EFFICIENCY OF CLASSIFICATION METHODS
BASED ON EMPIRICAL RISK MINIMIZATION

V. I. Norkin\textsuperscript{a} and M. A. Keyzer\textsuperscript{b}

UDC 519:234:24:85

A binary classification problem is reduced to the minimization of convex regularized empirical risk functionals in a reproducing kernel Hilbert space. The solution is searched for in the form of a finite linear combination of kernel support functions (Vapnik’s support vector machines). Risk estimates for a misclassification as a function of the training sample size and other model parameters are obtained.

Keywords: machine learning, classification, recognition, empirical risk minimization, support vector machine (SVM), consistency, rate of convergence.

INTRODUCTION

The paper discusses the theoretical efficiency of some (binary) classification methods such as the support vector machine/method (SVM) [1]. The classification problem is considered for a supervised-learning model, which is standard for statistical learning theory. Let there be a training sample of pair observations \((y_i, x_i), i = 1, \ldots, m\) of size \(m\), where \(x_i\) is the feature vector of the object \(i\) with values in a set \(X\), \(y_i\) is the label of the class from a discrete set \(Y\) to which the object \(i\) belongs. The statistical learning theory assumes that the pairs \((y_i, x_i)\) are independent random vectors with common unknown probabilistic distribution \(P\) on the set \(Y \times X\). By a classification problem is meant constructing a mapping (classifier) from \(X\) into \(Y\) based on a training sample. A measure of classifier efficiency is the average misclassification probability as a function of the training sample size and other model parameters. This quantity is called average Bayesian risk (in a restricted sense), which has a theoretical minimum. For a rational classification method, the misclassification risk should tend to the theoretical minimum as the size of the training sample increases, which means convergence (in probability or almost surely) of the classification method. Such classification methods are called consistent; however, consistency can be observed only for certain classes of distributions of the training sample.

One of the problems in the statistical classification theory is that the theoretical distribution of elements of the training sample is not known; therefore, it is impossible to verify formally whether the distribution of the training sample belongs to one class or another. Classification methods consistent on any distribution of training data could solve this problem. Such methods can reasonably be called universally consistent [2]. Whether there are universally consistent classification methods have been unknown for a long time. As late as in 1977, it was shown [3] that the \(k\)-nearest neighbor algorithm known since 1951 possessed this property. However, it has been found out [2] that universally consistent methods may converge (reduce the misclassification risk as the size of training sample increases) arbitrarily poorly on some distributions of training data; hence, there is no universally best (optimal) classification method. Thus, the statements on the optimality of some classification method or concerning the estimates of the rate of the convergence of misclassification risk to an irreducible minimum are true only for a certain class of distributions of training data.

This conclusion is also true for empirical risk minimization methods such as support vector machine [1]. Its linear version (the method of optimal separating planes) is detailed in [4, 5], and the nonlinear (the method of potential functions) in [6], the newest versions (support vector machines, SVM) being described in [1, 7, 8]. Nowadays, SVM successfully compete with the most developed machine classification systems and keeps being a subject of intensive theoretical analysis.

\textsuperscript{a}V. M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, norkin@i.com.ua. \textsuperscript{b}Centre for World Food Studies, Vrije Universiteit, Amsterdam, the Netherlands, m.a.keyzer@sow.vu.nl. Translated from Kibernetika i Sistemnyi Analiz, No. 5, pp. 93–105, September–October 2009. Original article submitted December 2, 2008.
The classical substantiation of the method is based on the uniform functional law of large numbers, and the estimates of the rate of convergence depend on the so-called VC-capacity (Vapnik–Chervonenkis) of the class of decision functions [1, 4, 5]. However, estimating the VC-capacity is generally a challenge; moreover, the class of feasible functions is not always of finite VC-capacity. Though some often used (quadratic, of absolute deviation) empirical risk functionals being minimized reflect the quality of the classification rule, their relation to the probability of correct classification is not obvious. The form of the available estimates of the rate of convergence in terms of confidential bounds for the risk does not allow comparing this method to other ones for which these estimates are obtained in terms of the convergence of mean risk.

The present paper analyzes the SVM method to solve binary classification problems in the context of the theory of ill-posed problems and estimates the rate of convergence of the method under rather general assumptions on the distribution of training data. These assumptions imply that some characteristics of data distribution (conditional medians and means) belong to a certain functional Hilbert space (with a reproducing kernel). The paper specifies the relationship between the risk functionals being used and misclassification probabilities. The rate of convergence of misclassification probability to a minimum is estimated, the estimates being dependent on data distribution but independent of the VC-capacity of the functional space. The uniform functional law of large numbers is not used. The estimates include unknown constants; therefore, they are unsuitable for quantitative conclusions but show how the average error of this classifier tends to the theoretical minimum. As a rule, the rate of convergence is of order \( \frac{1}{\sqrt{m}} \), where \( m \) is the number of elements in the training sample.

The presentation is structured as follows. The first section discusses classification methods based on the approximation of the exact solution of the minimization problem for classification risk. The second section considers an alternate approach to classification and shows how the minimization of misclassification probability can be reduced to minimization of a convex functional of risk. The third section describes a regularization method to minimize convex empirical risk functionals, and the fourth section analyzes it for convergence as the number of training examples increases. The fifth section interprets these results for classification problems. In conclusion, basic features of the SVM are discussed in view of the results obtained in the paper.

