has the density

$$g_\ell(x) = q_\ell \int_{\Omega_\ell} f(x, y)\, \mu(dy), \qquad \ell = 0, 1,$$

where the function $f(x, y)$ is a probability density for all $y \in \Omega$ and $q_\ell^{-1} = \mu(\Omega_\ell)$.

The maximum a posteriori probability estimators $\widehat{\Omega}_{\ell,n}$, $\ell = 0, 1$, for the sets Ω_ℓ, $\ell = 0, 1$, are constructed with the help of the learning sample. Under natural assumptions imposed on Ω_0 and Ω_1, we show that the estimators converge to some sets (possibly different from Ω_0 and Ω_1). If the mean frequencies π_ℓ of observations of the classes Ω_ℓ are equal to $\mu(\Omega_\ell)$, $\ell = 0, 1$, then the estimators are consistent in the sense that $\widehat{\Omega}_{\ell,n} \overset{n\to\infty}{\longrightarrow} \Omega_\ell$, $\ell = 0, 1$. We also discuss some results of numerical experiments showing the applicability of our classifier for solving the problems of the statistical classification.

1. Introduction

The Bayesian classification is one of the well-developed directions in the modern statistical analysis of data. This technique is applied not only in the mathematical statistics but also in other applied areas and still attracts the attention of many researchers. The methods of the Bayesian classification are used in applied science for analyzing data and images and for pattern recognition [1].

Different topics of the theory as well as methods of the statistical Bayesian estimation are presented in the papers [2, 8, 9, 10, 15]. Applications of those methods are described in [1, 16]. Among other modern methods of the classification we should like to mention the Vapnik methods of ordered minimization of risk and support vector machine [5, 6], random trees [3, 4], and neural networks [7].

A new Bayesian classifier of multidimensional data is presented in this paper. This classifier uses the information about geometrical forms and sizes of the sets generating the observations. In order to provide a better understanding of our results, below we compare this classifier with the classical Fisher classifier used in linear discriminant analysis.

The classical Fisher model assumes that every observation is a result of an additive distortion of one of two fixed points that are the classes generating the random observations. To construct the linear discriminant decision rule, one uses a learning sample,

2000 *Mathematics Subject Classification.* Primary 62C10; Secondary 90Bxx.
The first author was supported by the INTAS grant 04-77-7036.

evaluates the estimators $\overline{x}_0$ and $\overline{x}_1$ of these *classes* (actually, *points*) as the arithmetical means of the observations belonging to the corresponding classes, and then constructs the hyperplane separating these points [9]. It turns out that the estimators $\overline{x}_0$ and $\overline{x}_1$ are the maximum a posteriori estimators of the joint Gaussian density of the observations $p(x_0, x_1|X_1, \dots, X_n)$.

We consider the case where the classes Ω_ℓ, $\ell = 0, 1$, generating the observations of the sample are disjoint subsets of the k-dimensional cube $\Omega = [0, 1]^k$ (this assumption is imposed for the simplicity of our statements), partitioning the cube in such a way that $\Omega_0 \cap \Omega_1 = \varnothing$ and $\Omega_0 \cup \Omega_1 = \Omega$. We also assume that every random vector of the sample is a result of the observation of the *classes* (*subsets*, in fact) with a random error determined by an integral Bayesian density. The decision rule is constructed as the maximum a posteriori joint density estimator $p(\Omega_0, \Omega_1|X_1, \dots, X_n)$ over all admissible configurations of the subsets Ω_0 and Ω_1.

Similar methods are quite often used in various applications [12, 14] for the estimation (not for the classification however) of qualitative data, images or composite structures, say.

The results of numerical experiments described below show that the proposed approach can be applied even for the classification of the data generated by the nonconvex sets Ω_0 and Ω_1, in which case the linear discriminant decision rule may have a significant error. The difficulties arising when evaluating multidimensional integrals in our method can be avoided if one considers the regular discrete lattice $[0, \frac{1}{N}, \dots, \frac{N-1}{N}, 1]^k$ instead of the unit cube $[0, 1]^k$. However, modern software such as Mathematica or Matlab allows one to perform the necessary calculations even in the continuous case.

2. Description of the Bayesian classifier

In this section, we describe the Bayesian model underlying the proposed classifier. When describing the model we do not treat the most general setting; we rather choose simple and intuitively clear assumptions and restrictions.

Following the tradition, we denote random vectors and other random objects by uppercase letters, while their values are denoted by lowercase letters. For example, the equality $U = u$ means that a random vector U assumes a value u. Throughout in the paper we assume that all random vectors and measures are defined with respect to the Borel σ-algebra $\mathcal{B}_{\mathbb{R}^k}$ and that all the sets appearing in the text belong to this σ-algebra.

We consider the classification problem with a learning sample for observations of independent random vectors X_i of $\mathbb{R}^k$, $k \geq 1$. It is assumed that the observations belong to one of the two classes denoted by the numbers 0 and 1.

Every observation X_i belonging to the class $\ell = 0, 1$ has a Bayesian density. To define the notion of a Bayesian density we need some notation.

Let Ω_0 and Ω_1 be two disjoint subsets of the unit cube $[0, 1]^k$, $\Omega_0 \cup \Omega_1 = [0, 1]^k$, generating observations of each class and let μ be the a priori probability measure on Ω.

Remark 2.1. One can use other sets instead of $[0, 1]^k$. The unit cube is chosen just for the sake of simplicity of statements.

The Bayesian conditional density of an observation X_i given that it is a member of a class ℓ is defined by

$$g_\ell(x) = q_\ell \int_{\Omega_\ell} f(x, y)\, \mu(dy), \qquad \ell = 0, 1,$$

where $f(x, y)$, as a function of x, is a probability density for all $y \in [0, 1]^k$ and where $q_\ell^{-1} = \mu(\Omega_\ell)$.

Example 2.1. In some applications, it is reasonable to consider the Gaussian density with parameters (y, T) as $f(x, y)$, that is,

$$f(x,y) = (2\pi)^{-k/2} \det(T)^{-1/2} \exp\left\{-\frac{1}{2}\left\langle T^{-1}(x-y), x-y\right\rangle\right\},$$

and the Lebesgue measure on $[0,1]^k$ as the a priori probability μ. This in particular means that subsets of Ω_ℓ that are of equal volumes generate observations with equal probabilities.

It is further assumed that a learning sample

$$(x_1, \xi_1), \dots, (x_n, \xi_n), \qquad x_i \in \mathbb{R}^k, \quad \xi_i \in \{0,1\},$$

is known. Each member (x_i, ξ_i) of the learning sample is a pair constituted by the value x_i of the random vector X_i accompanied with the result of teacher classification ξ_i. The numbers $\xi_i \in \{0,1\}$ are understood as values of independent 0–1 random variables Ξ_i.

