Machine Learning (2019) 108:381–423
https://doi.org/10.1007/s10994-018-5742-0

Rethinking statistical learning theory: learning using statistical invariants

Vladimir Vapnik[1,2] · Rauf Izmailov[3]

Received: 2 April 2018 / Accepted: 25 June 2018 / Published online: 18 July 2018

Abstract
This paper introduces a new learning paradigm, called Learning Using Statistical Invariants (LUSI), which is different from the classical one. In a classical paradigm, the learning machine constructs a classification rule that minimizes the probability of expected error; it is data-driven model of learning. In the LUSI paradigm, in order to construct the desired classification function, a learning machine computes statistical invariants that are specific for the problem, and then minimizes the expected error in a way that preserves these invariants; it is thus both data- and invariant-driven learning. From a mathematical point of view, methods of the classical paradigm employ mechanisms of strong convergence of approximations to the desired function, whereas methods of the new paradigm employ both strong and weak convergence mechanisms. This can significantly increase the rate of convergence.

Keywords Intelligent teacher · Privileged information · Support vector machine · Neural network · Classification · Learning theory · Regression · Conditional probability · Kernel function · Ill-Posed problem · Reproducing Kernel Hilbert space · Weak convergence

Mathematics Subject Classification 68Q32 · 68T05 · 68T30 · 83C32

1 Introduction

It is known that Teacher–Student interactions play an important role in human learning. An old Japanese proverb says “Better than thousand days of diligent study is one day with a great teacher.” What is it exactly that great Teachers do? This question remains unanswered.

Editors: Alexander Gammerman, Vladimir Vovk, Henrik Boström, Lars Carlsson.

✉ Rauf Izmailov
rizmailov@perspectalabs.com

Vladimir Vapnik
vladimir.vapnik@gmail.com

[1] Columbia University, New York, NY, USA

[2] Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

[3] Perspecta Labs, Basking Ridge, NJ, USA

At first glance, it seems that the information that a Student obtains from his interaction with a Teacher does not add too much to the standard textbook knowledge. Nevertheless, it can significantly accelerate the learning process.

This paper is devoted to mechanisms of machine learning, which include elements of Teacher–Student (or Intelligent Agent—Learning Machine) interactions. The paper demonstrates that remarks of a Teacher, which can sometimes seem as trivial (e.g., in the context of digit recognition, "in digit 'zero', the center of the image is empty" or "in digit 'two', there is a tail in the lower right part of the image"), can actually add a lot of information that turns out to be essential even for a large training data set.

The mechanism of Teacher–Student interaction presented in this paper is not based on some heuristic. Instead, it is based on rigorous mathematical analysis of the machine learning problem.

In 1960, Eugene Wigner published the paper "Unreasonable Effectiveness of Mathematics in Natural Sciences" Wigner (1960), in which he argued that mathematical structures "know" something about physical reality. Our paper might as well have the subtitle "Unreasonable Effectiveness of Mathematics in Machine Learning" since the idea of the proposed new mechanism originated in rigorous mathematical treatment of the problem, which only afterwards was interpreted as an interaction between an Intelligent Teacher and a Smart Student.[1]

While analyzing the setting of the learning problem, we take into account some details that were, for simplicity, previously omitted in the classical approach. Here we consider the machine learning problem as a problem of estimating the conditional probability function rather than the problem of finding the function that minimizes a given loss functional. Using mathematical properties of conditional probability functions, we were able to make several steps towards the reinforcement of existing learning methods.

1.1 Content and organization of paper

Our reasoning consists of the following steps:

1. We define the pattern recognition problem as the problem of estimation of conditional probabilities $P(y = k|x),\ k = 1, \dots, n$ (probability of class $y = k$ given observation x): example x_* is classified as $y = s$ if $P(y = s|x_*)$ is maximum. In order to estimate $\max_k P(y = k|x),\ k = 1, \dots, n$, we consider n two-class classification problems of finding $P_k(y^* = 1|x), k = 1, \dots, n$, where $y^* = 1$ if $y = k$ and $y^* = 0$ otherwise.

We start with introducing direct definitions of conditional probability function which differs from standard definition.[2] Let $x \in R^n$. In Sect. 2.2, we define the conditional probability as the solution $f(x) = P(y = 1|x)$ of the Fredholm integral equation of the first kind:

[1] The idea of the new approach is based on analysis of two mathematical facts:

(1) Direct definition of conditional probability and regression functions (Sect. 2.2).
(2) Existence of both strong and weak modes of convergence in Hilbert space, which became the foundation for two different mechanisms of generalization: the classical data-driven mechanism and the new intelligence-driven mechanism (Sect. 6).

[2] The standard definitions of conditional probability function for continuous x is as follows. Let the probability distribution be defined on pairs (x, y). If y takes discrete values from $\{0, 1, \dots, k\}$, the conditional probability $P(y = t|x)$ of $y = t$ given the vector x is defined as the ratio of two density functions $p(y = t, x)$ and $p(x)$:

$$P(y = t|x) = \frac{p(y = t, x)}{p(x)}, \quad y = \{0, 1, \dots, n\}.$$

$$\int \theta(x - x')f(x')dP(x') = P(y = 1, x), \tag{1}$$

where kernel $\theta(z)$ of the integral operator is defined as the *step-function*

$$\theta(z) = \begin{cases} 1, & \text{if } z \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In order to estimate conditional probability from data, we solve equation (1) in the situation where the cumulative distribution functions $P(x)$ and $P(y = 1, x)$ are unknown but iid data

$$(x_1, y_1), \ldots, (x_\ell, u_\ell)$$

generated according to $P(y, x)$ are given.

The advantage of definition (1) is that it is based on the cumulative distribution function, the fundamental concept of probability theory. This definition does not directly require (as in the classical case) the existence of density functions and their ratio.

2. In order to estimate the desired solution from data, we use the following standard inductive step (heuristics): we replace the unknown cumulative distribution functions $P(x)$ and $P(y = k, x)$ with their empirical estimates

$$P_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - x_i), \quad P_\ell(y = k, x) = \frac{1}{\ell}\sum_{j:\{y_j = k\}}^{\ell}\theta(x - x_j),$$

thus obtaining an empirical equation $A_\ell f = F_\ell$ in the form

$$\sum_{i=1}^{\ell}\theta(x - x_i)f(x_i) = \sum_{j=1}^{\ell} y_j\theta(x - x_j).$$

3. The inductive step of replacing the cumulative distribution function with the corresponding empirical distribution function is the main instrument of statistical methods. Justification of this step and analysis of its accuracy are main topics of classical statistics described in Glivenko–Cantelli Theorem (1933) and Kolmogorov–Dvoretzky–Kiefer–Wolfowitz–Massard bounds (1933–1990). The generalization of Glivenko–Cantelli theory, the Vapnik-Chervonenkis theory (VC-theory 1968) plays an important part in justification of learning methods. Results of these theories are outlined in Sect. 1.3.

4. The estimation of conditional probability function by solving Fredholm integral equation $Af(x') = F(x)$ is an *ill-posed* problem. For solving ill-posed problems, Tikhonov and Arsenin (1977) proposed the regularization method, which, under some general conditions, guarantees convergence (in the given metric) of the solutions to the desired function.

In our setting, we face a more difficult problem: we have to solve an ill-posed equation (1) where *both* the operator and the right-hand side of the equation are defined approximately $A_\ell f(x') \approx F_\ell(x)$. In 1978, Vapnik and Stefanyuk (1978) proved that, under some conditions on the operator A of the equation, the regularization method also converges to the desired solution in this case. Section 4.3 outlines the corresponding results.

5. According to Tikhonov's regularization method (Sect. 4.2), in order to solve the operator equation $A_\ell f(x') = F_\ell(x)$, one has to minimize the functional

$$R(f) = \rho^2(A_\ell f(x'), F_\ell(x)) + \gamma W(f(x')),$$

where one has to define the following:

1. the distance $\rho(A_\ell f(x'), F_\ell(x))$ between the function $A_\ell f(x')$ on the left-hand side and the function $F_\ell(x)$ on the right-hand side of the equation;

2. the set of functions $\{f(x')\}$ in which one is looking for the solution of the equation;
3. the regularization functional $W(f(x'))$ and regularization parameter $\gamma > 0$.

In this paper, we define the following elements:

1. we use L_2-metric for distance $\rho^2(A_\ell f(x'), F_\ell(x)) = \int (Af_\ell(x') - F_\ell(x))^2 d\mu(x)$;
2. we solve the equations in the sets of functions $f(x')$ that belong to Reproducing Kernel Hilbert Space (RKHS) of the kernel $K(x, x')$ (see Sect. 5.2).
3. we use the square of a function's norm $||f(x')||^2$ as the regularization functional.

6. In Sect. 5.1, using L_2-distance for estimating the approximation of the solution of the Fredholm equation based on the regularization method, we obtain the non-regularized (with $\gamma = 0$) empirical loss functional

$$R_\ell(f) = \sum_{i=1}^{\ell} (y_i - f(x_i))(y_j - f(x_j))V(i, j), \tag{3}$$

which has to be minimized. Here $V(i, j)$ are elements of the so-called V-matrix that are computed from the data. The classical (non-regularized) empirical loss functional has the form

$$R_\ell(f) = \sum_{i=1}^{\ell} (y_i - f(x_i))^2, \tag{4}$$

which defines the least squares method. The least squares method is a special case of (3), where V-matrix is replaced with the identity matrix I.

> The V-matrix in Eq. (3) has the following interpretation: when we are looking for the desired function, we take into account not only the residuals $\Delta_i = y_i - f(x_i)$, $i = 1, \ldots, \ell$ at the observation points x_i, but also the mutual positions $V(i, j)$ of the observation points x_i and x_j.
> The classical solution of the problem (i.e., the least squares method (4)) uses only information about the residuals Δ_i.

7. Section 5.2.1 shows that, for the Reproducing Kernel Hilbert Space of kernel $K(x, x')$, the function $f(x)$ has the representation (Representer theorem)

$$f(x) = \sum_{i=1}^{\ell} a_i K(x, x_i). \tag{5}$$

The square of norm of this function (which we use as a regularization functional) has the form

$$||f||^2_{rkhs} = \sum_{i,j}^{\ell} a_i a_j K(x_i, x_j). \tag{6}$$

In this paper, we use the following vector-matrix notations:

- ℓ-dimensional vector $Y = (y_1, \dots, y_\ell)^T$,
- ℓ-dimensional vector $A = (a_1, \dots, a_\ell)^T$ of parameters a_i,
- ℓ-dimensional vector-function $\mathcal{K}(x) = (K(x, x_1), \dots, K(x, x_\ell))^T$,
- $(\ell \times \ell)$-dimensional matrix V of elements $V(x_i, x_j),\ i, j = 1, \dots, \ell$,
- $(\ell \times \ell)$-dimensional matrix K of elements $K(x_i, x_j),\ i, j = 1, \dots, \ell$.

In these notations, (5) and (6) can be written as

$$f(x) = A^T \mathcal{K}(x) \quad \text{and} \quad ||f||^2_{rkhs} = A^T K A.$$

8. Section 5.3 shows that, in order to estimate the conditional probability function[3]

$$f(x) = A^T \mathcal{K}(x) \tag{7}$$

using the regularization method, one has to find the vector A that minimizes the functional

$$R(A) = A^T K V K A - 2A^T K V Y + \gamma A^T K A, \tag{8}$$

where the coordinates of vector Y are binary (one or zero).

The minimum of this functional has the form

$$A_V = (VK + \gamma I)^{-1} V Y, \tag{9}$$

where I is the identity matrix. We call such estimate the vSVM method. The classical (square loss) SVM estimate has the form

$$A_I = (K + \gamma I)^{-1} Y. \tag{10}$$

The difference between classical SVM approximation (10) and new vSVM approximation (9) is in its use of identity matrix I instead of V-matrix. When using V-matrix, one takes into account mutual positions of observed vectors $x \in X$.

9. Starting from Sect. 6, we consider the problem of the Teacher–Student interaction. We introduce the following mathematical model of interaction which we call *Learning Using Statistical Invariants* (LUSI).

Let $P(y = 1|x),\ x \in R^n,\ y \in \{0, 1\}$ be the desired conditional probability function. Consider m functions $\psi_s(x),\ s = 1, \dots, m$. There exist constants C_{ψ_s} such that the equalities

$$\int \psi_s(x) P(y = 1|x) dP(x) = C_{\psi_s}, \quad s = 1, \dots, m \tag{11}$$

hold true. We consider as estimates (approximations) $P(y = 1|x)$ the functions satisfying the equations

$$\int \psi_s(x) P_\ell(y = 1|x) dP(x) = C^\ell_{\psi_s}. \tag{12}$$

We say that a sequence of approximations $P_\ell(y = 1|x)$ converges to $P(y = 1|x)$ *in strong mode* if

$$\lim_{\ell \to \infty} ||P(y = 1|x) - P_\ell(y = 1|x)|| = 0,$$

[3] In Sect. 5.3, we consider a more general set of functions $f(x) = A^T \mathcal{K}(x) + c$, where c is a constant. In this introduction, in order to simplify the notations, we set $c = 0$.

in the metric of functions $P(y = 1|x)$. We also say that the sequence of approximations $P_\ell(y = 1|x)$ converges to $P(y = 1|x)$ in *weak mode* if the sequence of values C^ℓ_ψ converges to the value C_ψ with $\ell \to \infty$ for *any* $\psi(x) \in L_2$:

$$\left| \int \psi(x) P_\ell(y = 1|x)dx - \int \psi(x) P(y = 1|x)dx \right| = |C^\ell_\psi - C_\psi| \longrightarrow 0, \quad \forall \psi \in L_2.$$

It is known that strong convergence implies weak convergence.

The idea of LUSI model is to minimize the loss functional (8) in the subset of functions (7) satisfying (12). In order to find an accurate approximation of the desired function, we employ mechanisms of convergence that are based on both strong and weak modes. (The classical method employs only the strong mode mechanism).

The important role in LUSI mechanisms belongs to Teacher. According to the definition, the weak mode convergence requires convergence for *all* functions $\psi(x) \in L_2$. Instead, the Teacher replaces the infinite set of functions $\psi(x) \in L_2$ with a finite set $\mathcal{F} = \{\psi_s(x), \ s = 1, \ldots, m\}$.

Let the Teacher define functions $\psi_s(x), \ s = 1, \ldots, m$ in (11), which we call *predicates*. Suppose that the values C_{ψ_s}, which we call *expressions of predicate*, are known. This fact has the following interpretation: equations (11) describe d (integral) properties of the desired conditional probability function. Our goal is thus to find the approximation that has these properties.