### 1. BAYESIAN CLASSIFICATION METHODS

Let observation data be random pairs \((y, x)\) with the distribution \(P\). The scalar value \(y \in Y\) can take only discrete values (labels of classes), for example, \(y \in Y = \{0, 1\}\), and components of the \(n\)-dimensional vector \(x \in X\) (features) may be either discrete or continuous. A problem with \(s\) classes can be reduced in a standard way to a binary classification problem, where one class is one of the original classes and the second one is all the others. For any measurable function \(f(x): X \to \mathbb{R}^1\), the binary classification rule is defined by

\[
I_{1/2}(f(x)) = \begin{cases} 
1, & f(x) > 1/2, \\
0, & \text{otherwise.}
\end{cases}
\]  

(1)

The quality of the classification rule \(I_{1/2}(f(\cdot))\) is measured by the Bayesian risk, i.e., by the misclassification probability \(P\{I_{1/2}(f(x)) \neq y\}\), where \(y \in \{0, 1\}\). Recall [2] that the Bayesian risk attains the minimum value \(P^*\) on the decision rule specified by the conditional probability function \(p_1(x) = P\{y = 1|x\}\) yet unknown. In the case of many classes, where \(y \in Y = \{0, 1, 2, \ldots\}\), the optimal Bayesian classification strategy is to maximize (with respect to \(l \in \{0, 1, 2, \ldots\}\)) the conditional probability distribution \(p_1(x) = P\{y = l|x\}\) [10], which however is not known too.

Thus, a possible way of constructing optimal classifiers is to approximate the conditional probability function \(p_1(x) = P\{y = 1|x\}\) in the binary case or the distributions \(p_l(x) = P\{y = l|x\}\), \(l = 0, 1, \ldots\), in the general case. For example, in the \(k\)-nearest neighbor classification algorithm [1, Sec. 5], \(k\) observations \(\{x_i, i \in I_k(x)\}\) closest to the feature vector \(x\) are selected and classed, and the vector \(x\) is related to the class with maximum frequency. Denote such a classifier by \(g_k(x)\), its quality is measured by the misclassification probability \(L_k(m) = E_{\{(y_1, x_1), \ldots, (y_m, x_m)\}}P(g_k(x) \neq y)\), and the asymptotic quality by \(L_k^* = \lim_{m \to \infty} L_k(m)\). As is known from [1, Sec. 5], \(P^* \leq L_k^* \leq P^*(1 + 1/\sqrt{ke})\) for all distributions and even \(k\), where \(e\) is the basis of natural logarithms. Moreover, this classifier is universally consistent, i.e., \(L_k(m) \to P^*\) as \(m \to \infty\) and
Approximation of conditional probabilities designates expectation with respect to the measure \( p_x P_y x \) of a training sample, irrespective of the probabilistic distribution of sample units though \( p_x P_y x \) is the function of conditional mean (regression); therefore, to estimate it, standard approaches of regression analysis such as nonparametric methods can be applied [11]. Let \( \{ (y_i, x_i), i = 1, \ldots, m \} \) be a training sample, \( \rho(\cdot, \cdot) \) be a function of the distance between points in the feature space \( X \), \( k(\cdot) \) be a one-dimensional symmetric density of probabilities, and \( \theta_m \) be positive numbers. Then the Nadaraya–Watson kernel estimator [11, Sec. 5] of the regression function \( p_1(x) \) has the form

\[
\tilde{p}_1(x) = \sum_{i: y_i = 1} k \left( \frac{\rho(x, x_i)}{\theta_m} \right) / \sum_{i=1}^m k \left( \frac{\rho(x, x_i)}{\theta_m} \right),
\]

and the corresponding binary classifier is specified by (1) with \( f(x) = \tilde{p}_1(x) \).

In [12–14], the unknown conditional probability distribution \( \{ p_l(x), l = 0, 1, \ldots \} \) is approximated by the Bayesian estimate \( \{ \tilde{p}_l(x), l = 0, 1, \ldots \} \) under the (strong) assumption of conditional independence of features (components of a random vector \( x \) for objects from a fixed class \( l \)). In [12–14], estimates of the rate of convergence

\[
B(m) = E_{\{ (y_i, x_i), \ldots, (y_m, x_m) \}} \{ \arg \max_l p_l(x) \neq y \} \leq P^* + C / \sqrt{m}
\]

are obtained for such a classifier, where \( C \) is a universal constant independent of data distribution, and it is proved that they are unimprovable under the assumptions on the influence of the size of the training sample \( m \). The feature independence assumption, which is important for this estimate, is detailed in [10, Sec. 3.3].

As is known (see [2] and references therein), the Bayesian classification error can be expressed in the binary case in terms of the error \( \tilde{p}_1(x) - p_1(x) \) of the approximation of conditional probability \( p_1(x) = P \{ y = 1 | x \} \) as follows:

\[
P \{ I_{1/2} (\tilde{p}_1(x)) \neq y \} - P^* \leq 2E | \tilde{p}_1(x) - p_1(x) |, \tag{2}
\]

where the symbol \( E \) designates expectation with respect to the measure \( P \). This estimate gives a statistical substantiation for the classification methods based on the approximation of conditional probabilities \( p_l(x) = P \{ y = l | x \}, l = 0, 1 \).