The conditional density $g_\ell(x)$ can be written with the help of the Bayes formula as follows:

$$g_\xi(x) = p(x|\Omega_\xi, \xi) = \frac{p(\Omega_\xi|x,\xi)\, p(x|\xi)}{p(\Omega_\xi|\xi)}, \tag{1}$$

where the sets Ω_ℓ, $l = 0, 1$, are random and such that $p(\Omega_0|0) = 1 - p(\Omega_1|0)$. The assumption that the sets Ω_ℓ are random helps to describe the class of admissible sets in the classification problem. The density $p(\Omega_\xi|\xi)$ is assumed to be nonzero for admissible sets Ω_ℓ and zero for the other sets. We can translate the case of nonrandom classes (that is, if the admissible Ω_ℓ are equally favorable) to the above language by saying that Ω_ξ are uniformly distributed in the set of all admissible configurations.

It follows from (1) that the prior density is given by

$$p(\Omega_\xi|x,\xi) = \frac{p(x|\Omega_\xi,\xi)\, p(\Omega_\xi|\xi)}{p(x|\xi)}.$$

Let $\mathbf{x} = (x_1, \dots, x_n)$ be a vector of observations of the random vectors $X_1, \dots, X_n$ (recall that $x_i \in \mathbb{R}^k$), while $\xi = (\xi_1, \dots, \xi_n)$ are the results of teacher classification. Then the joint posterior density of the sample can be written as follows:

$$p(\Omega_\xi|\mathbf{x},\xi) = \frac{p(\mathbf{x}|\Omega_\xi,\xi)\, p(\Omega_\xi|\xi)}{p(\mathbf{x}|\xi)} = \prod_{j=1}^{n} \frac{p\left(x_j|\Omega_{\xi_j},\xi_j\right) p\left(\Omega_{\xi_j}|\xi_j\right)}{p(x_j|\xi_j)}$$

under the condition that the corresponding triples $(X_j, \Omega_{\Xi_j}, \Xi_j)$ are independent.

Since $p(x_j|\xi_j)$ does not depend on Ω_ξ and since $\Omega_1 = [0,1]^k \setminus \Omega_0$, one can define the maximum likelihood function $L(\Omega_0)$ by

$$L_{\mathbf{x}}(\Omega_0) = \frac{1}{n}\sum_{j=1}^{n} \ln\left(p\left(\Omega_{\xi_j}|\xi_j\right)\mu^{-1}\left(\Omega_{\xi_j}\right)\int_{\Omega_{\xi_j}} f(x_j,y)\,\mu(dy)\right).$$

The maximum a posteriori density estimator for Ω_0 is given by

$$\widehat{\Omega}_{0,n} = \operatorname*{argmax}_{\Omega_0} L_{\mathbf{x}}(\Omega_0).$$

The estimator is well defined if there exists an element $\widehat{\Omega}_{0,n}$ such that

$$L_{\mathbf{x}}(\widehat{\Omega}_{0,n}) = \max.$$

The natural restrictions usually imposed on the form of the sets Ω_ℓ in the practical classification problems guarantee that the latter condition holds. The general results can be stated for the case of a compact set $\Omega_0 \subset [0,1]^k$ if the likelihood function $L_{\mathbf{x}}(\Omega_0)$

is an upper semicontinuous map acting from the metric space of compact sets $\{\Omega_0 \subset [0,1]^k | \overline{\Omega_0} = \Omega_0\}$ equipped with the Hausdorff metric to the real line [11].

When solving concrete classification problems one usually deals with the sets of a simple topological and geometrical structure whose boundaries are also "simple". Below we determine the sets Ω_0 with the help of a finite-dimensional vector parameter $\theta \in [a,b]^m$. We also assume that the function $L_{\mathbf{x}}(\theta)$ is continuous with respect to the arguments $(\mathbf{x},\theta)$ everywhere except (possibly) for a set of zero Lebesgue measure.

The properties of the classifier $\widehat{\Omega}_{0,n}$ are described in the next section.

3. Properties of the classifier

As mentioned above, we state our results in the form convenient for practical applications of the classifier. Without loss of generality, we assume that the first n_0 observations are generated by the set Ω_0, while the other $n_1 = n - n_0$ observations are generated by the set Ω_1. Then the likelihood function is given by

$$\begin{aligned} L_{\mathbf{x}}(\Omega_0) = {} & \frac{1}{n}\sum_{j=1}^{n_0} \ln\left(p\left(\Omega_0|0\right) \mu\left(\Omega_0\right)^{-1} \int_{\Omega_0} f(x_j,y)\,\mu(dy)\right) \\ & + \frac{1}{n}\sum_{j=1}^{n_1} \ln\left(p\left(\Omega_1|1\right) \mu\left(\Omega_1\right)^{-1} \int_{\Omega_1} f(x_{j+n_0},y)\,\mu(dy)\right). \end{aligned} \tag{2}$$

If

$$\mathsf{E} \ln\left(\int_{\Omega_\ell} f(X_j,y)\,\mu(dy)\right) < \infty, \qquad \ell = 0,1, \tag{3}$$

then the strong law of large numbers holds; namely,

$$L_{\mathbf{x}}(\Omega_0) \overset{n\to\infty}{\longrightarrow} \mathsf{E}\, L_{\mathbf{X}}(\Omega_0) \quad \text{almost surely},$$

where the expectation is equal to

$$\begin{aligned} \mathsf{E}\, L_{\mathbf{X}}(\Omega_0) = {} & p_0\, \mathsf{E} \ln\left(p\left(\Omega_0|0\right) \mu\left(\Omega_0\right)^{-1} \int_{\Omega_0} f(X_1,y)\,\mu(dy)\right) \\ & + p_1\, \mathsf{E} \ln\left(p\left(\Omega_1|1\right) \mu\left(\Omega_1\right)^{-1} \int_{\Omega_1} f(X_{1+n_0},y)\,\mu(dy)\right) \end{aligned}$$

for $p_0 = 1 - p_1 = \mathsf{P}(\xi_1 = 0)$. Condition (3) holds for a wide class of distributions appearing in mathematical statistics (including, for example, uniform, Gaussian, and other distributions of exponential type [10]). The above limit relation together with some extra conditions imposed on the sets Ω_0 allows one to conclude that the estimator $\widehat{\Omega}_{0,n}$ is asymptotically stable as $n \to \infty$.

In what follows we assume that the sets Ω_0 permit a continuous parametrization such that

a) $\theta \in [-a,a]^d$, $d \geq 1$;
b) $\Omega_0(\theta') \neq \Omega_0(\theta'')$ for $\theta' \neq \theta''$;
c) $\mu\left(\Omega_0(\theta_m) \Delta \Omega_0(\theta)\right) \to 0$ as $\theta_m \to \theta$.