The idea is to identify the subset of functions $P(y = 1|x)$ for which expressions of predicates $\psi_s(x)$ are equal to C_{ψ_s}, and then to select the desired approximation from this subset.

In reality, the values C_{ψ_s} are not known. However, these values can be estimated from the data $(x_1, y_1), \ldots, (x_\ell, y_\ell)$. According to the Law of Large Numbers, the values C_{ψ_s} for the corresponding functions $\psi_s(x)$ can be estimated as

$$C^\ell_{\psi_s} \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i \psi_s(x_i) = \frac{1}{\ell} \Phi_s^T Y, \quad s = 1, \ldots, m,$$

where we have denoted $\Phi_s = (\psi_s(x_1), \ldots, \psi_s(x_\ell))^T$.

Since the structure of conditional probability has the form of (7), we can also estimate the left-hand side of (12). That estimate is $(\ell)^{-1} A^T K \Phi$. From estimates of left- and right-hand sides in (12), one obtains equations (we call them *invariants*)

$$A^T K \Phi_s^T \approx Y^T \Phi_s, \quad s = 1, \ldots, m. \tag{13}$$

The method of LUSI is to find the vector A that defines the approximation (7) by minimizing functional (8) in the set of vectors A satisfying[4] (13).

10. The problem of minimizing the functional (8) subject to the constraints (13) has a closed-form solution. In order to obtain it, one has to do the following:

(1) Compute vector A_V as in (9).
(2) Compute d vectors

$$A_s = (VK + \gamma I)^{-1} \Phi_s, \ s = 1, \ldots, m.$$

[4] In other words, if one wants to find a rule for identification of ducks, the first thing one has to do is to find a set of rules that do not contradict the basic "duck test" of identification, i.e., birds that look like a duck, swim like a duck, and quack like a duck. Then one has to select the rule for identification of ducks *within* this set of rules.

(3) Compute vector-parameters A of estimate (7)

$$A = A_V - \left(\sum_{s=1}^{m} \mu_s A_s \right),$$

where vector-parameters μ_s are the solution of some linear system of equations determined by the data (Sect. 6.2). Vector A is the sum of two estimates:

(1) the estimate A_V, which is obtained from the standard learning scenario by the vSVM method (*data-driven part of estimate*), and
(2) the term shown in parentheses, which is the correction term based on invariants (13) (*intelligence-driven part of estimate*). We call the sum A the vSVM *with m invariants* estimate, denoted as vSVM & I$_m$.

11. The introduction of predicate-functions $\psi_s(x),\ s = 1, \ldots, m$ for constructing invariants reflects the Teacher's intellectual contribution to the Teacher–Student interaction; some examples of possible functions for invariants are provided in Sect. 6.5. The first task of the Student in this interaction is to understand which invariants are suggested by the Teacher, and the second task is to choose the best one from the admissible set of functions that preserve these invariants.

There is an important difference between invariants and features in classical learning models. With increasing number of invariants, the capacity of the set of functions from which Student has to choose the desired one *decreases* (and as a result, according to VC bounds, this leads to a more accurate estimate). In contrast to that, with increasing number of features, the capacity of the set of admissible functions *increases* (and thus, according to VC bounds, this requires more training examples for an accurate estimate[5]).

12. Section 6.4 contains examples that illustrate the effectiveness of the ideas of LUSI and remarks on implementation of SVM& I$_d$ algorithms.

1.2 Phenomenological model of learning

The general mathematical theory of learning was first introduced about fifty years ago. That theory was created for the following phenomenological model:

> In some environment X, there exists a generator G of random events $x_i \in X$. This generator G generates events x randomly and independently, according to some unknown probability measure $P(x)$.
>
> In this environment, a classifier A operates, which labels the events x; in other words, on any event x_i produced by generator G, classifier A reacts with a binary signal $y_i \in \{0, 1\}$. The classification y_i of events x_i is produced by classifier A according to some unknown conditional probability function $P(y = 1|x)$.
>
> The problem of learning is formulated as follows: given ℓ pairs
>
> $$(x_1, y_1), \ldots, (x_\ell, y_\ell) \tag{14}$$
>
> containing events x_i produced by generator G and classifications y_i produced by classifier A, find, in a given set of indicator functions, the one that minimizes probability of discrepancy between classifications of this function and classifier A.

[5] In Vapnik and Izmailov (2017), we showed that, in data-driven estimates, ℓ examples $(x_i, y_i), i = 1, \ldots, \ell$ can provide no more than ℓ bits of information. However, using one invariant in intelligence-driven estimates, ℓ examples can provide more than ℓ bits of information (Sect. 6.3).

This phenomenological model describes the basic learning problem, namely, the so-called two-class pattern recognition problem. Its generalizations to an arbitrary finite number $n > 2$ of classes $y \in \{a_1, \dots, a_n\}$ are just technical developments of this simplified model; they are addressed later in the paper.

1.3 Risk minimization framework

The first attempt to construct the general learning theory was undertaken in the late 1960s—beginning of 1970s. At that time, the learning was mathematically formulated as the problem of expected risk minimization (Vapnik and Chervonenkis 1974; Wapnik and Tscherwonenkis 1979).

1.3.1 Expected risk minimization problem

In a given set of functions $\{f(x)\}$, find the one that minimizes the expected risk functional

$$R(f) = \int L(y, f(x))dP(x, y) \tag{15}$$

(here $L(y, f(x))$ is a given *Loss Function*), under the assumption that probability measure $P(x, y)$ is unknown however iid data (14) generated according to the phenomenological model $P(x, y) = P(y|x)P(x)$ described above is available. In this paper, we focus on the special case of $\{f(x)\}$ being a set of indicator functions (this case is called *two-class classification rule*) and then generalize the obtained results for estimating conditional probability functions (real-valued functions $0 \leq f \leq 1$). In this paper, we use the loss function

$$L(y, f) = (y - f(x))^2. \tag{16}$$

For pattern recognition case, the function $f(x)$ that minimizes the functional (16) belongs to a set of indicator functions. The function $f \in \{f(x)\}$ that minimizes functional (15) with loss (16) will have the smallest probability of error among all functions in $\{f(x)\}$.

1.3.2 Empirical loss minimization solution

In order to find such functions, the method of minimizing the empirical risk functional

$$R_\ell(f) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(y_i, f(x_i)) \tag{17}$$

was introduced. The function $f_\ell(x) \in \{f(x)\}$ which minimizes (17) is considered as the function f for which the value $R(f)$ of functional (14) (i.e., the value of the expected error) is close to the minimum: $R(f_\ell) \approx \min_f R(f)$.

The choice of empirical risk minimization method is based on the following consideration: since the cumulative distribution function $P(y, x)$ is unknown but data (14) is given, one approximates the unknown cumulative distribution function by its empirical estimate, the joint *empirical cumulative distribution function*

$$P_\ell(y, x) = \sum_{i=1}^{\ell} \theta(y - y_i)\theta(x - x_i), \tag{18}$$

where the one-dimensional step function is defined as

$$\theta(z) = \begin{cases} 1, & \text{if } z \geq 0, \\ 0, & \text{if } z < 0, \end{cases} \tag{19}$$

and the multi-dimensional step function ($x = (x^1, \ldots, x^n)$) is defined as

$$\theta(x - x_i) = \prod_{k=1}^{n} \theta(x^k - x_i^k). \tag{20}$$

Using approximation (18) in the expected loss functional (14), we obtain the empirical loss functional (17). In Sect. 3, we argue that empirical cumulative distribution function is a good approximation for the cumulative distribution function. The minimization of functional (17) is the main inductive instrument in estimation theory.

1.3.3 Problems of theory

The theory of empirical risk minimization was developed to address the following questions:

1. When is the method of empirical loss minimization consistent?
2. How close is the value of expected loss to the minimum across the given set of functions?
3. Is it possible to formulate a general principle that is better than the empirical loss minimization?
4. How to construct algorithms for estimation of the desired function?

1.3.4 Main results of VC theory

Forty years ago, a theory that addresses all four questions above was developed; it was the so-called *Statistical Learning Theory* or *VC Theory*.

1. The VC theory defines the necessary and sufficient conditions of consistency for both (i) the case when the probability measure $P(x)$ of the generator G in the phenomenological model is unknown (in this case, the necessary and sufficient conditions are valid for *any* probability measure $P(x)$), and (ii) for the case when the probability measure $P(x)$ is known (in this case, the necessary and sufficient conditions are valid for a *given* probability measure $P(x)$).
   In both of these cases, the conditions for consistency are described in terms of the capacity of the set of functions. In the first case (consistency for any probability measure), the VC dimension of the set of indicator functions is defined (it has to be finite). In the second case (consistency for the given probability measure), the VC entropy of the set of function for the given probability measure is defined (VC entropy over number of observations has to converge to zero).
2. When VC dimension of a set of indicator functions is finite, the VC theory provides bounds on the difference between the real risk that exists for the function f_ℓ that minimizes empirical risk functional (15) (the value $R(f_\ell)$) and its empirical estimate (17) (i.e., the value $R_\ell(f_\ell)$). That difference depends on the ratio of VC dimension h to number of observations ℓ. Specifically, with probability $1 - \eta$, the bound

$$R(f_\ell) - R_\ell(f_\ell) \leq T\left(\frac{h - \ln \eta}{\ell}\right) \tag{21}$$

holds, which implies the bound

$$R(f_\ell) - \inf_f R(f) \leq T_* \left(\frac{h - \ln \eta}{\ell} \right). \tag{22}$$

In (21) and (22), $T(h/\ell)$ and $T_*(h/\ell)$ are known monotonically decreasing functions.

3. VC theory introduces a generalization of the empirical risk minimization method, the so-called method of *Structural Risk Minimization* (SRM), which is the basic instrument in statistical inference methods. In the SRM method, a nested set of functions

$$S_1 \subset S_2 \subset \ldots \subset S_n \subset \ldots \tag{23}$$

is constructed on the set of given function $\{f\}$; here S_k is the subset of all functions $\{f\}$ that have VC dimension $h^{(k)}$. For subset S_k, the bound (21) has the form

$$R(f_\ell^{(S_k)}) \leq R_\ell(f_\ell^{(S_k)}) + T \left(\frac{h^{(k)} - \ln \eta}{\ell} \right). \tag{24}$$

In order to find the function that provides the smallest guaranteed risk for the given observations, one has to find both the subset S_k of the structure (23) and the function f in this subset that provides the smallest (over k and $f \in S_k$) value of the right-hand side of inequality (24). *Structural Risk Minimization (SRM) method is a strongly universally consistent risk minimization method.*

One can consider the SRM method as a mathematical realization of the following idea: *Chose the simplest*[6] *function (one from the subset with the smallest VC dimension) that classifies training data well.*

4. Based on VC theory, an algorithm for solving pattern recognition problems was developed—it was the Support Vector Machine (SVM) algorithm which realized the SRM method.

In this paper, we consider an SVM type of algorithm for square loss function (16) as the baseline algorithm for comparisons. We introduce two new ideas which form the basis for a new approach. One idea (presented in Part One of paper) is technical and another one (presented in Part Two) is conceptual.

Part One: The *V*-matrix estimate

In this first part, we describe our technical innovations that we use in the second part for constructing algorithms of inference.

2 The observation that defines the *V*-matrix method

Our first idea is related to the mathematical understanding of a learning model that is different from the phenomenological scheme described in Sect. 1.2.

In the classical approach to the learning methods (analyzed by the VC theory), we ignored the fact that the desired decision rule is related to the conditional probability function used

[6] The definition of function simplicity is not trivial. In order to formalize the concept of simplicity for a set of functions, K. Popper introduced the concept of falsifiability of the set of functions (Popper 1934). However, the mathematical formalization of his idea contained an error. The corrected formalization of the falsifiability concept leads to the concept of VC-dimension—see Corfield et al. (2005, 2009) for details.

by Classifier A. We just introduced some set of indicator functions and defined the goal of learning as finding the one that guarantees the smallest expected risk of loss (15) in this set of functions.

Now we take into account the fact that the best rule for classification within n classes has the form

$$r(x) = \operatorname{argmax}_{s\in\{1,\ldots,n\}}(P(y=1|x),\ldots,P(y=s|x),\ldots,P(y=n|x)),$$

where $P(y=s|x)$ is the probability of class $y=s$ given observation x. For a two-class pattern recognition problem (where $y\in\{0,1\}$), this rule has the structure

$$r(x)=\theta\left(P(y=1|x)-\frac{1}{2}\right), \tag{25}$$

where $\theta(z)$ is the step function. Thus, the rule $r(x)$ classifies vector x as belonging to the first class ($y=1$) if the conditional probability of the first class exceeds 0.5.

We consider the problem of estimating the conditional probability function $P(y=s|x)$ using data (14) as the main problem of learning. The construction of the classification rule (25) based on the obtained conditional probability function is then a trivial corollary of this solution.

2.1 Standard definitions of conditional probability and regression functions

In statistical theory, the basic properties of random events $x\in R^n$ are described by their cumulative distribution function. The cumulative distribution function of variable $x=(x^1,\ldots,x^n)$ is defined as the probability that a random vector $X=(X^1,\ldots,X^n)$ does not exceed the vector x (coordinate-wise):

$$P(x)=P(X^1\le x^1,\ldots,X^n\le x^n).$$

The probability density function (if it exists) is defined as derivative of $P(x)$:

$$p(x)=\frac{\partial^n P(x)}{\partial x^1\cdots\partial x^n}. \tag{26}$$

Consider pair (x,y), where $x\in R^n$ and $y\in R^1$ are continuous variables. The ratio of two density functions

$$p(y|x)=\frac{p(y,x)}{p(x)} \tag{27}$$

is called *conditional density function*. The expectation function

$$r(x)=\int yp(y|x)dy \tag{28}$$

over y for any fixed x is called *regression function*.

If y is discrete, say $y\in\{0,1\}$, the ratio of two densities $p(x,y)$ and $p(x)$

$$p(y=1|x)=\frac{p(y=1,x)}{p(x)}=\frac{p(x|y=1)p(y=1)}{p(x)} \tag{29}$$

is called *conditional probability function*. Function $p(y=1|x)$ defines probability of classification $y=1$ given vector x. In this definition, the factor $p(y=1)$ is the probability of event $y=1$ in trials with (x,y).

The basic problem of statistics is as follows: given the set of iid pairs

$$(x_1, y_1), \dots, (x_\ell, y_\ell)$$

generated according to $P(x, y)$, estimate the conditional probability function.