2. BINARY CLASSIFICATION PROBLEM RELATED TO THE OPTIMIZATION OF CONVEX RISK FUNCTIONALS

Another approach to developing classification methods is to reduce a classification problem to a convex optimization problem for a risk functional [9, Sec. 4.2]. In what follows, we will consider the cases not presented in the review [9]. For example, as is known from [2], \( p_1(x) = \arg \min_f E (y - f(x))^2 \). If \( f(x) \) is an approximate solution of a quadratic risk minimization problem, the corresponding decision rule is defined by (1), and the classification quality estimate by (2). This approach to binary classification is discussed in detail in [15]. Moreover, the statistical theory of classification and learning uses risk functionals of the form

\[
R_\epsilon(f) = E \max \{ 0, | y - f(x) | - \epsilon \}, \ \epsilon \geq 0,
\]

such as \( R_0(f) = E | y - f(x) | = L_1(f) \) [1]. Their application is somewhat substantiated by the following estimate [2]:

\[
P \{ I_{1/2} (f(x)) \neq y \} - \min_f P \{ I_{1/2} (f(x)) \neq y \} \leq 2E | y - f(x) | - \min_f E | y - f(x) |, \tag{3}
\]

where the minimization is over the set of Borel functions on \( X \).

The theorem below estimates the quality of the classifier that minimizes the quadratic risk functional \( L_2(f) = E (y - f(x))^2 \), the estimate differing from (2).
THEOREM 1. Let $F$ be a set of Borel functions on $x \in X$ such that $p_1(x) = P \{ y = 1 | x \in F \}$. Then the following estimate holds for any function $f(\cdot) \in F$:

$$
P \{ I_{1/2} (f(x)) \neq y \} - \min_{f \text{-measurable}} P \{ I_{1/2} (f(x)) \neq y \} \leq 2\sqrt{L_2(f)} - \min_{f \in F} L_2(f).$$

(4)

**Proof.** Let us represent

$$
P \{ I_{1/2} (f(x)) \neq y \} = E_x \{ P \{ I_{1/2} (f(x)) \neq y | x \} \},$$

where $P \{ | x \}$ and $E \{ | x \}$ are the conditional probability and conditional expectation for a fixed component $x$ of the random vector $(y, x)$; $E_x$ is expectation with respect to the distribution of the component $x$. Let us consider the functions $p_1(x) = P \{ y = 1 | x \}$, $p_0(x) = P \{ y = 0 | x \} = 1 - p_1(x)$, and $e(h, x) = E \{ (h - y)^2 | x \}$.

The following relationships are true:

$$
\begin{align*}
    r(h, x) &= P \{ I_{1/2} (h) \neq y | x \} = \begin{cases} 
        p_0(x) = 1 - p_1(x), & h > 1/2, \\
        p_1(x), & h \leq 1/2.
    \end{cases} \\
    e(h, x) &= p_1(x)(h - 1)^2 + (1 - p_1(x))h^2 = h^2 - 2hp_1(x) + p_1(x) \\
    &= (h - p_1(x))^2 + p_1(x)(1 - p_1(x)) = (h - p_1(x))^2 + e^*(x), \\
     e^*(x) &= p_0(x)p_1(x),
\end{align*}
$$

whence it follows that

$$
\begin{align*}
    r(h, x) &\geq r(p_1(x), x) = \min \{ p_0(x), p_1(x) \} = r^*(x), \\
    e(h, x) &\geq e(p_1(x), x) = p_0(x)p_1(x) = e^*(x)
\end{align*}

(5)

(6)

for any $h \in R^1$. If $p_1(x) \leq 1/2$, then

$$
r(h, x) - r^*(x) = \begin{cases} 
        0, & h \leq 1/2, \\
        p_0(x) - p_1(x), & h > 1/2.
    \end{cases}
$$

If $p_1(x) > 1/2$, then

$$
r(h, x) - r^*(x) = \begin{cases} 
        p_1(x) - p_0(x), & h \leq 1/2, \\
        0, & h > 1/2.
    \end{cases}
$$

Let $p_1(x) \leq 1/2$. For $h \leq 1/2$, $r(h, x) - r^*(x) = 0 \leq 2(e(h, x) - e^*(x))^{1/2}$. For $h > 1/2$,

$$
e(h, x) - e^*(x) = (h - p_1(x))^2 \geq (1/4)(1 - 2p_1(x))^2
$$

$$
= (1/4)(p_0(x) - p_1(x))^2 = (1/4)(r(h, x) - r^*(x))^2.
$$

Thus, for $p_1(x) \leq 1/2$ and all $h$,

$$
r(h, x) - r^*(x) \leq 2\sqrt{e(h, x) - e^*(x)}.
$$

(7)

For the case $p_1(x) > 1/2$, the inequality can be proved similarly. Substituting $h = f(x)$ into (5), (6), and (7) and taking the expectation with respect to $x$ yield

$$
P \{ I_{1/2} (f(x)) \neq y \} \geq P \{ I_{1/2} (p_1(x)) \neq y \} = P^*, \quad E \{ (f(x) - y)^2 \} \geq E \{ (p_1(x) - y)^2 \},
$$

753
for any measurable function $f(x)$. To derive this inequality, the Jensen inequality for the concave function $\sqrt{\cdot}$ was used. Therefore, if $p_{1}(\cdot) \in F$, then

$$P\{I_{1/2}(f(x)) \neq y\} - P\{I_{1/2}(p_{1}(x)) \neq y\} \leq 2\sqrt{E\{(f(x)-y)^2\} - E\{(p_{1}(x)-y)^2\}}$$

for any function $f(\cdot) \in F$, which proves the statement.