Certainly this is not the most general condition imposed on the parametrization of the sets $\Omega_0(\theta)$. Nevertheless this case is useful for solving problems of the multidimensional classification where consideration is restricted to the polynomial surfaces of low orders separating Ω_0 and Ω_1. The surfaces often are of the third, second, or even first order or they are described with a finite family of functions.

Remark 3.1. The restrictions on the form of the sets Ω_0 are dictated by the nature of the problem itself. For example, it is natural to assume that the multivariate data for patients form the classes with smooth boundaries. Another reason for introducing these restrictions is the dimension of the data to be classified. If the dimension k of the cube $[0,1]^k$ is large, then the parametric surfaces depend on a large number of parameters. For example, a surface of the second order separating the classes in $\mathbb{R}^k$ is determined by $(k^2+k+2)/2$ parameters. Therefore a large number of observations are needed to statistically justify the estimator of Ω_0.

The parametrization of the sets Ω_ℓ allows one to obtain the conditional densities

$$p(\theta|\ell) = p\left(\Omega_\ell|\ell\right), \qquad \ell = 0, 1.$$

For example, if θ determines the boundary between the sets, then it is natural to assume that the random variables θ and ξ are independent. The conditional densities transform to the unconditional densities $p(\theta|\ell) = p(\theta)$ in this case. The latter assumption is not used in the further reasoning. However, it is important in some classification problems appearing in practice. Let $0 < \gamma < 1$ be a real number and let $\ell = 0, 1$. Put

$$\Upsilon = \{\theta|\gamma \le \mu\left(\Omega_0(\theta)\right) \le 1-\gamma\}, \qquad I_\ell(\theta, X_j) = \int_{\Omega_\ell(\theta)} f(X_j, y)\, \mu(dy),$$

$$L_{\mathbf{x}}(\theta) = L_{\mathbf{x}}(\Omega_0(\theta)), \qquad \theta_n^* = \operatorname*{argmax}_{\theta\in\Upsilon} L_{\mathbf{x}}(\theta), \qquad \theta^* = \operatorname*{argmax}_{\theta\in\Upsilon} \mathsf{E}\, L_{\mathbf{X}}(\theta).$$

The parametrization of the sets Ω_0 reduces the problem of the evaluation of the maximum a posteriori density to the calculation of θ_n^*.

Remark 3.2. Let θ_{true} be a value of the parameter corresponding to the true partition. Then, for all Borel functions h,

$$\mathsf{E}\, h(X_j) = \mu(\Omega_\ell(\theta_{\text{true}}))^{-1} \int h(x) I_\ell(\theta_{\text{true}}, x)\, dx,$$

where ℓ is the number of the class containing X_j, since random vectors X_j are generated by the sets $\Omega_\ell(\theta_{\text{true}})$, $\ell = 0, 1$.

To prove the convergence of $\widehat{\Omega}_{0,n}$ to $\widehat{\Omega}_0$, we need the following results.

Theorem 3.1. *Assume that the regularity conditions a), b), and c) hold for the parametrization of the sets $\Omega(\theta)$ (see p. 26). We also assume that*

d) *random vectors X_j have density such that*

$$\sup_{x,y\in[0,1]^k} |f(x,y)| \le r \quad \textit{and} \quad f(x,y) \ge c_1 \exp\left(-|x|^b\right)$$

for some $r, c_1, b > 0$;

e) *the moment*

$$\mathsf{E}\, |X_j|^b = c_2 < \infty, \qquad \ell = 0, 1,$$

is finite;

f) *the conditional densities $p\left(\Omega_\ell(\theta)|\ell\right)$, $\ell = 0, 1$, are continuous with respect to θ and $p\left(\Omega_\ell(\theta)|\ell\right) > 0$ for $\theta \in \Upsilon$.*

Then the trajectories of the likelihood function $L_{\mathbf{x}}(\theta)$ are stochastically equicontinuous on Υ for all real $0 < \gamma < 1$ in the sense that

$$\lim_{h\to 0} \sup_n \mathsf{P}\left(\sup_{\theta_1,\theta_2\in\Upsilon,\, \|\theta_1-\theta_2\|\le h} |L_{\mathbf{X}}(\theta_1) - L_{\mathbf{X}}(\theta_2)| > \varepsilon \right) = 0 \tag{4}$$

for all $\varepsilon > 0$.

Proof. Since the parametrization of the set $\Omega_0(\theta)$ is stochastically continuous (see property c) on p. 26), the set Υ is compact. This together with the continuity of $L_{\mathbf{x}}(\theta)$ in θ for all $\mathbf{x}$ implies that the expression

$$\sup_{\theta_1,\theta_2\in\Upsilon,\,\|\theta_1-\theta_2\|\le h} |L_{\mathbf{X}}(\theta_1)-L_{\mathbf{X}}(\theta_2)|$$

is a random vector. The compactness of the set Υ gives immediately the uniform convergence $\mu\left(\Omega_0(\theta_m)\Delta\Omega_0(\theta)\right)\to 0$ as $\theta_m\to\theta$ and $p\left(\Omega_\ell(\theta_m)|\ell\right)\to p\left(\Omega_\ell(\theta)|\ell\right)$. Therefore, for any $\delta>0$, one can choose $h>0$ such that

$$\mu\left(\Omega_\ell(\theta_1)\Delta\Omega_\ell(\theta_2)\right)<\delta \quad\text{and}\quad \left|p\left(\Omega_\ell(\theta_1)|\ell\right)-p\left(\Omega_\ell(\theta_2)|\ell\right)\right|<\delta, \qquad \ell=0,1,$$

for all $\|\theta_1-\theta_2\|<h$. Then

$$\begin{aligned}&\left|\ln\left(p\left(\Omega_0(\theta_1)|0\right)\mu\left(\Omega_0(\theta_1)\right)^{-1}I_0(\theta_1,X_j)\right)-\ln\left(p\left(\Omega_0(\theta_2)|0\right)\mu\left(\Omega_0(\theta_2)\right)^{-1}I_0(\theta_2,X_j)\right)\right|\\ &\qquad\le\left|\ln\left(\frac{p\left(\Omega_0(\theta_1)|0\right)}{p\left(\Omega_0(\theta_2)|0\right)}\right)\right|+\left|\ln\left(\frac{\mu\left(\Omega_0(\theta_2)\right)}{\mu\left(\Omega_0(\theta_1)\right)}\right)\right|+\left|\ln\left(\frac{I_0(\theta_1,X_j)}{I_0(\theta_2,X_j)}\right)\right|\end{aligned}$$

for $\|\theta_1-\theta_2\|<h$.