In this paper, we estimate the conditional probability function. We assume that $x \in \mathcal{R}^d$, where $\mathcal{R}^d = [a_1, c_1] \times \cdots \times [a_d, c_d]$. To simplify the notations, we assume, without loss of generality, that $a_1 = \ldots = a_s = \ldots = a_d = 0$, so that $\mathcal{R}^d = \{x \in [0, \mathbf{c}]^d\}$, where $\mathbf{c} = (c_1, \dots, c_d)^T$.

2.2 Direct definitions of conditional probability function

In this section, we consider another definitions of the conditional probability function. We define the conditional probability function $f(x)$ as the solution of the Fredholm equation

$$\int_{\mathcal{R}^n} \theta(x - x') f(x') dP(x') = P(y = 1, x). \tag{30}$$

Formally, this definition of the conditional probability function does not require the existence of density function. If derivatives of functions of $P(x)$ and $P(y = 1, x)$ exist, taking derivative over x on the left- and right-hand sides of Eq. (30), we obtain

$$f(x)p(x) = p(y = 1, x),$$

which is the definition of conditional probability (29).

Similarly, we define the regression function $f(x)$ as the solution of Fredholm equation

$$\int_{\mathcal{R}^n} \theta(x - x') f(x') dP(x') = \int_{\mathcal{R}^n} \theta(x - x') \int y dP(y, x'). \tag{31}$$

Indeed, taking derivatives of (31) over x, we obtain

$$f(x)p(x) = \int yp(y, x)dy \Longrightarrow f(x) = \int y \frac{p(y, x)}{p(x)} dy = \int yp(y|x)dy,$$

which is the definition of regression (28). In this paper, however, we consider only conditional probability estimation problem.

In order to estimate the conditional probability function or the regression function, we need to find the solution of Eqs. (30) or (31), Note that cumulative distribution functions that define these equations are unknown but corresponding iid data $(x_1, y_1), \dots, (x_\ell, y_\ell)$ are given.

2.3 Estimation of conditional probability function for classification rule

This paper is devoted to the estimation of the solutions of Eq. (30). To find the classification rule $r(x)$ in two-class classification problem, we use the estimated conditional probability function

$$r(x) = \theta\left(P(y = 1|x) - \frac{1}{2}\right). \tag{32}$$

At first glance, such approach to pattern recognition problem seems to be an overkill since we connect the solution of a relatively simple problem of finding an indicator function to a much more difficult (ill-posed) problem of estimating the conditional probability functions. The explanation why it is nevertheless a good idea is the following:

1. The conditional probability function directly controls many different statistical invariants that manifest themselves in the training data. Preserving these invariants in the rule is equivalent to incorporating some prior knowledge about the solution. Technically, this allows the learning machine to extract additional information from the data—the information that cannot be extracted directly by existing classical methods. Section 6 is devoted to the realization of this idea.
2. Since the goal of pattern recognition is to estimate the classification rule (32) (not the conditional probability function), one can relax the requirements of accuracy of the conditional probability estimation. We really need to have an accurate estimate of the function in the area $x \in X$ where values $P(y = 1|x)$ are close to 0.5; conversely, we can afford to have less accurate estimates in the area where $|P(y = 1|x) - 0.5|$ is large. This means that the cost of error of deviation of the estimate $P_\ell(y = 1|x)$ from the actual function $P(y = 1|x)$ can be monotonically connected to the variance $\phi(x) = P(y = 1|x)(1 - P(y = 1|x))$ (the larger is the variance, the bigger is the cost of error). This fact can be taken into account when one estimates conditional probability functions.

As we will see in Sect. 6, the model of estimating the conditional probability function controls more important factors of learning than the classical expected risk minimization model considered in the VC theory.

2.4 Problem of inference from data

Statistical inference problem defined in Sect. 2.2 requires solving the Fredholm integral equation

$$Af = F, \tag{33}$$

where operator A has the form

$$Af = \int \theta(x - x')f(t)dP(t). \tag{34}$$

This operator maps functions $f \in E_1$ to functions $Af \in E_2$. In the pattern recognition problem, the operator A maps the set $\mathcal{P}$ of non-negative bounded functions $0 \leq f(x) \leq 1$ into the set of non-negative bounded functions $0 \leq F(x) = P(y = 1, x) \leq P(y = 1)$, where $P(y = 1)$ is the probability of class $y = 1$. The problem is, for any function $F(x) = P(y = 1, x)$ in the right-hand side of (30), to find the solution of this equation in the set $\mathcal{P}$.

Note that the solution of operator equations in a given set of functions is, generally speaking, an ill-posed problem. The problem of statistical inference is to find the solution defined by the corresponding equation [(30) or (31)] when both the operator and the right-hand side of the equation are unknown and have to be approximated from the given data (14).

In all cases, we approximate the unknown cumulative distribution functions $P(x)$, function $P(y = 1, x) = P(x|y = 1)P(y = 1)$ and $P(x, y)$ using *Empirical Cumulative Distribution Functions*

$$P_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - x_i), \tag{35}$$

$$P_\ell(1,x) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\theta(x-x_i), \tag{36}$$

$$P_\ell(x,y) = \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(x-x_i)\theta(y-y_i). \tag{37}$$

Putting (36) and (37) into Eq. (30) instead of unknown elements $P(x)$ and $P(y=1,x)$, and putting (35) and (37) into Eq. (31) instead of unknown elements $P(x)$ and $P(y,x)$, we obtain (taking into account that the derivative of $\theta(x-x_i)$ is δ-function $\delta(x-x_i)$) the following approximation to the Eq. (30) and the same approximation to the Eq. (31):

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \theta(x-x_i)f_1(x_i) \approx \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\theta(x-x_i). \tag{38}$$

> ***In the remainder of the paper, we use equation*** (38) ***only for estimating conditional probability function***[7] (***where*** $y \in \{0,1\}$).

Similarly, in order to estimate the conditional probability $P(y=0|x)$, one has to solve the equation

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \theta(x-x_i)f_0(x_i) \approx \frac{1}{\ell}\sum_{i=1}^{\ell} (1-y_i)\theta(x-x_i). \tag{39}$$

The problem of the new approach to learning is:

Find the approximation to the desired function by solving Eq. (38).

Theory of this learning paradigm has to answer the following four questions:

1. Why is the replacement of Cumulative Distribution Functions with their approximation (30) a good idea?
2. How to solve the ill-posed problem of statistical inference when both the operator and the right-hand side of the equation are defined approximately?
3. How to incorporate the existing statistical invariants into solutions?
4. What are constructive algorithms for inference?

The next sections are devoted to the answers to these questions.

3 Main claim of statistics: Glivenko–Cantelli theorem

Consider the following approximation of a cumulative distribution function $P(z)$ obtained for iid observations $z_1,\ldots,z_\ell$ generated according to $P(z)$:

$$P_\ell(z) = \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(z-z_i).$$

In 1933, Glivenko and Cantelli proved the main theorem of statistics:

[7] In the case of regression function (where $y \in R^1$), the right-hand side of (38) converges to the right-hand side of Eq. (31) not as fast as the empirical cumulative function converges to the actual one in the right-hand side of Eq. (30).

Theorem *Empirical Distribution Functions $P_\ell(z)$ converge to the actual cumulative distribution function $P(z)$ uniformly with probability 1:*

$$\sup_z |P_\ell(z) - P(z)| \longrightarrow^{P}_{\ell\to\infty} = 0.$$

In the same 1933, Kolmogorov found the asymptotically exact rate of convergence for the case $z \in R^1$. He showed that the equality

$$P\left[\lim_\ell \sqrt{\ell} \sup_x |P_\ell(z) - P(z)| > \varepsilon\right] = 2\sum_{k=1}^{\ell}(-1)^{k-1}\exp -2\varepsilon^2 k^2, \quad \forall\varepsilon \tag{40}$$

holds true. Later (in 1956–1990), Dvoretzky–Kiefer–Wolfowitz–Massard found the sharp non-asymptotic bound with its right-hand side coinciding with the first term of Kolmogorov equality:

$$P\left[\sup_x |P_\ell(x) - P(x)| > \varepsilon\right] \le 2\exp -2\varepsilon^2\ell, \quad \forall\varepsilon. \tag{41}$$

The generalization of the bound for $z \in R^n$ was obtained in 1970 using the VC theory:

$$P\left[\sup_z |P_\ell(z) - P(z)| > \varepsilon\right] \le 2\exp\left\{-\left(\varepsilon^2 - \frac{n\ln\ell}{\ell}\right)\ell\right\}, \quad \forall\varepsilon. \tag{42}$$

In order to obtain constructive equations for pattern recognition problem, we replaced the cumulative distribution functions in (30), (31) with the corresponding empirical distribution functions.

4 Solution of ill-posed problems

In this section, we outline regularization principles that we apply to the solution of the described problems.

4.1 Well-posed and ill-posed problems

Let A be a linear operator which maps the elements f of a metric space E_1 into the elements F of a metric space E_2. We say that problem of solving operator equation

$$Af = F \tag{43}$$

in the set $\{f\}$ is *well-posed* if the solution *exists*, is *unique*, and is *continuous*. That is, if the functions F_1 and F_2 of the right-hand side of Eq. (43) are close in the metric of space E_2 (i.e., $\rho_{E_2}(F_1, F_2) \le \varepsilon$), they correspond to solutions f_1 and f_2 that are close in the metric of space E_1 (i.e., $\rho_{E_1}(f_1, f_2) \le \delta$). The problem is called *ill-posed* if at least one of the three above conditions is violated. Below we consider the ill-posed problems where unique solutions exist, but the inverse operator

$$f = A^{-1}F$$

could be discontinuous. Solving of inference problems defined by the Fredholm equation

$$\int_0^1 \theta(x-t)f(t)dP(x) = P(y=1, x)$$

is ill-posed. The problem of statistical inference requires *to solve ill-posed Fredholm equations when both right-hand side F and operator $A$of* Eq. (43) *are defined approximately.*

4.2 Regularization of ill-posed problems

The solution of ill-posed problems is based on the following lemma.

Lemma (Lemma about inverse operator) *If A is a continuous one-to-one operator A defined on* compact set $\mathcal{M}$ *of functions* $\{f\}$*, then the inverse operator A^{-1} is continuous on the set $\mathcal{N} = A\mathcal{M}$.*

Consider a continuous non-negative functional $W(f)$ and the set of functions defined by the constant $C > 0$ as

$$\mathcal{M}_C = \{f : W(f) \leq C\}. \tag{44}$$

Let the set of functions $\mathcal{M}_C$ be convex and compact for any C. Suppose that the solution of operator equation belongs to compact sets $\mathcal{M}_C$ with $C \geq C_0$.

The idea of solving ill-posed problems is to choose an appropriate compact set (i.e., to choose a constant C^*) and then solve Eq. (43) on the compact set of functions[8] *defined by C^*. In other words, to minimize the square of distance in E_2 space*

$$\rho = \rho^2_{E_2}(Af, F) \tag{45}$$

over functions f subject to the constraint

$$W(f) \leq C^*. \tag{46}$$

The equivalent form of this optimization problem is Tikhonov's regularization method. In this regularization method, the functional

$$R(f) = \rho^2_{E_2}(F(x), Af) + \gamma W(f) \tag{47}$$

is minimized, where $\gamma > 0$ is the regularization constant.

The expression (47) is the Lagrangian functional for the optimization problem minimizing (45) subject to (46), where parameter γ is defined by the parameter C that defines the chosen compact set (44). The parameter $\gamma = \gamma^*$ in (47) should chosen be such that the equality

$$W(f_*) = C^*$$

holds true for the solution f^* of the minimization problem.

In 1963, Tikhonov proved the following theorem:

Theorem *Let E_1 and E_2 be metric spaces and suppose that, for $F \in E_2$, there exists a solution of the equation*

$$Af = F$$

that belongs to the set $f \in \{W_{E_1}(f) \leq C\}$ for $C > C_0$. Let the right-hand side F of this equation be approximated with F_δ such that $\rho(F, F_\delta) \leq \delta$. Suppose that the values of (regularization) parameters $\gamma(\delta)$ are chosen such that

$$\gamma(\delta) \longrightarrow 0, \quad \textit{for } \delta \longrightarrow 0$$

[8] Note that this idea of solving ill-posed problems is the same as in structural risk minimization in VC theory. In both cases, a structure is defined on the set of functions. When solving well-posed problems, elements of structure should have finite VC-dimension. When solving ill-posed problems, elements of structure should be compact sets.

$$\lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} \leq r \leq \infty.$$

Then the elements $f_{\gamma(\delta)}$ *minimizing the functional*

$$R(f) = \rho^2_{E_2}(Af, F(\delta)) + \gamma(\delta) W_{E_1}(f)$$

converge to the exact solution as $\delta \longrightarrow 0$.

4.3 Generalization for approximately defined operator

Let our goal be to solve the operator equation

$$Af = F, \quad f \in E_1, \; F \in E_2,$$

when we are given a sequence of random approximations F_ℓ and random operators A_ℓ, $\ell = 1, 2, \ldots$.

Consider, as solutions of the problem, the sequence of functions f_ℓ, $\ell = 1, 2, \ldots$ minimizing the regularization functional

$$R(f) = \rho^2_{E_2}(A_\ell f, F_\ell) + \gamma W_{E_1}(f).$$

We define the distance between operators as

$$||A - A_\ell|| = \sup_f \frac{||A_\ell f - Af||}{W^{1/2}(f)}.$$

Theorem (Vapnik and Stefanyuk 1978). *For any* $\varepsilon > 0$ *and any* $C_1, C_2 > 0$ *there exists a value* $\gamma_0 > 0$ *such that the inequality*

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} < P\{\rho_{E_2}(F_\ell, F) > C_1\sqrt{\gamma_\ell}\} + P\{||A_\ell - A|| > C_2\sqrt{\gamma_\ell}\}$$

holds true for any $\gamma_\ell \leq \gamma_0$.

Corollary *Let the right-hand side* F_ℓ *of the equation converge to* F *with rate* $r_F(\ell)$ *and let operator* A_ℓ *converge to* A *with rate* $r_A(\ell)$. *Then there exists a function*

$$r_0(\ell) = \lim_{\ell \to \infty} \max\{r_F(\ell), r_A(\ell)\} = 0$$

such that the sequence of solutions converges if

$$\lim_{\ell \to \infty} \frac{r_0(\ell)}{\sqrt{\gamma_\ell}} = 0$$

and $\gamma_\ell \longrightarrow 0$ *as* $\ell \longrightarrow \infty$.