Let us consider the problem \[2\]

$$L_{1}(f) = E |f(x) - y| \rightarrow \inf_{f \in F}$$

where $F$ is a set of Borel functions on $x \in X$ such that $g_{1}(\cdot) \in F$, where

$$g_{1}(x) = \begin{cases} 
1, & p_{1}(x) > 1/2, \\
0, & p_{1}(x) \leq 1/2,
\end{cases} \quad p_{1}(x) = P\{y = 1|x\}.$$

The theorem below generalizes estimate (3) and establishes a relationship between the Bayesian risk and the convex functional $L_{1}(f) = E |f(x) - y|$. \[8\]

**Theorem 2.** Let $F$ be a set of Borel functions on $x \in X$ such that $g_{1}(\cdot) \in F$ or $P(\cdot) \in F$, where $\mu(\cdot)$ is any conditional median of the distribution $P[\cdot|x]$ for a fixed $x$. Then the estimate

$$P\{I_{1/2}(f(x)) \neq y\} - \min_{f \text{-measurable}} P\{I_{1/2}(f(x)) \neq y\} \leq 2(R(f) - \min_{f \in F} R(f))$$

holds for any function $f(\cdot) \in F$, where $R(f) = L_{1}(f) = E |f(x) - y|$. \[9\]

**Proof.** For the case where $F$ is the set of all measurable functions on $x \in X$, the theorem is available in [2] (without factor 2 on the right-hand side of (9)). Let us represent

$$P\{I_{1/2}(f(x)) \neq y\} = E_{x}\{P\{I_{1/2}(f(x)) \neq y|x\}\},$$

$$E |f(x) - y| = E_{x}\{E |f(x) - y| |x\}.$$}

Consider the functions $p_{1}(x) = P\{y = 1|x\}$ and $a(h, x) = E \{|h - y||x\}$. The following representations are true:

$$r(h, x) = P\{I_{1/2}(h) \neq y|x\} = \begin{cases} 
1 - p_{1}(x), & h > 1/2, \\
p_{1}(x), & h \leq 1/2,
\end{cases}$$

$$a(h, x) = E \{|h - y||x\} = p_{1}(x)|h - 1| + (1 - p_{1}(x))|h|.$$ Denote $r^{*}(x) = \min\{p_{1}(x), 1 - p_{1}(x)\}$. For any conditional median $\mu(\cdot)$,

$$\mu(x) \in \begin{cases} 
1, & p_{1}(x) > 1/2, \\
[0, 1], & p_{1}(x) = 1/2, \\
0, & p_{1}(x) \leq 1/2,
\end{cases}$$

and, in particular, $g_{1}(x)$ is the conditional median of the distribution $P$ for a fixed $x$, whence

$$r(h, x) \geq r(p_{1}(x), x) = r^{*}(x),$$

$$a(h, x) \geq a(\mu(x), x) = r^{*}(x)$$

for any $h \in \mathbb{R}$. Let us prove the inequality

$$r(h, x) - r^{*}(x) \leq 2(a(h, x) - r^{*}(x)).$$
Consider the functions
\[ \varphi(p, h) = \begin{cases} 1 - p, & h > 1/2, \\ p, & h \leq 1/2 \end{cases} \]
and
\[ \psi(p, h) = p |h - 1| + (1 - p) |h| = \begin{cases} p - h, & h \leq 0, \\ p + h - 2ph, & 0 \leq h \leq 1, \\ h - p, & h \geq 1. \end{cases} \]

Let us show that \( \varphi(p, h) - p \leq 2(\psi(p, h) - p) \) holds for \( p \leq 1/2 \leq 1 - p \). Indeed,
\[ \varphi(p, h) - p = 0 \leq 2(\psi(p, h) - p) = -2h \text{ for } h \leq 0; \]
\[ \varphi(p, h) - p = 0 \leq 2(\psi(p, h) - p) = 2h(1 - 2p) \text{ for } 0 \leq h \leq 1/2; \]
\[ \varphi(p, h) - p = 1 - 2p \leq 2(\psi(p, h) - p) = 2h(1 - 2p) \text{ for } 1/2 < h \leq 1; \]
\[ \varphi(p, h) - p = 1 - 2p \leq 2(\psi(p, h) - p) = 2h(1 - 2p) \text{ for } 1 < h. \]

It can similarly be verified that \( \varphi(p, h) - (1-p) \leq 2(\psi(p, h) - (1-p)) \) holds for \( 1-p \leq 1/2 \leq p \). Thus, inequality (12) has been proved.

Substituting \( h = f(x) \) into (10), (11), and (12) and taking the expectation with respect to \( x \), we obtain
\[
P \{I_{1/2}(f(x)) \neq y\} \geq P \{I_{1/2}(p_1(x)) \neq y\} = P^*,
\]
\[
E |f(x) - y| \geq E |\mu(x) - y|,
\]
\[
P \{I_{1/2}(f(x)) \neq y\} - P^* \leq 2(E |f(x) - y| - E |\mu(x) - y|) \tag{13}
\]
for any Borel function \( f(x) \). The required inequality (9) follows from (13).

The theorem has been proved.

Thus, the minimization of the functional \( L_1(f) = E |f(x) - y| \) over the set of Borel functions \( F \) such that \( g_1(\cdot) \in F \) or \( \mu(\cdot) \in F \) automatically minimizes the functional of Bayesian risk by virtue of (9).