Since $|\ln(1+s)|\le|s|/(1-|s|)$ for all $-1<s<1$,

$$\left|\ln\left(\frac{\mu\left(\Omega_0(\theta_2)\right)}{\mu\left(\Omega_0(\theta_1)\right)}\right)\right|\le\left|\ln\left(1+\frac{\mu\left(\Omega_0(\theta_2)\right)-\mu\left(\Omega_0(\theta_1)\right)}{\mu\left(\Omega_0(\theta_1)\right)}\right)\right|\le\frac{\delta}{\gamma-\delta} \tag{5}$$

for $\theta_1,\theta_2\in\Upsilon$ and sufficiently small $\delta>0$. Similarly one can prove the following inequality:

$$\left|\ln\left(\frac{p\left(\Omega_0(\theta_1)|0\right)}{p\left(\Omega_0(\theta_2)|0\right)}\right)\right|\le\frac{\delta}{c_3-\delta}, \tag{6}$$

where the inequality $c_3=\inf_{\theta\in\Upsilon}p\left(\Omega_\ell(\theta)|\ell\right)>0$ follows from the compactness of the set Υ and assumption f) of the theorem. Inequalities (5) and (6) remain true if Ω_1 is substituted for Ω_0. Thus

$$\begin{aligned}&\mathsf{P}\left(\sup_{\theta_1,\theta_2\in\Upsilon,\,\|\theta_1-\theta_2\|\le h}|L_{\mathbf{X}}(\theta_1)-L_{\mathbf{X}}(\theta_2)|>\varepsilon\right)\\ &\qquad\le\mathsf{P}\left(\sup_{\substack{\theta_1,\theta_2\in\Upsilon\\ \|\theta_1-\theta_2\|\le h}}\left|\frac{1}{n}\sum_{j=1}^{n_0}\ln\left(\frac{I_0(\theta_1,X_j)}{I_0(\theta_2,X_j)}\right)+\frac{1}{n}\sum_{j=n_0+1}^{n}\ln\left(\frac{I_1(\theta_1,X_j)}{I_1(\theta_2,X_j)}\right)\right|>\frac{\varepsilon}{2}\right)\end{aligned} \tag{7}$$

for sufficiently small δ.

Similarly to (5), we use assumption d) of the theorem and prove that

$$\left|\ln\left(\frac{I_\ell(\theta_1,X_j)}{I_\ell(\theta_2,X_j)}\right)\right|\le\frac{\delta re^{|X_j|^b}}{c_1\gamma-\delta re^{|X_j|^b}}\ . \tag{8}$$

Then, for an arbitrary $R>0$,

$$\begin{aligned}&\mathsf{E}\sup_{\theta_1,\theta_2\in\Upsilon,\,\|\theta_1-\theta_2\|\le h}\left|\ln\left(\frac{I_\ell(\theta_1,X_j)}{I_\ell(\theta_2,X_j)}\right)\right|\\ &\qquad=\mathsf{E}\,1_{(|X_j|\le R)}\sup_{\theta_1,\theta_2\in\Upsilon,\,\|\theta_1-\theta_2\|\le h}\left|\ln\left(\frac{I_\ell(\theta_1,X_j)}{I_\ell(\theta_2,X_j)}\right)\right|\\ &\qquad\quad+\mathsf{E}\,1_{(|X_j|>R)}\sup_{\theta_1,\theta_2\in\Upsilon,\,\|\theta_1-\theta_2\|\le h}\left|\ln\left(\frac{I_\ell(\theta_1,X_j)}{I_\ell(\theta_2,X_j)}\right)\right|,\end{aligned} \tag{9}$$

where $1_{(\cdot)}$ is the indicator function.

Estimate (8) implies that the first term on the right hand side of (9) does not exceed

$$\delta r e^{R^b} \left(c_1\gamma - \delta r e^{R^b} \right)^{-1}.$$

We use the inequalities

$$\begin{aligned} &\mathsf{E}\, 1_{(|X_j|>R)} \sup_{\theta\in\Upsilon} |\ln\left(I_\ell(\theta, X_j)\right)| \\ &\quad \le \mathsf{E}\left(1_{|X_j|>R}\right) \sup_{\theta\in\Upsilon} \left(\left(1_{I_\ell(\theta,X_j)<1} + 1_{\ I_\ell(\theta,X_j)\ge 1}\right) |\ln\left(I_\ell(\theta, X_j)\right)| \right) \\ &\quad \le \mathsf{E}\, 1_{(|X_j|>R)} \left(\left(|\ln(c_1\gamma)| + |X_j|^b \right) + \ln(r) \right) \\ &\quad \le \mathsf{E}\, 1_{(|X_j|>R)} \left(c_4(c_1,\gamma,r) + |X_j|^b \right) \end{aligned} \tag{10}$$

to estimate the second term on the right hand side of (9). The latter sequence of inequalities follows from the property d). Now (9) and (10) imply that

$$\begin{aligned} &\mathsf{E} \sup_{\theta_1,\theta_2\in\Upsilon,\, \|\theta_1-\theta_2\|\le h} \left| \ln\left(\frac{I_\ell(\theta_1, X_j)}{I_\ell(\theta_2, X_j)} \right) \right| \\ &\qquad \le \frac{\delta r e^{R^b}}{(c_1\gamma - \delta r e^{R^b})} + 2\ \mathsf{E}\, 1_{(|X_j|>R)} \left(c_4(c_1,\gamma,r) + |X_j|^b \right). \end{aligned} \tag{11}$$

We apply the Chebyshev inequality to the right hand side of (7) and use estimate (11) to complete the proof. We obtain

$$\begin{aligned} &\mathsf{P}\left(\sup_{\theta_1,\theta_2\in\Upsilon,\, \|\theta_1-\theta_2\|\le h} |L_{\mathbf{X}}(\theta_1) - L_{\mathbf{X}}(\theta_2)| > \varepsilon \right) \\ &\qquad \le 4\varepsilon^{-1} \left(\frac{\delta r e^{R^b}}{(c_1\gamma - \delta r e^{R^b})} + \mathsf{E}\, 1_{(|X_j|>R)} \left(c_4(c_1,\gamma,r) + |X_j|^b \right) \right). \end{aligned} \tag{12}$$

Setting $R = (\ln(1/\delta)/2)^{1/b}$ and recalling Remark 3.2, we make sure that the right hand side of the latter inequality approaches zero as $\delta \to 0$ with the rate that does not depend on n. □

In what follows we use the following result that follows from estimates (11) and (12), since Υ is a compact set.

Proposition 3.1. *Let the assumptions of Theorem* 3.1 *hold. Then the expectation*

$$\mathsf{E}\left| \ln\left(\frac{I_\ell(\theta_1, X_j)}{I_\ell(\theta_2, X_j)} \right) \right|$$

is uniformly continuous in $\Upsilon \times \Upsilon$ *as a function of* θ_1 *and* θ_2 *on the set* $\theta_1 = \theta_2$ *in the sense that*

$$\lim_{h\to 0} \sup_{\theta_1,\theta_2\in\Upsilon,\, \|\theta_1-\theta_2\|\le h} \mathsf{E}\left| \ln\left(\frac{I_\ell(\theta_1, X_j)}{I_\ell(\theta_2, X_j)} \right) \right| = 0.$$

Proposition 3.1 and estimates (5) and (6) imply the following result.