5 Solution of stochastic ill-posed problems

In order to estimate conditional probability function, we find the solution of approximation (38) of Fredholm integral equation (30) using a regularization method (Sect. 4.2). In order to do this, we have to specify three elements of the minimization functional

$$R(f) = \rho^2_{E_2}(A_\ell f, F_\ell) + \gamma W_{E_1}(f). \tag{48}$$

We specify them in the following manner.

1. The distance $\rho^2_{E_2}(A_\ell f, F_\ell)$ in E_2 space: *We select $L_2(\phi)$ metric.*
2. The set of functions $\{f(x)\},\ x \in [0,1]^n$ containing the solution f_ℓ: *We select Reproducing Kernel Hilbert Space (RKHS) of kernel $K(x, x_*)$.*
3. The regularization functional $W_{E_1}(f)$ in space E_1: *We select the square of the norm of function in the RKHS.*

5.1 Distance in space L_2

We define the square of distance between two functions F_1 and F_2 that denote the left- and right-hand sides of Eq. (38) in the following way:

$$\rho^2(F_1, F_2) = \int (F_1(x) - F_2(x))^2 \phi(x) d\mu(x), \tag{49}$$

where $\phi(x) \geq 0$ is a given weight function and $\mu(x)$ is a probability measure defined on d-dimensional domain $\mathcal{C}$ consisting of vectors $(x_1, \ldots, x_d)$, where $0 \leq x_i \leq c_i$ for each $i = 1, \ldots, d$ and $c_1, \ldots, c_d$ are non-negative constants. We select both the weight function and the probability measure later. Using this metric, we compute the square of the distance between the left- and right-hand sides of our approximations (38) of Fredholm equations:

$$\rho^2 = \int_{\mathcal{C}} \left(\sum_{i=1}^{\ell} \theta(x - x_i) f(x_i) - \sum_{j=1}^{\ell} y_j \theta(x - x_j) \right)^2 \phi(x) d\mu(x), \tag{50}$$

where $f \in \{f\}$, $y \in \{0, 1\}$. For estimation of conditional probability function, this can be rewritten as follows:

$$\rho^2 = \sum_{i,j=1}^{\ell} f(x_i) f(x_j) V(i,j) - 2 \sum_{i,j=1}^{\ell} f(x_i) y_j V(i,j) + \sum_{i,j=1}^{\ell} y_i y_j V(i,j), \tag{51}$$

where the constant $V(i, j)$ denotes

$$V(i,j) = \int_{\mathcal{C}} \theta(x - x_i) \theta(x - x_j) \phi(x) d\mu(x). \tag{52}$$

In the d-dimensional case $x = (x^1, \ldots, x^d)$, we have

$$V(i,j) = \int_{\mathcal{C}} \left(\prod_{k=1}^{d} \theta(x^k - x_i^k) \theta(x^k - x_j^k) \right) \phi(x) d\mu(x). \tag{53}$$

For the special case where $x \in [(0, \ldots, 0), \mathcal{C}]$, $\mu(x) = x$, and $\phi(x) = 1$, this expression has the form

$$V(i,j) = \prod_{k=1}^{d} (c_k - \max(x_i^k, x_j^k)). \tag{54}$$

Expression (54) defines V-matrix in *multiplicative* form. For high-dimensional problems, computation of (54) may be difficult; therefore we also define an *additive* form of V-matrix along with (54). Consider, instead of distance (50), the following expression

$$\rho^2 = \int_{\mathcal{C}} \sum_{k=1}^{d} \left(\sum_{i=1}^{\ell} f(x_i^k) \theta(x_s^k - x_i^k) - \sum_{j=1}^{\ell} y_j \theta(x_s^k - x_j^k) \right)^2 \phi(x) d\mu(x), \tag{55}$$

where x^k is the k-th coordinate of vector x and $f(x_i)$ is the value of function in the point $x_i = (x_i^1, \dots, x_i^d)$. For this distance, elements $V(i, j)$ of V-matrix have the form

$$V(i, j) = \sum_{k=1}^{d} (c_k - \max(x_i^k, x_j^k)). \tag{56}$$

In our computations we using multiplicative form of V-matrix.

Matrix V is a symmetric nonnegative matrix; the maximum value of any column (or any row of V) is the value on the intersection of that column (row) with the diagonal of V.

Algorithmic Remark. With increase of dimensionality, it becomes more difficult to realize the advantages of the V-matrix method. This is due to the fact that mutual positions of the vectors in high-dimensional spaces are not expressed as well as in low-dimensional spaces.[9] The mathematical manifestation of this fact is that, in a high-dimensional space, V-matrix can be ill-conditioned and, therefore, require regularization. We used the following regularization method: (1) transform the V-matrix to its diagonal form in the basis of its eigenvectors using the appropriate orthonormal mapping T; (2) add a small regularizing value to its diagonal elements (in our experiments, 0.001 was usually sufficient), and (3) use inverse mapping T^{-1} to transform the regularized V-matrix to its original basis.

5.2 Reproducing Kernel Hilbert space

We are looking for solutions of our inference problems in the set of functions $f(x, \alpha), \alpha \in \Lambda$ that belong to Reproducing Kernel Hilbert Space associated with kernel $K(x, x')$, where $K(x, x')$ is a continuous positive semi-definite function of variables $x, x' \in X \subset R^n$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j) c_i c_j \geq 0 \tag{57}$$

for any $\{x_1, \dots, x_n\}$ and $\{c_1, \dots, c_n\}$. Consider linear operator

$$Af = \int_a^b K(x, s) f(s)\, ds \tag{58}$$

mapping elements $f(s)$ into elements $Af(x)$ in space H.

In 1909, Mercer showed that, for any continuous positive semi-definite kernel, there exists an orthonormal basis $e_i(x)$ consisting of eigenfunctions of $K(x, x')$ of operator (58), and the corresponding sequence of nonnegative eigenvalues λ_i such that kernel $K(x, x')$ has the representation

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \tag{59}$$

where the convergence of the sequence is absolute and uniform.

We say that the set Φ of functions $f(x)$ belongs to Reproducing Kernel Hilbert Space (RKHS) associated with kernel $K(x, x')$ if the inner product $< f_1, f_2 >$ between functions f_1, and f_2 of this set is such that for any function $f(x) \in \Phi$ the equality

$$f(x') =< K(x, x'), f(x) > \tag{60}$$

[9] Recall that almost all the points in a high-dimensional ball belong to an area where all the points are close to the surface of the ball.

holds true. That is, the inner product of functions from Φ with kernel $K(x, x')$ (where variable x' is fixed) has the reproducing property.

Consider the parametric set Φ of functions

$$f_c(x) = \sum_{i=1}^{\infty} c_i e_i(x), \quad c = (c_1, c_2, \ldots) \in R^{\infty}. \tag{61}$$

According to representation (61), kernel $K(x, x')$ as a function of variable x belongs to set Φ (the values $\lambda_i \phi_i(x')$ can be considered as parameters c_i of expansion.)

In order to define Reproducing Kernel Hilbert Space for set (61), we introduce the following inner product between two functions $f_b(x)$, $f_d(x)$, defined by parameters $b = (b_1, b_2, \ldots)$ and $d = (d_1, d_2, \ldots)$ in (61):

$$\langle f_a(x), f_b(x) \rangle = \sum_{i=1}^{\infty} \frac{b_i d_i}{\lambda_i}. \tag{62}$$

It is easy to check that, for such inner product, reproducing property (60) of functions from Φ holds true and the square of the norm of function $f_b(x) \in \Phi$ is equal to

$$||f_b(x)||^2 =< f_b(x), f_b(x) >= \sum_{i=1}^{\infty} \frac{b_i^2}{\lambda_i}. \tag{63}$$

5.2.1 Properties of RKHS

The following three properties of functions from RKHS make them useful for function estimation problems in high-dimensional spaces:

1. Functions from RKHS with bounded square of norms

$$\sum_{i=1}^{\infty} \frac{b_i^2}{\lambda_i} \leq C \tag{64}$$

   belong to a compact set and therefore the square of the norm of function can be used as regularization functional (see Lemma in Sect. 4.2).

2. The function that minimizes empirical loss functional (51) in RKHS, along with its parametric representation (61) in *infinite*-dimensional space of parameters c, has another parametric representation in ℓ-dimensional space $\alpha = (\alpha_1, \ldots, \alpha_\ell) \in R^\ell$, where ℓ is the number of observations:

$$f(x, \alpha) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x). \tag{65}$$

   (This fact constitutes the content of the so-called *Representer Theorem*).

3. The square of the norm of the chosen function, along with representation (63), has the representation of the form

$$||f(x, \alpha)||^2 =< f(x, \alpha), f(x, \alpha) >= \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j). \tag{66}$$

   This representation of the function in RKHS is used to solve the inference problem in high-dimensional space.

5.2.2 Properties of Kernels

Kernels $K(x, x')$ (also called *Mercer kernels*) have the following properties:

(1) Linear combination of kernels $K_1(x, x')$ and $K_2(x, x')$ with non-negative weights is the kernel

$$K(x, x') = \alpha_1 K_1(x, x') + \alpha_2 K_2(x, x'), \quad \alpha_1 \geq 0, \alpha_2 \geq 0. \tag{67}$$

(2) Product of the kernels $K_1(x, x')$ and $K_2(x, x')$ is the kernel

$$K(x, x') = K_1(x, x')K_2(x, x'). \tag{68}$$

In particular, the product of kernels $K(x^k, x'^k)$ defined on coordinates x^k of vectors $x = (x^1, \ldots, x^m)$ is a multiplicative kernel in m-dimensional vector space $x \in R^m$:

$$K(x, x') = \prod_{s=1}^{m} K_s(x^s, x'^s). \tag{69}$$

(3) Normalized kernel is the kernel

$$K_*(x, x') = \frac{K(x, x')}{\sqrt{K(x, x)K(x', x')}}. \tag{70}$$

5.2.3 Examples of Mercer Kernels

(1) **Gaussian kernel** in $x \in R^1$ has the form

$$K(x, x') = \exp\{-\delta(x - x')^2\}, \quad x, x' \in R^1, \tag{71}$$

where $\delta > 0$ is a free parameter of the kernel. In m-dimensional space $x \in R^m$, Gaussian kernel has the form (69)

$$K(x, x') = \prod_{k=1}^{m} \exp\{-\delta(x^k - x'^k)^2\} = \exp\{-\delta|x - x'|^2\}, \quad x, x' \in R^n. \tag{72}$$

(2) **INK-spline kernel** (*spline with infinite numbers of knots*). INK-spline kernel of order d was introduced in Vapnik (1995). For $x \in [0, c]$, it has the form

$$K(x, x') = \int_0^c (x - t)_+^d (x' - t)_+^d dt = \sum_{r=0}^{d} \frac{C_d^r}{2d - r + 1} [\min(x, x')]^{2d-r+1} |x - x'|^r, \tag{73}$$

where we have denoted $(z)_+ = \max(z, 0)$. In particular, INK-spline kernel of order 0 has the form

$$K_0(x, x') = \min(x, x'). \tag{74}$$

This INK-spline is used to approximate piecewise continuous functions. INK-spline kernel of order 1 has form

$$K_1(x, x') = \frac{1}{2}|x - x'| \min(x, x')^2 + \frac{\min(x, x')^3}{3}. \tag{75}$$

This INK-spline is used to approximates smooth functions. Its properties are similar to those of cubic splines that are used in the classical theory of approximations.

The described two types of kernels reflect different ideas of function approximation: local approximation of the desired function (Gaussian kernel) and global approximation of the desired function (INK-spline kernels).

5.3 Basic solution of inference problems

We solve our inference problem using functions

$$f(x) = \psi(x) + c, \tag{76}$$

where $\psi(x)$ belongs to RKHS of kernel $K(x, x')$ and $c \in R^1$ is the bias.

In order to find the solution of our integral equation, we minimize the regularization functional (48). For the solution in RKHS with bias, we use the representation (see Sect. 5.2.1)

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x) + c, \tag{77}$$

where $x_i, \ i = 1, \ldots, \ell$ are vectors from the training set. For the regularization term, we use the square of the norm of the function $\phi(x)$ in RKHS described in representation (66) as

$$W(f) = ||f(x)||^2 = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j). \tag{78}$$

In order to solve our ill-posed problem of inference, we minimize the functional (48). Taking into account explicit expression of distance (51) and regularizer (66), we obtain

$$\begin{aligned} R(\alpha) = & \sum_{i,j=1}^{\ell} f(x_i) f(x_j) V(i, j) + \sum_{i,j=1}^{\ell} y_i y_j V(i, j) \\ & -2 \sum_{i,j=1}^{\ell} f(x_i) y_j V(i, j) + \sum_{i,j=1}^{\ell} y_i y_j V(i, j) + \gamma_\ell \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j) \end{aligned} \tag{79}$$

in the set of functions (77).

Matrix-vector notations. We use the following notations:

1. $(\ell \times 1)$-dimensional matrix A of elements α:

$$A = (\alpha_1, \ldots, \alpha_\ell)^T;$$

2. ℓ-dimensional vector-function $\mathcal{K}(x)$:

$$\mathcal{K}(x) = (K(x_1, x), \ldots, K(x_\ell, x))^T;$$

3. $(\ell \times \ell)$-dimensional matrix K with elements $K(x_i, x_j)$:

$$K = ||K(x_i, x_j)||, \quad i, j = 1, \ldots, \ell;$$

4. $(\ell \times \ell)$-dimensional matrix V of elements $V(i, j)$:

$$V = ||V(i, j)||, \quad i, j = 1, \ldots, \ell;$$

5. ℓ-dimensional vector of elements y_i of training set:

$$Y = (y_1, \ldots, y_\ell)^T;$$

6. ℓ-dimensional vector of 1:

$$1_\ell = (1, \ldots, 1)^T;$$

Using these notations, we rewrite the functional (79) in the form

$$R(A) = (KA + c1_\ell)^T V(KA + c1_\ell) - 2(KA + c1_\ell)^T VY + \gamma_\ell A^T KA + Y^T VY. \quad (80)$$

5.4 Closed-form solution of minimization problem

In order to find the solution

$$f(x) = A^T \mathcal{K}(x) + c \quad (81)$$

of our inference problem in the set of functions that belong to RKHS of kernel $K(x, x')$, we have to minimize functional (80) over parameters A and c (since the term $Y^T VY$ does not depend on the parameters A, c, we drop it). The necessary conditions of minimum are

$$\begin{cases} \dfrac{\partial R(A,c)}{\partial A} \Longrightarrow VKA + cV1_\ell - VY + \gamma_\ell A = 0, \\ \dfrac{\partial R(A,c)}{\partial c} \Longrightarrow 1_\ell^T VKA + c1_\ell^T V1_\ell - 1_\ell^T VY = 0. \end{cases} \quad (82)$$

From the first equation of (82) we obtain

$$(VK + \gamma I)A = VY - cV1_\ell,$$

so

$$A = (VK + \gamma I)^{-1}(VY - cV1_\ell). \quad (83)$$

We then compute vectors

$$A_b = (VK + \gamma I)^{-1} VY, \quad A_c = (VK + \gamma I)^{-1} V1_\ell. \quad (84)$$

According to (83), the desired vector A has the form

$$A = A_b - cA_c. \quad (85)$$

From second equation of (82) and (85), we obtain the equation to define c:

$$1_\ell^T VK(A_b - cA_c) + c1_\ell^T V1_\ell - 1_\ell^T VY = 0,$$

where A_b and A_c are defined by (84). We have

$$\left[1_\ell^T VKA_b - 1_\ell^t V1_\ell\right] c \;=\; \left[1_\ell^T VKA_c - 1_\ell^T VY\right]$$

and the value of bias c is thus

$$c = \frac{\left[1_\ell^T VKA_b - 1_\ell^T VY\right]}{\left[1_\ell^T VKA_c - 1_\ell^T V1_\ell\right]}. \quad (86)$$

Putting c in (85), we obtain the desired parameters A. The desired function f has the form (81).