The quadratic risk functional \( L_2(f) \) is known to be minimum for the function of conditional mean \( m(x) = \int_R y P(dy|x) \) of the distribution \( P \). For nonquadratic risk functionals, it is less obvious that their minima are associated with some characteristics of distribution; however, such a relationship can be established for a functional of average absolute deviation (often used in statistical learning theory).

**THEOREM 3.** In the problem of minimization of the risk functional
\[
R(f) = E_{(X, Y)} \max \{(1-\delta)(f(x) - y), \delta(y - f(x))\}
\]
over all the measurable functions \( f(x) \), the minimum is attained on conditional \( \delta \)-quantiles of distribution \( P \), i.e., on functions \( q(x) \) such that \( P \{y \leq q(x) \mid x\} \geq \delta \). In particular, for \( \delta = 0.5 \), the risk functional has the form \( R(f) = (1/2)E |f(x) - y| \) and is minimum on the conditional medians \( \mu(x) \) of the distribution \( P \{\cdot \mid x\} \).

This theorem was formulated in [16, 17]; in the context of stochastic minimax problems, this fact was established in [18, 19]; it is detailed in [20]. Note that \( \delta \)-quantile and median of distribution may be not unique in the general case.

If there is a priori reason to assume that conditional medians of the distribution \( P(\cdot \mid x) \) belong to some class of functions, for example, to a Hilbert space \( H \), then \( F = H \) can be put in (8). This case suggests that there is no error of the approximation (medians) by functions from \( H \). In the general case, the approximation error exists and is estimated in [2, 8, 11, 21].

In solving classification problems, \( \varepsilon \)-insensitive risk functionals of the form
\[
R_\varepsilon(f) = E \max \{0, |f(x) - y| - \varepsilon\}
\]
are often used in [1].
It is easily seen that the functional $L_1(f) = E|f(x) - y|$ is related with $R_0(f)$ by $L_1(f) - \varepsilon \leq R_0(f) \leq L_1(f)$ uniformly with respect to all the Borel functions $f$; therefore, under conditions of Theorem 2, (9) yields the relation

$$\text{P}\{I_{1/2}(f(x)) \neq y\} - \min_{f \in F} \text{P}\{I_{1/2}(f(x)) \neq y\} \leq 2(R_0(f) - \min_{f \in F} R_0(f)) + 2\varepsilon.$$  

Using $\varepsilon$-insensitive risk functionals simplifies the classifier [1] though worsens the classification accuracy by $2\varepsilon$.

Classification problems often employ functionals of the form $R(f) = E\varphi(y - f(x))$ [9], where labels of classes $y \in \{\pm 1\}$, $\varphi(\cdot)$ are a nonnegative convex nondecreasing loss function such that $\lim_{t \to -\infty} \varphi(t) = 0$, and $\varphi(0) = 1$. They are also used to estimate the risk of correct classification similar to (4) and (9).

Note that the feature space $X$ is often discrete in classification problems, for example, it may consist of nodes of a unit cube [6, Ch. III, Sec. 1.3]. In this case, the function $f(x), x \in X$, is specified by a finite, probably very large, number of values, i.e., is a high-dimensional vector.

3. OPTIMIZATION OF REGULARIZED EMPIRICAL RISK FUNCTIONALS AND THE SUPPORT VECTOR MACHINE

As is shown in Sec. 2, a binary classification problem can be reduced to minimizing a convex risk functional. In the general case, it has the form

$$R(f) = Ec(y, f(x)) \rightarrow \min_{f \in F},$$

where $c(y, f(x))$ is a loss function, for example, $c(y, f(x)) = (y - f(x))^2$, $c(y, f(x)) = |y - f(x)|$, $c(y, f(x)) = \max\{0, 1 - yf(x)\}$; $F$ is an admissible class of functions. Denote the set of solutions of problem (14) by $F^*$. As is also shown in the previous section, minimum in such problems may be attained on some characteristic of the distribution of the random vector of observations $z = (y, x)$, for example, of the function of conditional mean $p_1(x)$ or conditional median $x(x)$. If it is reasonable to assume that these characteristics belong to some class of functions $F$, for example, to a subset of a Hilbert space of functions $H$, we may put in (14) that $F \subseteq H$. The statistical learning theory employs various classes of functions (classical Hilbert spaces with given basis, neural-network superpositions, trees, etc. [2]), for example, so-called reproducing Hilbert spaces of functions $H_k$ generated by the kernel $k$.

**Definition 1** (reproducing Hilbert space). A Hilbert space $H_k(X)$ of functions defined on a closed set $X \subseteq \mathbb{R}^n$ is called a reproducing Hilbert space (RHS) if there exists a function of two vector variables $k(\cdot, \cdot)$, defined on the Cartesian product $X \times X$ and having the following properties:

(a) $k(\cdot, x) \in H_k(X)$ $\forall x \in X$;

(b) $f(x) = \langle f, k(\cdot, x) \rangle_k$ $\forall f \in H_k(X)$, $\forall x \in X$ (the reproducing kernel property).

The RHS theory is presented in [7, 21, 22, 23]. In particular, the set of functions $\left\{ f(x) = \sum_k \alpha_k k(\mathbf{x}_i, x) \right\}$ from the RHS $H_k = H_k(X)$ (where $\mathbf{x}_i$ is an arbitrary finite set of points from $X$ and $\left\{\alpha_k\right\}$ is an arbitrary finite set of numbers) is known to be dense in $H_k(X)$.