Proposition 3.2. *Let a random variable* X' *have the density* $\mu(\Omega_0(\theta_{\text{true}}))^{-1} I_0(\theta_{\text{true}}, \mathbf{x})$ *and let a random variable* X'' *have the density* $\mu(\Omega_1(\theta_{\text{true}}))^{-1} I_1(\theta_{\text{true}}, \mathbf{x})$. *Let the assumptions of Theorem* 3.1 *hold. Then the expectation*

$$\begin{aligned} \mathsf{E}\, L_{\mathbf{x}}(\theta) = p_0\, \mathsf{E} \ln\Big(p\left(\Omega_0(\theta)|0\right) \mu\left(\Omega_0\right)^{-1} I_0(\theta, X') \Big) \\ + p_1\, \mathsf{E} \ln\Big(p\left(\Omega_1(\theta)|1\right) \mu\left(\Omega_1\right)^{-1} I_1(\theta, X'') \Big) \end{aligned}$$

is uniformly continuous in θ *on* Υ.

The supremum $\sup_\theta \xi(\theta,\omega)$ of trajectories of a stochastic process is not necessarily a random variable in the general case; nevertheless $\sup_{\theta\in\Upsilon} L_{\mathbf{x}}(\theta)$ is a random variable.

Let

$$\Theta_n^*(\mathbf{x}) = \left\{\theta\in\Upsilon \middle| L_{\mathbf{x}}(\theta) = \max_{\theta\in\Upsilon} L_{\mathbf{x}}(\theta)\right\}, \qquad \Theta^{**} = \left\{\theta\in\Upsilon \middle| \mathsf{E}\, L_{\mathbf{X}}(\theta) = \max_{\theta\in\Upsilon} \mathsf{E}\, L_{\mathbf{X}}(\theta)\right\}.$$

Proposition 3.3. *Let the assumptions of Theorem* 3.1 *hold. Then* $\sup_{\theta\in\Upsilon} L_{\mathbf{x}}(\theta)$ *is a random vector. Moreover the supremum is attained in the sense that* $\theta_n^* = \operatorname{argmax}_{\theta\in\Upsilon} L_{\mathbf{x}}(\theta)$ *exists. The supremum* $\sup_{\theta\in\Upsilon} \mathsf{E}\, L_{\mathbf{X}}(\theta)$ *is attained, too. The sets* $\Theta_n^*(\mathbf{x}), \Theta^{**}$ *are compact.*

Proposition 3.3 follows from the continuity of $L_{\mathbf{x}}(\theta)$ with respect to θ for all $\mathbf{x}$ and from the compactness of the set Υ.

The above conditions do not imply that the sets $\Theta_n^*(\mathbf{x})$ and Θ^{**} are unique as well as they do not imply that the sets are finite.

Example 3.1. Consider the case of two-dimensional observations. Then

$$\Omega_0(\theta)\cup\Omega_1(\theta) = [0,1]^2.$$

Assume that the admissible partitions are such that one of the sets is a square $\Omega_0(\theta)\subset[0,1]^2$ centered at $v\in[0,1]^2$ and whose side is of length s. Then we parameterize the sets with the parameter $\theta=(v,s)$. It is clear that this parametrization satisfies conditions a), b), and c) on p. 26. As the Bayesian measure μ we take the Lebesgue measure which is invariant with respect to shifts of the sets. The a priori density of partitions into the classes is taken such that the density $p\,(\Omega_0(v,s)|0) > 0$ is constant for all admissible v and s if s is fixed, that is, $p\,(\Omega_0(v_1,s)|0) = p\,(\Omega_0(v_2,s)|0)$ and $p\,(\Omega_0((1/2,1/2),s)|0)\le\varepsilon$. Assume that the sample is generated by the true partition for which $\theta_{\text{true}} = ((1/2,1/2),s_0)$. If ε is sufficiently small, then the set of maxima of Θ^{**} does not contain any point of the form $((1/2,1/2),s)$. Therefore the maximums are not attained in the nonsymmetric squares (v^*,s^*), $v^*\ne(1/2,1/2)$. On the other hand, since the set $\Omega_\ell((1/2,1/2),s_0)$ is symmetric, there are several squares of this kind.

A similar example shows that the set $\Theta_n^*(\mathbf{x})$ may contain more than one element in the general case. Although the above example is of theoretical value only, it clearly shows that the classification problems in practice require additional conditions to guarantee that the sets $\Theta_n^*(\mathbf{x})$ and Θ^{**} are singletons. An example is given at the end of this section, where the maximum a posteriori probability estimator is not unique; moreover it is not consistent.

The following result on the uniform closeness of trajectories of the likelihood function to its mathematical expectation is a common tool for studying the empirical extremums [5].

Theorem 3.2. *Let the assumptions of Theorem* 3.1 *hold. Then the trajectories of the likelihood function* $L_{\mathbf{x}}(\theta)$ *converge to* $\mathsf{E}\, L_{\mathbf{X}}(\theta)$ *in probability in the uniform metric, that is,*

$$\mathsf{P}\left(\sup_{\theta\in\Upsilon} |L_{\mathbf{x}}(\theta) - \mathsf{E}\, L_{\mathbf{X}}(\theta)| > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{13}$$

Proof. The bounds constructed in the proof of Theorem 3.1 and Proposition 3.1 imply that the trajectories $L_{\mathbf{x}}(\theta)$ are stochastically equicontinuous with respect to θ and also that $\mathsf{E}\, L_{\mathbf{X}}(\theta)$ is uniformly continuous in the compact set Υ, whence we conclude that $\sup_{\theta\in\Upsilon} |L_{\mathbf{x}}(\theta) - \mathsf{E}\, L_{\mathbf{X}}(\theta)|$ is measurable and that, for all $\varepsilon,\delta>0$, there exists a finite set $\Upsilon(\varepsilon)$ such that

$$\mathsf{P}\left(\sup_{\theta\in\Upsilon}\left|L_{\mathbf{x}}(\theta) - L_{\mathbf{x}}(\tilde\theta(\theta))\right| > \frac{\varepsilon}{3}\right) < \frac{\delta}{3}, \qquad \sup_{\theta\in\Upsilon}\mathsf{E}\left|L_{\mathbf{X}}(\theta) - L_{\mathbf{X}}(\tilde\theta(\theta))\right| < \frac{\delta}{3}$$

for all n and for a point $\tilde{\theta}(\theta) \in \Upsilon(\varepsilon)$ closest to θ. Then