Remark There exists one more prior knowledge about the conditional probability function: this function is *non-negative and bounded by 1*. We can easily construct a function that is non-negative and bounded at all the points of observation. In order to enforce such a solution, it is sufficient to minimize the functional (80) subject to additional constraints

$$0_\ell \leq KA + c1_\ell \leq 1_\ell. \tag{87}$$

This inequality is defined coordinate-wise where 0_ℓ is ℓ-dimensional vectors of zeros and 1_ℓ is ℓ-dimensional vector of ones. Unfortunately, one cannot obtain closed-form solution of the optimization problem with these constraints.

5.5 Dual estimate of conditional probability

Let $y \in \{0, 1\}$. The estimate of the conditional probability $P(y = 1|x)$ has the form

$$P(y = 1|x) = A_\ell^T \mathcal{K}(x) + c. \tag{88}$$

Similarly, the conditional probability for class $y = 0$ is

$$P(y = 0|x) = A_\ell^{*T} \mathcal{K}(x) + c^*, \tag{89}$$

where parameters A^* and c^* are obtained using formulas (85), (86), with binary vector $Y_0 = 1_\ell - Y$ (the vector-indicator of class $y = 0$).

According to the definition, for these conditional probability functions, the equality

$$P(y = 1|x) + P(y = 0|x) = 1 \tag{90}$$

holds for all $x \in X$.

When one estimates the conditional probability function, this equality forms a prior knowledge.

Below we estimate two conditional probabilities for which equality (90) holds true only for vectors x of the training set. We replace equality (90) with the following equality:

$$(KA + c1_\ell) + (KA^* + c^*1_\ell) = 1_\ell. \tag{91}$$

Taking into account (91), we estimate parameters A, A^*, c, c^* of Eqs. (88), (89). In order to do this, we minimize the functional which is the sum of two functionals of type (80) for estimation of parameters of conditional probabilities (88), (89):

$$\begin{aligned} R(A, A^*, c, c^*) = {} & (KA + c1_\ell)^T V(KA + c1_\ell) - 2(KA + c1_\ell)^T VY + 2\gamma A^T KA \\ & + (KA^* + c^*1_\ell)^T V(KA^* + c^*1_\ell) - 2(KA^* + c^*1_\ell)^T VY^* \\ & + 2\gamma_* A^{*T} KA^*, \end{aligned} \tag{92}$$

where $\gamma, \gamma_* > 0$ and $Y^* = 1_\ell - Y$. Using the property $KK^+K = K$ of pseudo-inverse K^+ of matrix K and expression [obtained from (91)]

$$KA^* = -(KA + (c + c^* - 1)1_\ell),$$

we can rewrite the last term of functional (92) as

$$A^{*T} KA^* = A^{*T} AKK^+KA^* = (KA + (c + c^* - 1)1_\ell)^T K^+(KA + (c + c^* - 1)1_\ell)$$

and the functional (92) itself as

$$\begin{aligned} R(A,c,c^*) = {} & (KA + c1_\ell)^T V(KA + c1_\ell) - 2(KA + c1_\ell)^T VY + 2\gamma A^T KA \\ & + (KA + (c-1)1_\ell)^T V(KA + (c-1)1_\ell) + 2(KA + (c-1)1_\ell)^T V(1_\ell - Y) \\ & + 2\gamma_*(KA + (c+c^*-1)1_\ell)^T K^+(KA + (c+c^*-1)1_\ell). \end{aligned} \tag{93}$$

In order to find the solution of this optimization problem, we minimize functional (93) with respect to A, c, and c^*. The necessary conditions of minimum are

$$\begin{aligned} & (VK + \gamma I + \gamma_* K^+K)A + (V + \gamma_* K^+)1_\ell c + \gamma_* K^+ 1_\ell c^* - (VY + \gamma_* K^+ 1_\ell) = 0 \\ & 1_\ell^T (V + \gamma_* K^+)KA + 1_\ell^T (V + \gamma_* K^+)1_\ell c + \gamma_* 1_\ell^T K^+ 1_\ell c^* - 1_\ell^T (VY + \gamma_* K^+ 1_\ell) = 0 \\ & 1_\ell^T K^+ KA + 1_\ell^T K^+ 1_\ell c + 1_\ell^T K^+ 1_\ell c^* - 1_\ell^T K^+ 1_\ell = 0 \end{aligned} \tag{94}$$

In order to simplify the expressions, we introduce the following notations:

$$\begin{aligned} & \mathcal{L} = (VK + \gamma I + \gamma_* K^+K), \quad \mathcal{M} = (V + \gamma_* K^+), \\ & \mathcal{N} = VY + \gamma_* K^+ 1_\ell, \quad \mathcal{P} = \frac{1_\ell^T K^+ K}{1_\ell^T K^+ 1_\ell}. \end{aligned} \tag{95}$$

From (94) and (95), we obtain

$$\mathcal{L}A = \mathcal{N} - (\mathcal{M}1_\ell)\, c - \gamma_* \left(K^+ 1_\ell\right) c^*.$$

The equivalent expression is

$$A = \mathcal{L}^{-1}\mathcal{N} - c\mathcal{L}^{-1}(\mathcal{M}1_\ell) - \gamma_* c^* \mathcal{L}^{-1}\left(K^+ 1_\ell\right). \tag{96}$$

We then compute vectors

$$A_V^* = \mathcal{L}^{-1}\mathcal{N}, \quad A_c^* = \mathcal{L}^{-1}(\mathcal{M}1_\ell), \quad A_{c^*} = \mathcal{L}^{-1}\left(K^+ 1_\ell\right). \tag{97}$$

(these vectors can be computed in one run) and define A as

$$A = A_V^* - cA_c^* - \gamma_* c^* A_{c^*}. \tag{98}$$

Using expression (98) for A in (97) and using the notations (95), we obtain

$$\begin{array}{lll} c\left[1_\ell^T \mathcal{M} 1_\ell - 1_\ell^T \mathcal{M} K A_c\right] & + c^*\gamma_* \left[1_\ell^T K^+ 1_\ell - 1_\ell^T \mathcal{M} K A_{c^*}\right] & = \left[1_\ell^T \mathcal{N} - 1_\ell^T \mathcal{M} K A_b\right] \\ c\left[1 - \mathcal{P} A_c\right] & + c_* \left[1 - \gamma_* \mathcal{P} A_{c^*}\right] & = \left[1 - \mathcal{P} A_b\right] \end{array} \tag{99}$$

From equations (99), we find parameters c, c^* that define vector A in (98) and the desired conditional probability (88).

Note that dual estimate takes into account how well the matrix K is conditioned: the solution of optimization problem uses expressions K^+K and K^+1_ℓ [see formulas (96), (97)].

Remark Since condition (91) holds true for the dual estimate of conditional probability, in order to satisfy inequalities (87), it is sufficient to satisfy the inequality

$$KA + c1_\ell \geq 0_\ell. \tag{100}$$

Therefore, in order to obtain the dual estimate of conditional probability using available general prior knowledge, one has to minimize the functional (93) subject to constraint (100).

Part Two: Intelligence-driven learning: learning using statistical invariants

In this section, we introduce a new paradigm of learning called *Learning Using Statistical Invariants* (LUSI) which is different from the classical data-driven paradigm.

The new learning paradigm considers a model of Teacher–Student interaction. In this model, the Teacher helps the Student to construct statistical invariants that exist in the problem. While selecting the approximation of conditional probability, the Student chooses a function that preserves these invariants.

6 Strong and weak modes of convergence

The idea of including invariants into a learning scheme is based on the mathematical structure of Hilbert space. In Hilbert space, relations between two functions ($f_1(x)$ and $f_2(x)$ have two numerical characteristics:

1. The distance between functions
$$\rho(f_1, f_2) = ||f_i(x) - f_2(x)||$$
that is defined by the metric of the space L_2 and
2. The inner product between functions
$$R(f_1, f_2) = (f_1(x), f_2(x))$$
that has to satisfy the corresponding requirements.

The existence of two different numerical characteristics implies two different modes of convergence of the sequence of functions from $f_\ell(x) \in L_2$ to the desired function $f_0(x)$:

1. The strong mode of convergence (convergence in metrics)
$$\lim_{\ell \to \infty} ||f_\ell(x) - f_0(x)|| = 0 \quad \forall x$$
2. The weak mode of convergence (convergence in inner products)
$$\lim_{\ell \to \infty} (f_\ell(x) - f_0(x), \psi(x)) = 0, \quad \forall \psi(x) \in L_2$$
(note that convergence has to take place for *all* functions $\psi(x) \in L_2$).

It is known that the strong mode of convergence implies the weak one. Generally speaking, the reverse is not true.

In the first part of the paper, we showed that in classical (data-driven) paradigm, we can estimate the conditional probability function solving the ill-posed problem of the approximatively defined Fredholm equation (30), obtaining the V-matrix estimate (see Sect. 5.4).

In the second part of the paper, we consider new learning opportunities using interaction with Teacher. The goal of that interaction is to include mechanisms of both weak and strong convergence in learning algorithms.

Here is the essence of the new mechanism. According to the definition, weak convergence has to take into account *all functions* $\psi(x) \in L_2$. The role of Teacher in our model is to replace this *infinite* set of functions with a specially selected finite set of *predicate*-functions $\mathcal{P} = \{\psi_1(x), \dots, \psi_m(x)\}$ that can describe property of the desired conditional probability function and restrict the scope of weak convergence only to the set of predicate functions $\mathcal{P}$.

6.1 Method of learning using statistical invariants

Let us describe a method of extracting specific intelligent information from data by preserving statistical invariants.

Estimation of conditional probability function. Suppose that Teacher has chosen m predicate functions

$$\psi_1(x), \ldots, \psi_m(x). \tag{101}$$

The Teacher believes that these functions describe important (integral) properties of desired conditional probability function defined by the equalities (invariants)

$$\int \psi_s(x)P(y=1|x)dP(x) = \int \psi_s(x)dP(y=1,x) = a_s, \quad s=1,\ldots,m, \tag{102}$$

where a_s is the expected value of $\psi_s(x)$ with respect to measure $P(y=1,x)$.

Suppose now that values $a_1,\ldots,a_m$ are known. Then we can formulate the following two-stage procedure of estimating the desired function [solving Fredholm equation (30)]:

1. Given pairs $(\psi_k(x), a_k),\ k=1,\ldots,m$ (predicates and their expectation for desired condition probability function), find the set of conditional probability functions $\mathcal{F} = \{P(y=1|x)\}$ satisfying equalities (102) (*preserving the invariants*).
2. Select, in the set of functions $\mathcal{F}$ satisfying invariants (102), the function that is the solution of our estimation problem (the function of form (88) with parameters that minimize (80) (or (92))).

In other words, the idea is to use data to look for approximation $P_\ell(y=1|x)$ in the subset of functions that preserve m invariants (102) formulated by the Teacher's predicates.

In our model, the Teacher suggests only functions (101) and does not provide values $a_s,\ s=1,\ldots,m$. However, for any function $\psi_s(x)$, using training data $(x_i,y_i),\ i=1,\ldots,\ell$, we can estimate the corresponding value a_s, by obtaining the following approximation for invariants (102):

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\psi_s(x_i)P_\ell(y=1|x_i) \approx a_s \approx \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\psi_s(x_i), \quad s=1,\ldots,m. \tag{103}$$

Remark Expression (103) defines equality only approximately. When one adds invariants sequentially, one can take this fact into account. Suppose that we have constructed the following approximation using m predicates:

$$P_\ell^m(y=1|x) = \sum_{i=1}^{\ell} A^m\mathcal{K}(x) + c_m \tag{104}$$

In order to chose the next predicate $\psi_{m+1}(x)$, we consider the value

$$\mathcal{T} = \frac{\left|\sum_{i=1}^{\ell}\psi_{m+1}(x_i)P_\ell^m(y=1|x_i) - \sum_{i=1}^{\ell} y_i\psi_{m+1}(x_i)\right|}{\sum_{i=1}^{\ell} y_i\psi_{m+1}(x_i)} \tag{105}$$

If $\mathcal{T} \geq \delta$ (where δ is a small positive threshold), we consider the new invariant defined by predicate $\psi_{m+1}(x)$. Otherwise, we treat this approximation as equality and do not add this invariant.

To use vector-matrix notations, we define the set of ℓ-dimensional vectors

$$\Phi_s = (\psi_s(x_1),\ldots,\psi_s(x_\ell))^T, \quad s=1,\ldots,m+1. \tag{106}$$

In these notations, the expression (103) with m predicates chosen for invariant as

$$\Phi_s^T KA + c\Phi_s^T 1_\ell = \Phi_s^T Y, \quad s = 1, \dots, m. \tag{107}$$

Consider a new predicate $\psi_{m+1}(x)$, the corresponding vector Φ_{m+1} and expression (105) as

$$\mathcal{T} = \frac{|\Phi_{m+1}^T KA^m + c_m \Phi_{m+1}^T 1_\ell - \Phi_{m+1}^T Y|}{Y^T \Phi_{m+1}} \tag{108}$$

We treat the expression (107) as equality[10] if $\mathcal{T} \leq \delta$.