In classification problems, the distribution $P(\cdot)$ of observations is usually not known completely, but there exists a set of independent observations $\{z_i = (y_i, x_i), i = 1, \ldots, m\}$ of a vector random variable $z = (y, x)$ with the distribution $P(\cdot)$, which is called a training sample in the statistical learning theory. This allows approximating the unknown distribution $P(\cdot)$ with the empirical distribution $P_m(\cdot)$, and the risk functional $R(f) = Ec(y, f(x))$ with the loss function $c(y, f)$ by the empirical mean (empirical risk) $\tilde{R}_m(f) = (1/m)\sum_{i=1}^{m} c(z_i, f(x_i))$.

The minimization problem for the risk functional (14) may generally be ill-posed, i.e., have ambiguous solutions, be unstable with respect to perturbations of the functional. In the statistical classification learning theory, the original risk functional $R(f)$ is replaced with a random approximation $\tilde{R}_m(f)$, i.e., its stochastic perturbation $R(f) + \delta_m(f)$ is considered, where $\delta_m(f) = \tilde{R}_m(f) - R(f)$. Therefore, to find approximate solutions, the Tikhonov method of regularization in a functional (Hilbert) space $H$ is applied [24, 25]. Let us consider the method of regularization in RHS for certain
(empirical) random perturbations of the functional and for general convex (not only quadratic) risk functionals, which reduces to a family of minimization problems for regularized empirical risk

\[ \tilde{R}_m(f) + \lambda \| f \|_k^2 = \frac{1}{m} \sum_{i=1}^{m} c(y_i, f(x_i)) + \lambda \| f \|_k^2 \to \inf_{f \in H_k}, \]

where \( H_k \) is an RHS generated by the kernel \( k \). The regularized problem (15) in RHS turns out to be reduced to finite-dimensional optimization, and for piecewise-linear loss functions, to the quadratic optimization with linear constraints. By virtue of the so-called theorem on representing the solution in an RHS [7, Theorem 4.2; 26], the solution of problem (15) exists and can be represented as

\[ f_m^\lambda(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x), \]

where \( \alpha^m = \{\alpha_i\} \) is an unknown set of real numbers and \( \{x_i\} \) is a known set of observation points. Substituting expression (16) into (15) and using the reproducing kernel property, we arrive at the following finite-dimensional optimization problem:

\[ R_m(\alpha^m) = \frac{1}{m} \sum_{i=1}^{m} c(y_i, \sum_{j=1}^{m} \alpha_j k(x_i, x_j)) + \lambda \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \to \min_{\alpha^m}. \]

If the loss function \( c(\cdot, \cdot) \) is convex and nonnegative, and the matrix \( \{k(x_i, x_j)\} \) is positive definite, this problem has a unique solution \( f_m^\lambda \). In the solution of problem (17), since there is a quadratic penalty in the objective function, most of the coefficients in (16) may be equal to zero. The vectors \( x_i \) corresponding to nonzero coefficients of expansion (16) are called support vectors, and the classification method based on the solution of problems (15)–(17) is called support vector machine [1, 7].

Note that for nonsmooth piecewise-linear loss functions such as \( c(y, f(x)) = |y - f(x)| \), problem (17) is convex and nonsmooth; however, using additional variables can easily reduce it to a quadratic programming problem with linear constraints. The numerical implementation of the method is detailed, for example, in [7, 27].

4. SVM CONVERGENCE AS THE NUMBER OF OBSERVATIONS INCREASES WITHOUT LIMIT

Let us consider the asymptotic properties (as \( m \to \infty \) and \( \lambda \to 0 \)) of solutions \( f_m^\lambda(x) \) of the regularized empirical risk minimization problem (15). In [1, 4, 5], the convergence \( R(f_m^\lambda) \to \inf_{f \in F} R(f) \) is analyzed on the assumption that the class of functions \( F \) is of bounded capacity. The approach is based on establishing the conditions whereby empirical approximations of the risk functional \( R_m^\lambda(f) = \frac{1}{m} \sum_{i=1}^{m} c(z_i, f(x_i)) \) converge to its true value \( R(f) = Ec(z, f(x)) \) uniformly in \( f \in F \), i.e., \( \sup_{f \in F} |R_m^\lambda(f) - R(f)| \to 0 \) as \( m \to \infty \). However, the approximating class of functions is not always of finite capacity (of finite dimension in the Vapnik–Chervonenkis sense [4]). Weaker requirements for the uniform, on a class of functions, convergence of empirical means to the risk functional can be formulated in terms of the Rademacher class complexity [9, Sec. 3]. Note that the condition of uniform convergence of the approximations \( R_m^\lambda(f) \) to \( R(f) \) is not necessary for the minima to converge [28]. Therefore, we will follow another approach based on the stability of regularized solutions \( f_m^\lambda(x) \) with respect to separate observations. Such an approach was used in [7, Sec. 12.1; 29, 30, 31], where the convergence of risk estimates in probability was analyzed. In contrast to these studies, the present paper establishes the conditions for \( \lambda = \lambda(m) \), for which the estimates converge with unit probability \( f_m^{\lambda(m)}(x) \) uniformly in \( x \in X \) to the minimum \( f^* \) of the risk functional \( R(f) \) with the minimum norm. In this sense, the resultant classifiers are asymptotically stable.
Assumption 1 (Properties of loss functions). The loss function \( c(y, \cdot) \) is nonnegative, convex, and Lipschitzian in the second argument with a constant \( L_y \) on the set

\[ \Phi = \{ f(x) | f \in F, x \in X \}. \]

Assumption 2 (Properties of kernels). The reproducing kernel \( k(\cdot, \cdot) \) satisfies the condition \( \sup_{x \in X} |k(x, x)| = K^2 < +\infty \).