$$\begin{aligned}
\mathsf{P}\left(\sup_{\theta\in\Upsilon}|L_{\mathbf{x}}(\theta)-\mathsf{E}\,L_{\mathbf{X}}(\theta)|>\varepsilon\right) &\le \mathsf{P}\left(\sup_{\theta\in\Upsilon}\left|L_{\mathbf{x}}(\tilde{\theta}(\theta))-\mathsf{E}\,L_{\mathbf{X}}(\tilde{\theta}(\theta))\right|>\frac{\varepsilon}{3}\right)+\frac{2\delta}{3}\\
&\le \mathsf{P}\left(\sup_{\theta\in\Upsilon(\varepsilon)}|L_{\mathbf{x}}(\theta)-\mathsf{E}\,L_{\mathbf{X}}(\theta)|>\frac{\varepsilon}{3}\right)+\frac{2\delta}{3} \qquad (14)\\
&\le \sum_{j=1}^{|\Upsilon(\varepsilon)|}\mathsf{P}\left(|L_{\mathbf{x}}(\theta_j)-\mathsf{E}\,L_{\mathbf{X}}(\theta_j)|>\frac{\varepsilon}{3}\right)+\frac{2\delta}{3}.
\end{aligned}$$

The law of large numbers implies that every term on the right hand side of (14) converges to zero. Thus one can choose $n(\varepsilon)$ such that the right hand side of (14) does not exceed δ for all $n > n(\varepsilon)$. □

Since Υ is compact, one can choose a version of the estimator $\theta_n^* = \theta_n^*(\mathbf{x})$ that is a random vector for all sets $\Theta_n^*(\mathbf{x})$ consisting of points of maximum of the maximum likelihood function $L_{\mathbf{x}}(\theta)$ (these sets, as we already know, are compact for all $\mathbf{x}$).

Proposition 3.4. *Let the assumptions of Theorem* 3.1 *hold. Then there exists a measurable version of the estimator* $\theta_n^* = \theta_n^*(\mathbf{x})$.

Proposition 3.4 follows from Pfanzagl's Theorem 3.10 in [13].

When solving classification problems in practice, one usually deals with the case where the set Υ of values of the parameter θ is finite. In such a case, one can use, for example, an integer parametrization of Υ and take

$$\theta_n^*(\mathbf{x}) = \min\Theta_n^*(\mathbf{x})$$

as a measurable version.

Theorem 3.2 implies the convergence in probability of maximum values of the likelihood function to the maximum of $\mathsf{E}\,L_{\mathbf{X}}(\theta)$; moreover this theorem implies the attraction of θ_n^* to the set Θ^{**} in probability in the sense that the Euclidean distance

$$\rho(\theta_n^*,\Theta^{**}) = \inf_{u\in\Theta^{**}}\|\theta_n^*-u\|$$

between θ_n^* and the set Θ^{**} approaches zero in probability as $n\to\infty$.

Proposition 3.5. *Let the assumptions of Theorem 1 hold. Then, for all* $\varepsilon > 0$,

$$\mathsf{P}\left(|L_{\mathbf{x}}(\theta_n^*)-\mathsf{E}\,L_{\mathbf{X}}(\theta^{**})|>\varepsilon\right)\overset{n\to\infty}{\longrightarrow}0, \qquad (15)$$

where $\theta^{**}\in\Theta^{**}$.

The estimators θ_n^* *are attracted to the set* Θ^{**} *in probability, that is,*

$$\mathsf{P}\left(\rho(\theta_n^*,\Theta^{**})>\varepsilon\right)\overset{n\to\infty}{\longrightarrow}0.$$

Proof. It is easy to see that

$$L_{\mathbf{x}}(\theta^{**})-\mathsf{E}\,L_{\mathbf{X}}(\theta^{**})\le L_{\mathbf{x}}(\theta_n^*)-\mathsf{E}\,L_{\mathbf{X}}(\theta^{**})\le L_{\mathbf{x}}(\theta_n^*)-\mathsf{E}\,L_{\mathbf{X}}(\theta_n^*).$$

Thus

$$|L_{\mathbf{x}}(\theta_n^*)-\mathsf{E}\,L_{\mathbf{X}}(\theta^{**})|\le\sup_{\theta\in\Upsilon}|L_{\mathbf{x}}(\theta)-\mathsf{E}\,L_{\mathbf{X}}(\theta)|.$$

Now we conclude that the limit (15) exists by Theorem 3.2.

We prove the second statement by contradiction. Assume that there are $\varepsilon_0,\delta>0$ and a subsequence of random vectors $\theta_{n_k}^* = \theta_{n_k}^*(\mathbf{X})$ such that

$$\mathsf{P}\left(\rho(\,\theta_{n_k}^*,\Theta^{**})>\varepsilon_0\right)\ge\delta.$$

If ε_1 is sufficiently small, then

$$\begin{aligned}&\mathsf{P}\left(\left|L_{\mathbf{x}}(\theta^*_{n_k}) - \mathsf{E}\, L_{\mathbf{X}}(\theta^{**})\right| > \varepsilon_1\right)\\ &\quad\geq \mathsf{P}\Big(\rho(\theta^*_{n_k}, \Theta^{**}) > \varepsilon_0,\\ &\qquad\left|L_{\mathbf{x}}(\theta^*_{n_k}) - \mathsf{E}\, L_{\mathbf{X}}(\theta^*_{n_k})\right| \leq \varepsilon_1, \left|\mathsf{E}\, L_{\mathbf{X}}(\theta^*_{n_k}) - \mathsf{E}\, L_{\mathbf{X}}(\theta^{**})\right| > 2\varepsilon_1\Big).\end{aligned} \tag{16}$$

Since Θ^{**} is a compact set and $\mathsf{E}\, L_{\mathbf{X}}(\theta)$ is continuous, we have

$$\left|\mathsf{E}\, L_{\mathbf{X}}(\theta^*_{n_k}) - \mathsf{E}\, L_{\mathbf{X}}(\theta^{**})\right| > 2\varepsilon_1,$$
$$\mathsf{P}\left(\left|L_{\mathbf{x}}(\theta^*_{n_k}) - \mathsf{E}\, L_{\mathbf{X}}(\theta^*_{n_k})\right| \leq \varepsilon_1\right) \overset{n\to\infty}{\longrightarrow} 1$$

for sufficiently small ε_1 and $\rho(\theta^*_{n_k}, \Theta^{**}) > \varepsilon_0$ by Theorem 3.2. Therefore the right hand side of inequality (16) approaches $\delta > 0$, which contradicts (15). □

Note that the classifier constructed above is inconsistent in the general case in the sense that $\Omega_0(\theta^*_n) \nrightarrow \Omega_0(\theta_{\text{true}})$. Nevertheless, $\rho(\,\theta^*_n, \Theta^{**}) \to 0$ in probability as $n \to \infty$. Note also that a classifier minimizing the number of wrongly classified observations should be of this kind in general. It is clear that the classifier should have a "bias" toward the class $\Omega_\ell(\theta_{\text{true}})$ containing a larger portion of observations of another class.