> The method of LUSI suggests to use an approximation function of the form (88), where parameters A and c are the solution of the following quadratic optimization problem: minimize functional (80) (or (92) for dual estimate) subject to constraints (107).

Rules for t-class pattern recognition problem. Consider t-class pattern recognition problem. Our goal is to estimate n conditional probability functions $P(y = s|x),\ s = 1, \dots, t$ in order to combine them into the rule

$$r(x) = \operatorname{argmax}(P(y = 1|x), \dots, P(y = s|x), \dots, P(y = t|x)).$$

In order to define appropriate invariants for the conditional probability function $P(y = k|x)$, consider the following two-class classification problem: we define the elements of class $y = s$ as class $y^* = 1$ and define elements of all other classes $y \neq s$ as class $y^* = 0$. One obtains parameters A_V^s and c^s for estimating the corresponding conditional probability by using formulas described above with the correponding vector $Y_s = (y_1^*, \dots, y_\ell^*)^T$.

6.2 Closed-form solution of intelligence-driven learning

The optimization problem of minimizing functional (80) subject to the constraints (105) has a closed-form solution. Consider Lagrangian for this problem:

$$\begin{aligned} L(A, c, \mu) = {} & (KA + c1_\ell)^T V(KA + c1_\ell) - 2(KA + c1_\ell)^T VY \\ & + \gamma_\ell A^T KA + 2\sum_{s=1}^{m} \mu_s (A^T K\Phi_s + c1_\ell^T \Phi_s - Y^T \Phi_s). \end{aligned} \tag{109}$$

As before, we obtain the following Lagrangian conditions:

$$\begin{aligned} \frac{\partial L(A, c, \mu)}{\partial A} & \Longrightarrow VKA + \gamma A + cV1_\ell - VY + \sum_{s=1}^{m} \mu_s \Phi_s = 0 \\ \frac{\partial L(A, c, \mu)}{\partial c} & \Longrightarrow 1_\ell^T VKA + 1_\ell^T V1_\ell c - 1_\ell^T VY + \sum_{s=1}^{m} \mu_s 1_\ell^T \Phi_s = 0 \\ \frac{\partial L(A, c, \mu)}{\partial \mu_k} & \Longrightarrow A^T K\Phi_k + c1_\ell^T \Phi_k - Y^T \Phi_k = 0, \quad k = 1, \dots, m \end{aligned} \tag{110}$$

[10] Instead of (107), it is more accurate to consider the inequality constraints

$$-\delta\mathcal{T} \leq \Phi_s^T KA + c\Phi_s^T 1_\ell - \Phi_s^T Y \leq \delta\mathcal{T}, \quad s = 1, \dots, m.$$

In this case, one has to solve the following quadratic optimization problem: to minimize (80) (or (92)) subject to these inequality constraints.

From the first line of (110), we obtain the expression

$$(VK + \gamma I)A = VY - cV1_\ell - \sum_{k=1}^{m} \mu_k \Phi_k \tag{111}$$

and the expression

$$A = (VK + \gamma I)^{-1}(VY - cV1_\ell - \sum_{s=1}^{m} \mu_s \Phi_s). \tag{112}$$

We compute $(m + 2)$ vectors (this can be done simultaneously in one run)

$$A_V = (VK + \gamma I)^{-1} VY, \tag{113}$$

$$A_c = (VK + \gamma I)^{-1} V1_\ell \tag{114}$$

$$A_s = (VK + \gamma I)^{-1} \Phi_s, \quad s = 1, \dots, n. \tag{115}$$

The desired vector A has the expression

$$A = A_V - cA_c - \sum_{s=1}^{m} \mu_s A_s. \tag{116}$$

Putting expression (116) back into the last two lines of (110), we note that, in order to find coefficient c and m coefficients μ_s of expansion (116), we have to solve the following system of $m + 1$ linear equations:

$$c[1_\ell^T VKA_c - 1_\ell^T V1_\ell] + \sum_{s=1}^{m} \mu_s [1_\ell^T VKA_s - 1_\ell^T \Phi_s] = [1_\ell^T VKA_V - 1_\ell^T VY] \tag{117}$$

$$c[A_c^T K\Phi_k - 1_\ell^T \Phi_k] + \sum_{s=1}^{m} \mu_s A_s^T K\Phi_k = [A_V^T K\Phi_k - Y^T \Phi_k], \quad k = 1, \dots, m. \tag{118}$$

Using estimated vector A and bias c, we obtain the desired function (88).

Summary. In this section, we have obtained a method for estimating conditional probability functions using invariants. The estimate has the form

$$f(x) = A^T \mathcal{K}(x) + c,$$

where the vector of coefficients of expansion A has the structure

$$A = (A_V - cA_c) - \left(\sum_{s=1}^{m} \mu_s A_s\right). \tag{119}$$

Vectors A_V, A_c and A_s, $s = 1, \dots, m$ are obtained using formulas (113), (114), (115), and coefficients c and μ_s of composition (119) are the solutions of linear equations (117), (118). We call this algorithm the vSVM algorithm with m invariants vSVM&I$_m$.

When estimating conditional probability function, one can also take into account additional prior knowledge (87). To find parameters A_V and c of approximation that take into account this information, one has to solve the following quadratic optimization problem: minimize the functional (80) subject to m equality constraints (107) and ℓ inequality constraints (87).

6.3 Dual intelligence-driven estimate of conditional probability

Dual intelligence-driven estimate of conditional probability requires minimization of functional (93) subject to equality type constraints (105).

In order to solve this optimization problem, we define the Lagrangian

$$\begin{aligned} L(A, c, c^*, \mu) = {} & (KA + c1_\ell)^T V(KA + c1_\ell) - 2(KA + c1_\ell)^T VY + 2\gamma A^T KA \\ & + (KA + (c-1)1_\ell)^T V(KA + (c-1)1_\ell) + 2(KA + (c-1)1_\ell)^T \\ & \times V(1_\ell - Y) + 2\gamma_*(KA + (c + c^* - 1)1_\ell)^T K^+(KA + (c + c^* - 1)1_\ell) \\ & + \sum_{s=0}^{m} \mu_s (A^T K \Phi_s + c1_\ell^T \Phi_s - Y^T \Phi_s), \end{aligned} \tag{120}$$

where μ_s are Lagrange multipliers. In order to find the solution of this optimization problem, we minimize this functional with respect to A, c, and c^*. The necessary conditions of minimum using notations (95) are

$$\begin{aligned} \frac{\partial R}{\partial A} & \Longrightarrow \mathcal{L}A + \mathcal{M}1_\ell c + \gamma_* K^+ 1_\ell c^* - VY - \gamma_* K^+ 1_\ell + \sum_{s=1}^{m} \mu_s \Phi_s = 0 \\ \frac{\partial R}{\partial c} & \Longrightarrow 1_\ell^T \mathcal{M} KA + 1_\ell^T \mathcal{M} 1_\ell c + \gamma_* 1_\ell^T K^+ 1_\ell c^* - 1_\ell^T VY - \gamma_* K^+ 1_\ell + \sum_{s=1}^{m} \mu_s \Phi_s = 0 \\ \frac{\partial R}{\partial c^*} & \Longrightarrow 1_\ell^T K^+ KA + 1_\ell^T K^+ 1_\ell c + 1_\ell^T K^+ 1_\ell c^* - 1_\ell^T K^+ 1_\ell = 0 \\ \frac{\partial R}{\partial \mu_s} & \Longrightarrow A^T K \Phi_s + c1_\ell^T \Phi_s - Y^T \Phi_s = 0 \end{aligned} \tag{121}$$

From the first equality of (121) we obtain

$$A = \mathcal{L}^{-1}\mathcal{N} - c\mathcal{L}^{-1}(\mathcal{M}1_\ell) - \gamma_* c^* \mathcal{L}^{-1}\left(K^+ 1_\ell\right) - \mathcal{L}^{-1} \sum_{s=1}^{m} \mu_s \Phi_s. \tag{122}$$

We then compute vectors

$$A_b = \mathcal{L}^{-1}\mathcal{N}, \quad A_c = \mathcal{L}^{-1}(\mathcal{M}1_\ell), \quad A_{c^*} = \mathcal{L}^{-1}\left(K^+ 1_\ell\right), \quad A_s = \mathcal{L}^{-1}\Phi_s \tag{123}$$

(these vectors can be computed in one run) and define A as

$$A = A_b - cA_c - \gamma_* c^* A_{c^*} - \sum_{s=1}^{m} \mu_s A_s. \tag{124}$$

Putting expression for A in the last three other lines of (121), we obtain the following linear system of equations for computing the unknown parameters c, c^*, and μ_s.

$$\begin{array}{llll} c1_\ell^T \mathcal{M}[1_\ell - KA_c] & +c^* \gamma_* 1_\ell^T \left[K^+ 1_\ell - \mathcal{M}KA_{c^*}\right] & -\displaystyle\sum_{s=1}^{m} \mu_s 1_\ell^T [\mathcal{M}KA_s - \Phi_s] & = 1_\ell^T [\mathcal{N} - \mathcal{M}KA_b] \\ c[1 - \mathcal{P}A_c] & +c_* \left[1 - \gamma_* \mathcal{P}A_{c^*}\right] & -\displaystyle\sum_{s=1}^{m} \mu_s \mathcal{P}A_s & = [1 - \mathcal{P}A_b] \\ c\Phi_s^T [1_\ell - KA_c] & +c_* \left[-\gamma_* \Phi_s^T KA_{c_*}\right] & -\displaystyle\sum_{s=1}^{m} \mu_s \Phi_s^T KA_s & = \Phi_s^T [Y - KA_b] \end{array}$$

6.4 LUSI methods: illustrations and remarks

In this section, we compare different methods of conditional probability estimate for two-class pattern recognition problems. First, we consider four estimation methods on one-dimensional examples; then we move to multi-dimensional examples. We apply the following four methods:

1. SVM estimate
2. vSVM estimate.[11]
3. SVM&I$_n$ estimate (SVM with n invariants). In this section, invariants are defined by the simple predicate functions[12]

$$\psi_0(x) = 1 \quad \text{and} \quad \psi_1(x) = x.$$

4 vSVM& I$_n$ estimate (vSVM with n invariants).

Invariants (107) with predicate $\psi_0(x)$ define the set of conditional probability estimates for which the frequency of elements of class $y = 1$ is the same as it was observed on the training data. Invariants (107) with predicate $\psi_1(x)$ define the set of conditional probability estimates for which the center of mass of elements x belonging to class $y = 1$ is the same as it was observed on the training data. In this section, we consider experiments with simple invariants, i.e, with invariants obtained based on functions $\psi_0(x)$ and $\psi_1(x)$ (preserving frequencies and means, respectively).

1. The experiments in one-dimensional space are presented in Fig. 1 (vertical axes denote conditional probability, while horizontal axes denote values of x). The first four rows of that figure show results of 12 experiments, where each row corresponds to one of four methods described above, and each column corresponds to one of three sizes of training data (48, 96, 192). In all the images, the actual conditional probability function $P(y = 1|x)$ is shown in blue color, while the obtained estimate $P_\ell(y = 1|x)$ is shown in black color. The images also show training data: representatives of class $y = 1$ are shown in red color, while representatives of class $y = 0$ are shown in green color. In order to demonstrate the robustness of the proposed methods, in all the experiments we use twice as many representatives of the class $y = 0$ than representatives of class $y = 1$: $(48 = 16 + 32; 96 = 32 + 64; 192 = 64 + 128)$. All the estimates were obtained using the INK-spline kernel (75).

The first row of Fig. 1 shows results of SVM estimates, (square loss SVM); the second row shows results of vSVM estimates; the third row shows results of SVM&I$_2$ estimates; the fourth row shows results of vSVM&I$_2$ estimates. The last two rows of Fig. 1 show results obtained by modified vSVM&I$_2$ and modified vSVM&I$_2$, which are described in the subdivision 4 after Table 1.

From Fig. 1, one can see that vSVM estimates are consistently better than SVM estimates and that adding the invariants consistently improves the quality of approximations. One can also see that approximations obtained using vSVM provide smoother functions.

2. Using approximation $P_\ell(y = 1|x)$ of the conditional probability function, we obtain the classification rule

$$r_\ell(x) = \theta\left(P_\ell(y = 1|x) - \frac{1}{2}\right)$$

[11] Square-loss SVM method uses (identity) I-matrix instead of V-matrix (see (9), (10)). To implement LUSI approach for square loss SVM methods, one has to use formulas for V-matrix method and replace V-matrix with (identity) I-matrix.

[12] These functions define values of zeroth order and first order moments of conditional probability function $P(y = 1|x)$.