Obviously, the loss functions \( c(y, f) = |y - f|, c(y, f) = \max \{0, 1 - yf\} \) satisfy Assumption 1 for any set \( \Phi \), and the function \( c(y, f) = (y - f)^2 \) satisfies this assumption for a bounded set \( \Phi \). Denote

\[ L = \max_{y \in Y} L_y, \quad C = \max_{y \in Y} c(y, 0). \quad (18) \]

The theorem below estimates the nonoptimality (on the average) of approximate solutions \( f_m^\lambda \) as a function of \( m \) and \( \lambda \).

**THEOREM 4** [32, 33]. Let a solution of problem (14) exist and functions \( f_m^\lambda \) be solutions of problem (15). Then on the assumptions made, the estimate

\[ E_m R(f_m^\lambda) \leq R(f^*) + 2 \frac{C + L \| f^* \|_\infty}{\sqrt{m}} + \frac{LK (5LK + 2\sqrt{AC})}{\lambda \sqrt{m}} + \lambda \| f^* \|_k^2 \quad (19) \]

holds for any \( \lambda > 0 \) and \( m \), where the expectation \( E_m \) is taken over all the samples \( \{z_1, \ldots, z_m\} \) with independent equally distributed observations, \( f^* \) is any solution of problem (14), \( \| f^* \|_\infty = \sup_{x \in X} |f(x)| \), and \( \| f^* \|_k \) is the norm of the function \( f^* \) in the space \( H_k \).

The theorem ensures that \( R(f_m^\lambda) \) converges on the average to the minimum value \( R(f^*) \) as \( \lambda(m) \to 0 \) and \( \sqrt{m} \lambda(m) \to 0 \) as \( m \to \infty \).

Let us specify the strong consistency conditions for estimates \( f_m^\lambda(x) \), i.e., the conditions of their convergence, uniformly in \( x \in X \), to a minimum \( f^*(x) \) of the risk functional \( R \) as \( \lambda = \lambda(m) \to 0 \) and \( m \to \infty \).

**Definition 2** [24]. A solution \( f^* \in F^* \) of the problem is called normal if it has minimum norm, \( \| f^* \|_k = \min_{f \in F^*} \| f \|_k \).

The next two theorems from [32, 33] provide sufficient conditions whereby the approximate solutions \( f_m^\lambda(m) \) converge uniformly and with unit probability to the normal solution \( f^* \in F^* \) of problem (14), i.e.,

\[ \lim_{m \to \infty} \sup_{x \in X} |f_m^\lambda(m)(x) - f^*(x)| = 0. \]

**THEOREM 5** (sufficient conditions for SVM strong consistency). Let a solution of problem (14) exist and Assumptions 1 and 2 be true. Let us consider a family of solutions \( f_m^\lambda(m) \) of problem (15), with \( \lim_{m \to \infty} \lambda(m) = 0 \). Then if \( \lim_{m \to \infty} m\lambda^2(m)/\ln m = \infty \), then \( R(f_m^\lambda(m)) \to R(f^*) \). If \( \lim_{m \to \infty} m\lambda^4(m)/\ln m = \infty \), then \( R(f_m^\lambda(m)) \to R(f^*) \) and solutions \( f_m^\lambda(m) \) of problem (15) converge to the normal solution \( f^* \) of problem (14) uniformly in \( x \in X \) and with unit probability as \( m \to +\infty \).

**THEOREM 6** (estimate of the rate of convergence of the SVM). Let in the conditions of the previous theorem \( \lambda(m) = \Lambda(\ln m)^e/m^{1/4} \), \( \Lambda > 0 \), \( 1/4 < e \leq 1 \), then statements of Theorem 5 are true and the following estimate holds:

\[ E_m R(f_m^\lambda(m)) - R(f^*) \leq 2 \frac{C + L \| f^* \|_\infty}{\sqrt{m}} + \frac{LK (5LK + 2\sqrt{AC})}{\Lambda(\ln m)^e \sqrt{m}} + \frac{\| f^* \|_k^2}{\Lambda(\ln m)^e \sqrt{m}}. \quad (20) \]
5. SVM EFFICIENCY IN SOLVING BINARY CLASSIFICATION PROBLEMS

Using the solution \( f_m^\lambda \) of problem (15), the binary classifier can be constructed as follows:

\[
I_{1/2}(f_m^\lambda(x)) = \begin{cases} 
1, & f_m^\lambda(x) > 1/2, \\
0, & \text{otherwise}.
\end{cases} \tag{21}
\]

For the given training sample, the efficiency of the classifier is measured by the misclassification probability

\[
\Delta_m^\lambda = P\{I_{1/2}(f_m^\lambda(x)) \neq y\} - \min_{f \in F} P\{I_{1/2}(f(x)) \neq y\},
\]

which can be estimated from above in terms of the differences \( [R(f_m^\lambda) - R(f^*)] \) according to inequalities (4) and (9) from Theorems 1 and 2 provided that the conditional medians and means belong to the feasible set \( F \) of the risk minimization problem (14). To obtain the mean misclassification probability, the expectation \( E_m\Delta_m^\lambda \) should be taken over all the independent training samples \( \{(y_j, x_j)\} \) of size \( m \). In turn, the mean value \( E_m\Delta_m^\lambda \) of the error of risk functional minimization over all the possible training samples can be estimated by inequalities (19) and (20) from Theorems 4 and 6. Thus, we arrive at the following results.