Now we consider a modified Bayesian classifier for the case where the a priori probabilities that the sets $\Omega_\ell(\theta_{\text{true}})$ occur coincide with their Bayesian measures; that is, $\mu(\Omega_\ell(\theta_{\text{true}})) = p_\ell$, and all admissible sets $\Omega_\ell(\theta)$ are uniformly distributed:

$$p\left(\Omega_\ell(\theta)|\ell\right) = \text{const}\,.$$

The modified Bayesian classifier uses sample estimators $\hat{p}_\ell = n_\ell/(n_0 + n_1)$ of the a priori probabilities as the normalizing factors for the densities $I_\ell(\theta, X_j)$. Then the modified likelihood function becomes of the following form:

$$\widetilde{L}_{\mathbf{x}}(\Omega_0) = \frac{1}{n}\sum_{j=1}^{n_0} \ln\left(\hat{p}_0^{-1} I_0(\theta, x_j)\right) + \frac{1}{n}\sum_{j=1}^{n_1} \ln\left(\hat{p}_1^{-1} I_1(\theta, x_j)\right),$$

and its expectation is equal to

$$\mathsf{E}\,\widetilde{L}_{\mathbf{x}}(\Omega_0(\theta)) = \mathsf{E}\,\hat{p}_0 \ln\left(\hat{p}_0^{-1} I_0(\theta, X_1)\right) + \mathsf{E}\,\hat{p}_1 \ln\left(\hat{p}_1^{-1} I_1(\theta, X_{n_0+1})\right).$$

Remark 3.3. We use additional a priori information in the definition of the modified classifier that the fraction of observations of every class in the sample is equal to the corresponding Bayesian measure of this class, that is,

$$\mu(\Omega_\ell(\theta_{\text{true}})) = p_\ell = \mathsf{E}\,\hat{p}_\ell.$$

It is not hard to prove the following result.

Proposition 3.6. *Let the assumptions of Theorem* 3.1 *hold. Then*

$$\mathsf{E}\,\widetilde{L}_{\mathbf{x}}(\Omega_0(\theta)) \leq \mathsf{E}\,\widetilde{L}_{\mathbf{x}}(\Omega_0(\theta_{\text{true}}))$$

for all $\theta \in \Upsilon$. *Therefore* θ_{true} *belongs to the set of maxima of the function* $\mathsf{E}\,\widetilde{L}_{\mathbf{x}}(\Omega_0(\theta))$.

Moreover

$$\mathsf{E}\,\widetilde{L}_{\mathbf{x}}(\Omega_0(\theta)) < \mathsf{E}\,\widetilde{L}_{\mathbf{x}}(\Omega_0(\theta_{\text{true}}))$$

for all $\theta \in \Upsilon$ *such that*

$$\text{mes}_{Leb}\big(\{x|I_0(\theta, x) \neq I_0(\theta_{\text{true}}, x)\}\big) > 0.$$

Proof. Since $\ln(1+s) = s + R(s)$ for $-1 < s$, where $R(s) \le 0$, we have

$$\begin{aligned}
&\mathsf{E}\,\widetilde{L}_{\mathbf{X}}(\Omega_0(\theta)) - \mathsf{E}\,\widetilde{L}_{\mathbf{X}}(\Omega_0(\theta_{\text{true}})) \\
&\quad\le \int \ln\left(\frac{I_0(\theta,x)}{I_0(\theta_{\text{true}},x)}\right) I_0(\theta_{\text{true}},x)\,dx + \int \ln\left(\frac{I_1(\theta,x)}{I_1(\theta_{\text{true}},x)}\right) I_1(\theta_{\text{true}},x)\,dx \\
&\quad\le \int \left(I_0(\theta,x) - I_0(\theta_{\text{true}},x) + I_1(\theta,x) - I_1(\theta_{\text{true}},x)\right)\,dx \qquad (17)\\
&\qquad + \int R\left(\frac{I_0(\theta,x) - I_0(\theta_{\text{true}},x)}{I_0(\theta_{\text{true}},x)}\right) I_0(\theta_{\text{true}},x)\,dx \\
&\qquad + \int R\left(\frac{I_1(\theta,x) - I_1(\theta_{\text{true}},x)}{I_1(\theta_{\text{true}},x)}\right) I_1(\theta_{\text{true}},x)\,dx
\end{aligned}$$

for all $\theta \in \Upsilon$. The first integral on the right hand side is equal to 0; hence

$$\big(I_\ell(\theta,x) - I_\ell(\theta_{\text{true}},x)\big) I_\ell^{-1}(\theta_{\text{true}},x) > -1$$

under the assumptions of Theorem 3.1. The second inequality of Proposition 3.6 follows from (17), since the second and third integrals on the right hand side of (17) are negative if $\text{mes}_{Leb}\left(\{x | I_0(\theta,x) \ne I_0(\theta_{\text{true}},x)\}\right) > 0$. □

Let

$$\widetilde{\Theta}_n^*(\mathbf{x}) = \left\{\theta \in \Upsilon \middle| \widetilde{L}_{\mathbf{x}}(\theta) = \max_{\theta\in\Upsilon} \widetilde{L}_{\mathbf{x}}(\theta)\right\}, \qquad \widetilde{\Theta}^{**} = \left\{\theta \in \Upsilon \middle| \mathsf{E}\,\widetilde{L}_{\mathbf{X}}(\theta) = \max_{\theta\in\Upsilon} \mathsf{E}\,\widetilde{L}_{\mathbf{X}}(\theta)\right\}.$$

The following result can be proved in the same way as Proposition 3.5.

Proposition 3.7. *Suppose the assumptions of Theorem 1 hold. Then*

$$\mathsf{P}\left(\left|L_{\mathbf{x}}(\widetilde{\theta}_n^*) - \mathsf{E}\,L_{\mathbf{X}}(\widetilde{\theta}^{**})\right| > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0$$

for all $\varepsilon > 0$ and $\widetilde{\theta}^* \in \widetilde{\Theta}_n^*(\mathbf{x})$ and $\widetilde{\theta}^{**} \in \widetilde{\Theta}^{**}$ for all estimators. Moreover $\theta_{\text{true}} \in \widetilde{\Theta}^{**}$.

The estimators $\widetilde{\theta}^*$ are attracted to the sets $\widetilde{\Theta}^{**}$ in probability, so that

$$\mathsf{P}\left(\rho(\widetilde{\theta}_n^*, \widetilde{\Theta}^{**}) > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Therefore $\widetilde{\theta}_n^* \overset{n\to\infty}{\longrightarrow} \theta_{\text{true}}$ in probability if the set $\widetilde{\Theta}^{**}$ is a singleton.