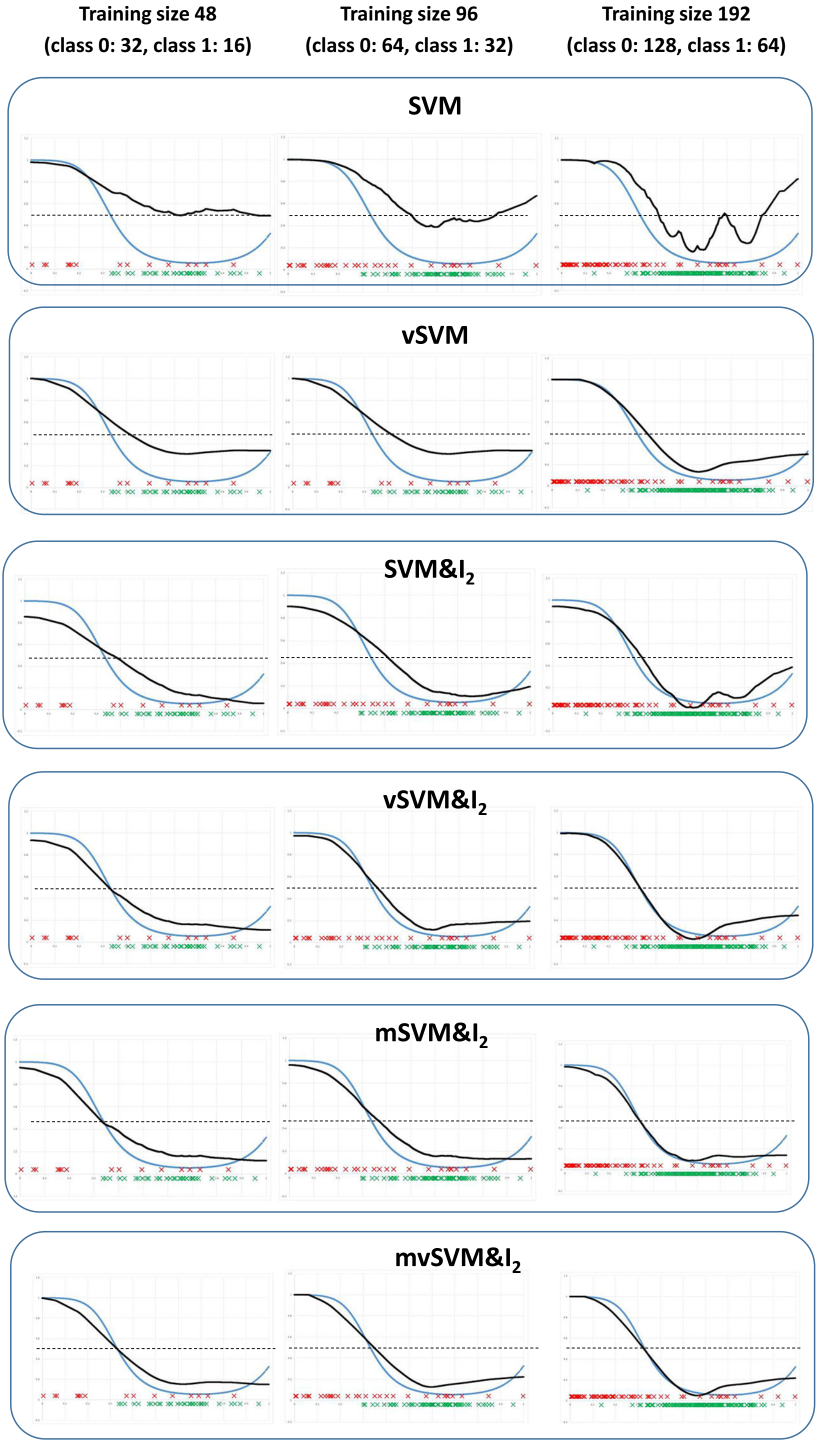

Fig. 1 Experiments with four estimates on three different sizes of training data

Table 1 Experiments with one-dimensional case

Training	SVM	vSVM	SVM&I$_2$	vSVM&I$_2$	mSVM&I$_2$	mvSVM&I$_2$
Distance to the true conditional probability in L_2 metric						
48	0.3756	0.2166	0.1432	0.1070	0.1064	0.0940
96	0.3212	0.1808	0.1207	0.0778	0.0950	0.0863
192	0.2273	0.1072	0.0689	0.0609	0.0461	0.0557
Error rate in %						
48	59.79	11.00	11.38	11.02	11.00	11.14
96	23.41	13.28	12.23	11.44	11.39	11.29
192	16.57	11.68	11.54	11.33	11.11	11.11
Relative error						
48	4.50	0.02	0.05	0.01	0.01	0.02
96	1.15	0.22	0.12	0.05	0.05	0.04
192	0.52	0.07	0.06	0.04	0.02	0.02

Error rate for Baysian rule is 10.87%

(Fig. 1 shows horizontal line $y = 1/2$. Geometrically, the rule classifies x as class $y = 1$ where curve $P_\ell(y = 1|x)$ is above this line; otherwise, it classifies x as class $y = 0$).

The quality of rule $r_\ell(x)$ is measured by the value of probability of error

$$P_{err}(r) = \int (y - r_\ell(x))dP(x, y).$$

When $P_\ell(y = 1|x)$ coincides with the actual conditional probability function $P(y = 1|x)$, the obtained rule $r_B(x)$ is called Bayesian. It gives the smallest value for the problem's error rate $P_{err}(r_B)$. For the problem shown in Fig. 1, Bayesian loss is $P_{err}(r_B) = 10.87\%$. Losses of rules $r_\ell(x)$ based on estimates of conditional probability are larger than that.

3. Numerical characteristics of accuracy of approximations are given in Table 1. It shows, for different sizes of training data, the following:

1. The distance between the estimate and the true function in L_2 metric

$$\rho_{L_2}(P, P_\ell) = \left(\int (P(y = 1|x) - P_\ell(y = 1|x))^2 d\mu(x)\right)^{1/2}.$$

2. The error rate $P_{err}(r_\ell)$ of rule $r(x) = \theta(P(y = 1|x) - 0.5)$ in percents, and
3. The values of relative (with respect to Bayesian rule) losses

$$\kappa_{err}(r_\ell) = \frac{P_{err}(r_\ell) - P_{err}(r_B)}{P_{err}(r_B)}.$$

4. From Fig. 1 and Table 1, one can see that, with increasing sophistication of conditional probability estimation methods, the obtained results are getting monotonically closer (in L_2 metric) to the true conditional probability function. However, this monotonic convergence does not always hold for the error rates provided by the constructed conditional probability functions. This is because our mathematical goal was to approximate the conditional probability function in L_2 metric (not the decision rule). A good estimate of conditional probability for a classification rule is the function that crosses the line $y = 0.5$ close to the point where this line is crossed by the true conditional probability; it does not have to be close to the true function in L_2 metric (see Fig. 1).

Estimation of conditional probability for constructing classification rule. We can take into account our ultimate goal of constructing the classification rule if we estimate the conditional probability more accurately in the area of x where $|P(y=1|x)-0.5|<\epsilon$ at the expense of accuracy of this estimate in the area x where $|P(y=1|x)-0.5|\gg\epsilon$. In order to do this, we note that square value of variance for a fixed x is

$$\sigma^2(x)=P(y=1|x)(1-P(y=1|x)).$$

It achieves the largest value 1/4 when $P(y=1|x)=1/2$ and monotonically decreases (up to zero) when deviation $|P(y=1|x)-0.5|$ increases. We can use this fact when estimating distance (50). Suppose we know the function

$$\phi(x)=\sigma^2(x)$$

(or some other monotonically increasing function of $\sigma^2(x)$, for instance, $\phi(\sigma^2(x)+\nu)$). Consider (50) in the equivalent form

$$\rho^2=\sum_{i,j=1}^{\ell}(y_i-f(x_i,\alpha))(y_j-f(x_j,\alpha))V(i,j)$$

and let us minimize the functional with the square of distance defined as

$$\rho^2=\sum_{i,j=1}^{\ell}(y_i-f(x_i,\alpha))(y_j-f(x_j,\alpha))\sigma(x_i)\sigma(x_j)V(i,j)$$

Using techniques of Sect. 5.1, we obtain modified V_M-matrix with the elements

$$V_M(i,j)=\sigma(x_i)\sigma(x_j)V(i,j)$$

instead of elements (54) (or (56)).

We can also construct modified I_M-matrix, which is a diagonal matrix with elements

$$I_M(j,j)=\sigma^2(x_j)$$

and $I(i,j)=0$ for $i\neq j$.

In reality, we do not know the function $P(y=1|x)$. However, we can use its estimate $P_\ell(y=1|x)$ to compute $\sigma(x_i)$. Therefore, in order to construct a special conditional probability function for subsequent creation of a decision rule, we can use the following two-stage procedure: in the first stage, we estimate (SVM&I$_n$ or vSVM&I$_n$) approximations of $P_\ell(y=1|x)$ and $\sigma^2(x)$, and, in the second stage, we obtain an estimate of the specialized conditional probability function (mSVM&I$_n$ or mvSVM&I$_n$) for the decision rules.

The last two rows of Fig. 1 and last two columns of Table 1 compare the rules obtained using SVM & I$_2$ estimates and vSVM&I$_2$ estimates with approximations obtained using $mSVM$ & I$_2$ and mvSVM&I$_2$ estimates (in both cases we used function $\phi(x)=\sigma^2(x)$). It is interesting to note that estimates based on modified weight-function $\phi(x)$ not only improve the error rates of corresponding classification rules but also provide better approximations of conditional probability functions in L_2 metric.

5. Our one-dimensional experiments demonstrate the following:

1. vSVM results are more accurate than SVM results.
2. Learning using statistical invariants can significantly improve performance. Both vSVM&I$_2$ and SVM &I$_2$ algorithms using 48 training examples achieve much better performance (which are very close to Bayesian) than just SVM or vSVM algorithms

Table 2 Experiments with multidimensional data

Data set	Training	Test	Features	SVM (%)	SVM&I$_{(n+1)}$ (%)
Diabetes	562	206	8	30.94	22.73
Bank marketing	445	4076	16	12.06	10.58
MAGIC	1005	18,015	10	19.03	15.10
Parkinsons	135	60	22	7.26	6.67
Sonar	160	48	60	12.48	12.40
Ionosphere	271	80	33	5.66	5.55
WPBC	134	60	33	25.48	23.02
WDBC	419	150	30	2.64	2.50

Table 3 Experiments with different sizes of training data

Diabetes			MAGIC		
Training	SVM (%)	SVM&I$_9$ (%)	Training	SVM (%)	SVM & I$_{11}$ (%)
71	32.42	27.52	242	20.51	17.35
151	29.97	24.56	491	20.93	15.91
304	31.35	23.78	955	18.89	15.19
612	30.43	23.30	1903	18.03	14.25

that use 192 examples. The result obtained based on mSVM&I$_2$ method are consistently better than the results obtained based on SVM&I$_2$ method.

3. The effect from adding invariants appears to be more significant than the effect of upgrading of SVM to vSVM.

The same conclusion can be derived from analysis of experiments for high-dimensional problems where we use the $(n + 1)$ invariants: one zeroth order moment and n first order moments, where n is dimensionality of the problem (see Table 2). In this table, for eight calibration datasets from Lichman (2013) (listed in the first column), we conducted 20 experiments. In each we create a random partition of the given dataset into training and test sets (their sizes are listed in the second and third column, respectively), and then apply two algorithms (SVM and SVM&I$_{(n+1)}$) for estimating conditional probability function. We then used tehse conditional probability functions to construct the classification rule. For all experiments we used RBF kernel (72). As the results show, SVM &I$_{(n+1)}$ is consistently better than SVM.

6. Table 3 compares results of SVM&I$_n$ with SVM for two datasets Diabetes and MAGIC using training data of different sizes. Experiments show that incorporation of invariants significantly improve SVM for all sizes of training sets. In order to achieve accurate results in these problems, SVM&I$_{(n+1)}$ requires significantly fewer training examples than SVM.

Remark When analyzing the experiments, it is important to remember that the Intelligent Teacher suggests to the Student "meaningful" functions for invariants instead of simple ones that we used here for illustrative purposes. In many of our experiments, the invariants were almost satisfied: the correcting parameters μ_s in (119) were close to zero. Some ideas about what form these "meaningful" invariants can take are presented in the next Sections.

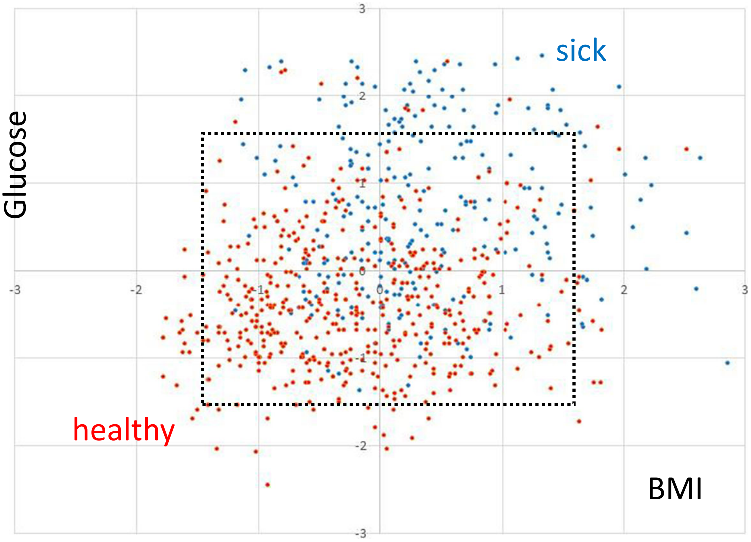

Fig. 2 Selected box $\mathcal{B}$ in the subspace of space X

6.5 Examples of invariants with non-linear predicate-functions

Example Table 2 shows that, using nine simple invariants (moments of zeroth and first order), SVM&I$_9$ achieved error rate 22.73% on "Diabetes". In order to obtain a better performance, we introduce an additional invariant constructed in some two-dimensional subspace of eight-dimensional space X of the problem (Fig. 2). We choose (using training data) the square box $\mathcal{B}$ (it is shown in Fig. 2) and consider the following predicate: $\psi_3(z) = 1$ if $x \in \mathcal{B}$ and $\psi_3(z) = 0$ otherwise. Our additional invariant is defined by this predicate $\psi_3(x)$. As a result, when using all 10 invariants, we decrease the error rate to 22.07%.

Note that function $\psi_3(x)$ was obtained due some intelligent input: we selected the square box intuitively, just looking at the positions of training points.[13] Perhaps, we can continue to improve performance (if we are still far from the Bayesian rate) by adding more invariants (choosing different subspaces, different boxes, and different functions[14]).

As we have already noted, there exists an important difference between idea of *features* that is used in classical algorithms and idea of *predicates* that is use for constructing *invariants*.

In order to construct an accurate rule, the classical paradigm recommends the introduction of special functions called features and construction of the rule based on these features. With increasing number of features one *increases the capacity* (the VC dimension) of the set of admissible functions from which one chooses the solution. According to the VC bounds, the larger is the capacity of that set, the more training examples are required for an accurate estimate.

With the increase of the number of invariants, one *decreases the capacity* of admissible set of functions, since the functions from which one chooses the solution preserve all the

[13] This is the same methodology that physicists use: "Find a situation (the box $\mathcal{B}$ in Fig. 2), where the existing model of the Nature (the approximation $P_n(y = 1|x)$) contradicts the reality (contradicts data (x_i, y_i) from the box $\mathcal{B}$) and then fix the model (obtain a new approximation $P_{n+1}(y = 1|x)$) which does not have the contradictions." Note that the most difficult part in model refinement is to find a contradiction situation.

[14] The choice the best position of the box can be done algorithmically.

invariants. According to the same VC bounds, the decrease of the capacity of admissible functions improves the performance while using the same number of training examples.

6.6 More examples of intelligence-driven invariants

The role of Intelligent Teacher in training processes described in this paper is to introduce functions $\psi_1(x), \ldots, \psi_n(x)$ for the problem of interest. The Student has to understand which functions the Teacher suggested and use them to create invariants. Below are some examples of functions that the Teacher can suggest for invariants.

Example 1 Suppose that the Teacher teaches Student to recognize digits by providing a number of examples and also suggesting the following heuristics: "In order to recognize the digit zero, look at the center of picture—it is usually light; in order to recognize the digit 2, look at the bottom of the picture—it usually has a dark tail" and so on.