**THEOREM 7** (SVM efficiency estimate for nonsmooth risk functional \( L_1(f) \)). Assume that the conditional median \( f^*(x) \) of the probabilistic distribution \( P \) of independent elements of the training sample \( \{(y_j, x_j)\} \) belongs to a subset \( F \) of a reproducing Hilbert space \( H_k \) with the reproducing kernel \( k \). For the binary classifier (21), where the function \( f_m^\lambda(x) \) is a solution of problem (15) with the loss function \( c(y, f) = |y - f| \) or \( c(y, f) = \max\{0, 1 - yf\} \), the misclassification error averaged over all the training samples \( \{(y_j, x_j)\} \) of size \( m \) can be estimated as follows:

\[
E_m\Delta_m^\lambda \leq 4 \frac{2C + L\| f^\ast \|_\infty}{\sqrt{m}} + \frac{2LK(5LK + 2\sqrt{2LC})}{\lambda \sqrt{m}} + \frac{2\| f^\ast \|_k^2}{m},
\]

where the constants \( L \) and \( C \) are defined in (18), and the constant \( K \) is defined in Assumption 2. For \( \lambda(m) = \Lambda \ln m / m^{1/4} \), \( \Lambda > 0 \), this estimate becomes

\[
E_m\Delta_m^\lambda \leq 4 \frac{2C + L\| f^\ast \|_\infty}{\sqrt{m}} + \frac{2LK(5LK + 2\sqrt{2LC})}{\Lambda \ln m \sqrt{m}} + \frac{2\| f^\ast \|_k^2}{\sqrt{m}},
\]

**THEOREM 8** (SVM efficiency estimate for a quadratic risk functional \( L_2(f) \)). Assume that the conditional mean \( p_1(x) = P\{y = 1|x\} = E\{y|x\} \) of the probabilistic distribution \( P \) of independent elements of the training sample \( \{(y_j, x_j)\} \) belongs to a subset \( F \) of a reproducing Hilbert space \( H_k \) with the reproducing kernel \( k \). For the binary classifier (21), where the function \( f_m^\lambda(x) \) is a solution of problem (15) with the quadratic loss function \( c(y, f) = (y - f)^2 \), the misclassification error averaged over all the training samples of size \( m \) can be estimated as follows:

\[
E_m\Delta_m^\lambda \leq \left( 2 \frac{2C + L\| p_1 \|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{2LC})}{\lambda \sqrt{m}} + \frac{\| p_1 \|_k^2}{\ln m} \right)^{1/2},
\]

where the constants \( L \) and \( C \) are defined in (18), and the constant \( K \) is defined in Assumption 2. For \( \lambda(m) = \Lambda \ln m / m^{1/4} \), \( \Lambda > 0 \), this estimate becomes

\[
E_m\Delta_m^\lambda \leq \left( 2 \frac{2C + L\| p_1 \|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{2LC})}{\Lambda \ln m \sqrt{m}} + \frac{\| p_1 \|_k^2}{\sqrt{m}} \right)^{1/2}.
\]

759
CONCLUSIONS

We can make several conclusions concerning the application of the SVM to solve binary classification problems.

When using the SVM, it is important to define correctly the class of functions $F$ and the space $H \supseteq F$ to which the conditional medians and conditional mean of the probabilistic distribution of elements of the training sample belong. In this case, it is said that there is no error of approximation of the median and the mean by functions from $F \subseteq H$. Since the theoretical distribution of training data is not known, and there is only a finite sample of observations with this distribution, choosing the space $H$ and its subset $F$ for a specific implementation of the SVM is not formalized. If $F = H = H_k$ is an RHS of functions, then constructing a classifier reduces to a quadratic programming problem.

The SVM is consistent (if the approximation error is absent), namely, if the regularization parameter $\lambda(m)$ is chosen according to the conditions $\lim_{m \to \infty} \lambda(m) = 0$ and $\lim_{m \to \infty} \lambda(m)/\sqrt{m} = \infty$, then the misclassification probability tends to a theoretical minimum (on the average and with respect to probability) for any distribution of training data. However, the resultant estimates of the rate of convergence of the mean classification error to the minimum contain unknown constants $(\| f^* \|_\infty, \| f^* \|_k, \| P_1 \|_\infty, \text{ and } \| P_1 \|_k^2)$ that depend on the probabilistic distribution of elements of the training sample.

The mean SVM misclassification probability converges to the minimum (as the size $m$ of the training sample increases) with the rate of convergence of order $\cong \sqrt{m}$ if the functional of absolute deviation $L_1(f)$ is used and of order $\cong m^{-1/2}$ if the quadratic risk functional $L_2(f)$ is used. The estimates of the rate of convergence do not explicitly contain the dimension of feature space (dimension of the vector $x$); however, this dimension may appear in the estimates through the constant $K$, which characterizes the reproducing kernel $k$ of the space $H_k$. For example, for a polynomial kernel of the form $k(x, x') = (1 + \langle x, x' \rangle)^q$, $q \geq 1$, and an $n$-dimensional vector $x$ with binary components, the constant has the form $K = (1 + n)^q$.

In conclusion, note that for stronger assumptions on the distribution $P$ of the training data, the rate of convergence of the SVM may be much higher than in Theorems 7 and 8, for example, of order $\cong 1/m$ [9].

REFERENCES