In conclusion we exhibit the results of numerical experiments performed with the help of Matlab 7.0. The results of classification by using the Bayesian classifier studied above are shown in the left part of Figure 1. The right part of Figure 1 contains the results of classification by the Gaussian mixtures classifier (GMM classifier). The black areas in the right part of Figure 1 separate the classes.

The classification in the first two examples is done with the help of separating lines, while a circle and parabola $y = ax^2 + bx + c$ are used for the third and fourth examples, respectively. The unusual choice of a circle or another closed curve for the classification can in fact be used for solving many typical problems in practice. For example, if patients are classified according to two characteristics, the blood pressure and body temperature, then the data of healthy patients form a rectangle inside a rectangle of the data of all patients.

We plan to study a Bayesian classifier with a density having a finite support elsewhere. The methods for constructing such a classifier are different from that used in this paper.

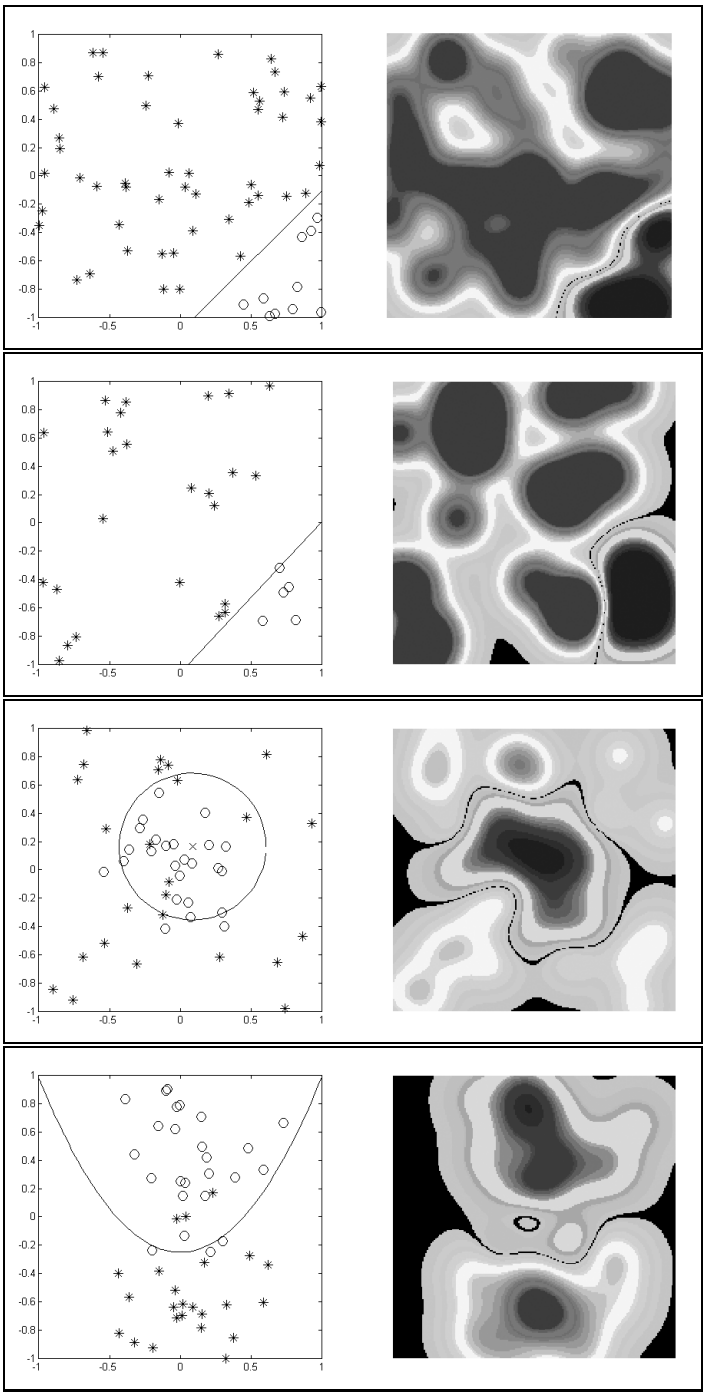

FIGURE 1. Examples of classification with the help of several admissible sets

Bibliography

1. S. A. Aivazyan, B. M. Buchshtaber, I. S. Enyukov, and L. D. Meshalkin, *Applied Statistics: Classification and Reducing of Dimension*, Finansy i Statistika, Moscow, 1989. (Russian)
2. A. A. Borovkov, *Mathematical Statistics*, Nauka, Moscow, 1984; English. transl., Taylor and Francis, Amsterdam, 1999. MR782295 (86i:62001)
3. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group, 1984. MR726392 (86b:62101)
4. L. Breiman, *Random Forests*, Technical report, Department of Statistics, University of California, Berkeley, CA, 1999.
5. V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979; English transl., Springer-Verlag, New York, 1982. MR672244 (84a:62043)
6. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998. MR1641250 (99h:62052)
7. S. Haykin, *Neural Networks: A Comprehensive Foundation*, Wiley, New York, 2005.
8. E. E. Zhuk and Yu. S. Kharin, *Stability in the Cluster Analysis of Multivariate Data*, Belgosuniversitet, Minsk, 1998. (Russian)
9. S. Zaks, *Theory of Statistical Inference*, John Wiley and Sons, New York, 1971. MR0420923 (54:8934a)
10. E. Lehmann, *Theory of Point Estimation*, Chapman and Hall, London, 1991. MR1143059 (93c:62003b)
11. G. Matheron, *Random Sets and Integral Geometry*, Wiley, New York, 1975. MR0385969 (52:6828)
12. V. V. Mottl' and I. B. Muchnik, *Hidden Markov Models in Structural Analysis of Signals*, Fizmatlit, Moscow, 1999. (Russian) MR1778152 (2001m:94014)
13. J. Pfanzagl, *On the measurability and consistency of minimum contrast estimates*, Metrika **14** (1969), 249–273.
14. D. Forsyth and J. Ponce, *Computer Vision. A Modern Approach*, Prentice Hall, New York, 2002.
15. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Elsevier Science and Technology Books, Amsterdam, 1990. MR1075415 (91i:68131)
16. M. I. Schlesinger and V. Hlavac, *Ten Lectures on Statistical and Structural Pattern Recognition*, Springer-Verlag, Berlin, 2002.

United Institute of Informatics Problems, National Academy of Sciences, Surganova Street 6, Minsk, 220012, Belarus'

E-mail address: zalesky@newman.bas-net.by

United Institute of Informatics Problems, National Academy of Sciences, Surganova Street 6, Minsk, 220012, Belarus'

Received 23/OCT/2006

Translated by N. SEMENOV