From the theory above, the Teacher wants the Student to construct specific predicates $\psi(x)$ to use them for invariants. However, the Student does not necessarily construct exactly the same predicate that the Teacher had in mind (the Student's understanding of concepts "center of the picture" or "bottom of the picture" can be different). Instead of $\psi(x)$, the Student constructs function $\widehat{\psi(x)}$. However, this is acceptable, since any function from L_2 can serve as a predicate for an invariant.

The Teacher explains which group of pixels constitutes the center of picture and the Student chooses "approximately" these pixels. Suppose that Student chose pixels $p_{t_1}, \ldots, P_{t_p}$ and lightness of pixel p_{t_1} is $x(p_{t_1})$. The lightness of center picture x_i is measured by the sum of lightness value of chosen pixels defined by inner product

$$\widehat{\psi(x_i)} = (z_0, x_i),$$

where z_0 is a binary vector which has coordinates that are equal to one for the chosen group of pixels that define the center of picture, while all other coordinates are equal to zero.[15] (Similarly, binary vector z_2 defines the concept of darkness of tail in the bottom of picture.) Using these functions and a number of examples, Student defines the invariants

$$A^T K \widehat{\Phi} + c 1_\ell^T \widehat{\Phi} = Y^T \widehat{\Phi},$$

where $\widehat{\Phi} = (\widehat{\psi}(x_1), \ldots, \widehat{\psi}(x_\ell))^T$. This equation describes idea of lightness of center of digit 0 (or darkness of tail in digit 2). Using these invariants, Student constructs a rule that takes Teacher's heuristics into account.

Generally speaking, the vector z could be defined as any real-valued defining any function that is linear in the pixel space.

Example 2 Suppose that a medical doctor teaches a medical student to classify some disease ($y = 1$). Suppose that the doctor believes that this disease is related to blood circulation. He suggests that some non-linear predicate function called "Blood Circulation Index"

$$\psi(x_i) = (H_b(x_i) - L_b(x_i)) Fr(x_i)$$

(the difference between systolic and diastolic blood pressures times heartbeat frequency) somehow expresses this disease. While demonstrating patients with disease $y = 1$ to the student, the doctor recommends that the student to pay attention to this index. The student

[15] Student can choose any appropriate weights.

observes ℓ patients (described by vectors $x_i, \ i = 1, \ldots, \ell$) and constructs vector Φ with coordinates $\Phi(x_i) = \psi(x_i), \ i = 1, \ldots, \ell$. The expression for the corresponding invariant has the form

$$\Phi^T K A \approx \Phi^T Y.$$

Next invariants are inspired by elements of machine learning techniques. In order to use this technique, we do not have to introduce the explicit form of predicate function. Instead, we introduce the algorithm of computing (using training data) the predicate function at any point of interest.

Example 3 (*local structure predicate*) Consider the following predicate-function. For point x', function $\psi_\rho(x_0)$ computes the expectation of the vectors of the first class in the sphere S with center x' and given radius ρ

$$\psi_\rho(x') = \int_S P(y = 1|x)dx$$

Then the value

$$R(\rho) = \int \psi_\rho(x')dP(x')$$

defines the local characteristics of conditional probability function $P(y = 1|x)$.

Let us compute, for any vector x_i of the training set the following characteristics: the number of examples of the first class that belong to sphere of radius ρ with center x_i. Let this value be $\psi_\rho(x_i)$. Consider vector $\Phi_\rho = (\psi_\rho(x_1), \ldots, \psi_\rho(x_\ell))^T$ to define the invariant. Choosing different values of radius ρ, one constructs different local characteristics of desired function.

Invariants with vectors $\Phi_{\rho_k}, \ k = 1, \ldots, n$ provide a description of the structure of the conditional probability function $P(y = 1|x)$. They can be taken into account when one estimates $P(y = 1|x)$.

Example 3a Consider the following values $\psi_k(x_i)$. For any vector x_i of training vector x_i, we compute the number of vectors of the first class among k nearest neighbors of vector x_i. In order to construct invariants, we use the vectors

$$\Phi_k = (\psi_k(x_1), \ldots, \psi_k(x_\ell))^T, \quad k = 1, 2, \ldots, n.$$

6.7 Example of a simple LUSI algorithm

Given data

$$(x_1, y_1), \ldots, (x_\ell, y_\ell)$$

consider the following learning method based on predicates $\psi_k(x)$ (for example, those defined in Example 3a).

Step 0. Construct vSVM (or SVM) estimate of conditional probability function (see Sects. 5.5 and 6.3).

Step 1. Find the maximal disagreement value $\mathcal{T}_s$ (108) for vectors

$$\Phi_k = (\psi_k(x_1), \ldots, \psi_k(x_\ell))^T, \quad k = 1, \ldots, s, \ldots.$$

Step 2. If the value $\mathcal{T}_s$ is large, i.e.,

$$\mathcal{T}_s = \text{argmax}(\mathcal{T}_1, \ldots, \mathcal{T}_s \ldots) > \delta,$$

add the invariant (107) with Φ_s; otherwise stop.

Step 3. Find the new approximation of the conditional probability function and go to step 1; otherwise stop.

Functions $\psi_k(x)$ can be intriduced in any subspace of space X.

7 Conclusion

In this paper, we introduced LUSI paradigm of learning which, in addition to the standard data-driven mechanism of minimizing risk, leverages an intelligence-driven mechanism of preserving statistical invariants (constructed using training data and given predicates). In this new paradigm, one first selects (using invariants) an admissible subset of functions which contains the desired solution and then chooses the solution using standard training procedures.

The important properties of LUSI are as follows: if the number ℓ of observations is sufficiently large, then (1) the admissible subset of functions always contains a good approximation to the desired solution regardless of the number of invariants used, and (2) the approximation to the desired solution is chosen by the methods that provide global minima to the guarantee risk[16].

LUSI method can be used to increase the accuracy of the obtained solution and to decrease the number of necessary training examples.

Acknowledgements This material is based upon work partially supported by AFRL under contract FA9550-15-1-0502. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We thank the anonymous reviewers, S. Sugrim, and P. Toccaceli for their careful reading of our paper and their remarks, which helped to improve the paper's readability.

Appendix 1: Invariants with respect to linear transformations

Let us estimate conditional probability function of the form

$$P(y=1|x) = A^T\mathcal{K}(x) + c,$$

Consider the linear transformation $x' = \mathcal{U}x$ of elements $x \in X$ into elements $x' \in X$ and let $\mathcal{U}^{-1}x'$ be inverse transformation. Suppose that our goal is to construct a function for which $P(y=1|x) = P(y=1|\mathcal{U}^{-1}x)$. For such a function,

$$\int P(y=1|x)\psi(x)dP(x) = \int P(y=1|\mathcal{U}^{-1}x)\psi(x)dP(x). \tag{125}$$

[16] The idea of two-stage learning is also realized in deep neural networks (DNN), where, at the first stage (using "deep architecture"), an appropriate network is constructed and then, at the second stage, using standard for NN training procedures, the solution is obtained. DNN, however, cannot guarantee either that the constructed network contains a good approximation to the desired function or that it can find the best solution for the given network.

Consider the functional

$$R = \int P(y=1|\mathcal{U}^{-1}x')\psi(x')dP(x').$$

By changing variables $x = \mathcal{U}^{-1}x'$, we obtain

$$R = \int P(y=1|\mathcal{U}^{-1}x')\psi(x')dP(x') = |U| \int P(y=1|x)\psi(\mathcal{U}x)dP(\mathcal{U}x).$$

Examples of transformation.

(1) Shift by vector h

$$R = \int P(y=1|x-h)\psi(x)dP(x) = \int P(y=1|x)\psi(x+h)dP(x+h).$$

(2) Similarity transformation $z = \mathcal{A}x = (a_1x^1, \dots, a_nx^n)$, where $a_1, \dots, a_n$ are parameters of coordinate transformation.

$$R = \int P(y=1|\mathcal{A}^{-1}x)\psi(x)dP(x) = |\mathcal{A}| \int P(y=1|x)\psi(\mathcal{A}x)dP(\mathcal{A}x),$$

where $|\mathcal{A}| = a_1a_2 \dots a_n$ (n is the dimensionality of X space).

(3) Rotation transformation. Let $\mathcal{R}x$ be a rotation transformation. For this transformation, $|\mathcal{R}| = 1$. Then

$$R = \int P(y=1|\mathcal{R}^{-1}x)\psi(x)dP(x) = \int f(x)\psi(\mathcal{R}x)P(\mathcal{R}x).$$

Consider the invariant defined by predicate-function $\psi_s(x),\ s = 1, \dots, m$

$$\int P(y=1|x)\psi_s(x)dP(x) = \int \psi(x)dP(y=1,x) = a_s \tag{126}$$

and suppose we would like to preserve this invariant for linearly transformed vectors $\mathcal{U}^{-1}x$. That is, according to (125), along with invariant (126), we would like to preserve the following invariant

$$\int P(y=1|\mathcal{U}^{-1}x)\psi_s(x)dP(x) = a_s.$$

We can rewrite this equation as

$$\int P(y=1|x)\psi_s(\mathcal{U}x)dP(\mathcal{U}x) = a_s,$$

where $\psi_s(\mathcal{U}x)$ is the corresponding predicate-function.

To find the approximation of function that preserves both statistical invariants defined by predicate-functions $\psi_s(x)$ and invariants defined by the same predicate with geometrically transformed vectors $\psi_s(\mathcal{U}x)$, we have to minimize the functional (80) subject to constraint

$$A^T K\Phi_s + c1_\ell^T\Phi_s = Y^T\Phi_s$$

and constraints

$$A^T K\Phi_s^{\mathcal{U}} + c1_\ell^T\Phi_s^{\mathcal{U}} = Y^T\Phi_s$$

where $\Phi_s^{\mathcal{U}} = (\psi(\mathcal{U}x_1), \dots, \psi(\mathcal{U}x_\ell))^T$.

Appendix 2: Pattern recognition in rebalanced environment

The classical setting of pattern recognition problem given in Sect. 1.3 considers the situation when generator G of vectors x is the same in both training and test sessions. One of the important modifications of this setting is the problem of obtaining a classification rule for generator G that is re-balanced for the test session.

Suppose that the new generator of random event x is such that conditional probabilities $P(y=k \mid x),\ k=1,\dots,n$ remain the same for both training and test data. However, the probabilities of classes $p_*(y=k)$ in the test session are different from the probabilities of classes $p(y=k)$ in the training session. *The probabilities* $p_*(y=k),\ k=1,\dots,n$ *are given* and $\sum_k p_*(k)=1$. The goal is, by using the data generated by $P(x)$, to find the rule of classification of vectors x produced by generator

$$P^*(x)=\sum_{k=1}^{n} P(x|y=k)p_*(y=k) \tag{127}$$

that was constructed using given probabilities $p_*(y=k)$. This replacement of distribution function $P(x)$ defined for training session with the function $P^*(x)$ defined for the test session we call *re-balancing* of data (environment).

The necessity of re-balancing the data appears, for example, when one tries to classify a rare disease (with small p) using real observations. In order to control the value of the error in a classification of rare diseases, one can re-balance the probability of the class of interest by increasing the value p. As we will see, this leads to a special estimate of V-matrix.

V-matrix for rebalanced pattern recognition problem

For simplicity, we consider two-class classification problem: suppose we are given training data

$$(x_1,y_1),\dots,(x_\ell,y_\ell),\quad x_i\in R^n;\, y\in\{0,1\} \tag{128}$$

defined by the probability distribution function

$$P(x)=p(y=1)P(x|y=1)+(1-p(y=1))P(x|y=0). \tag{129}$$

However, our goal is, using data (128), to construct a rule for classification in the environment defined by the probability distribution function

$$P^*(x)=p_*(y=1)P(x|y=1)+(1-p_*(y=1))P(x|y=0), \tag{130}$$

which is different from (129). While data (128) are generated according to distribution function $P(y,x)=P(y|x)P(x)$, we would like to estimate the conditional probability function for distribution $P_*(y,x)=P(y|x)P_*(x)$.

Let $\ell(1)$ and $\ell(0)$ be the number of elements of training data (128) belonging to the class $y=1$ and to the class $y=0$, respectively. Consider the empirical estimate of cumulative distribution function for the elements of the first class

$$P_{emp}(x|y=1)=\frac{1}{\ell(1)}\sum_{i=1}^{\ell} y_i\theta(x-x_i) \tag{131}$$

and for the elements of the second class

$$P_{emp}(x|y=0) = \frac{1}{\ell(0)} \sum_{i=1}^{\ell} (1-y_i)\theta(x-x_i). \tag{132}$$

Using (131) and (132), we construct re-balanced cumulative distribution function (129)

$$P^*_{emp}(x) = \frac{p_*}{\ell(1)} \sum_{i=1}^{\ell} y_i\theta(x-x_i) + \frac{1-p_*}{\ell(0)} \sum_{i=1}^{\ell} (1-y_i)\theta(x-x_i).$$

The estimate of probability function $P_{emp}(x, y=1)$ is

$$P^*_{emp}(x, y=1) = \frac{p_*}{\ell(1)} \sum_{i=1}^{\ell} y_i\theta(x-x_i).$$

We introduce the notations

$$a(y_i) = \begin{cases} \dfrac{p_*}{\ell(1)}, & \text{if } y_i = 1, \\ \dfrac{1-p_*}{\ell(0)}, & \text{if } y_i = 0. \end{cases}$$

In order to estimate the desired conditional probability functions for re-balanced data, we use the corrected estimates $P_{emp}(x)$ and $P_{emp}(y=1,x)$ instead of $P(x)$ and $P(y=1,x)$ in (129). We obtain the equation

$$\sum_{i=1}^{\ell} a(y_i)\theta(x-x_i)f(x_i) = \sum_{j=1}^{\ell} y_j a(y_j)\theta(x-x_j). \tag{133}$$

In order to keep our notations in the matrix form, we define the diagonal matrix S with diagonal elements $s_{ii} = a(y_i)$ ($s_{i,j} = 0$ if $i \neq j$). Solving Eq. (132) using the regularization method described in Sect. 5.3, we obtain the solution which is different from the one described in Sect. 5.4 solution only in its form of V matrix. For the re-balanced solution, V_r-matrix has the form

$$V_r = SVS,$$

where V is a standard estimate of V-matrix and matrix S defines the re-balancing effect. In order to preserve the invariants for the re-balanced data, the solution has to satisfy the equalities

$$\Phi_s^T KAS + c\Phi_s^T 1_\ell S = \Phi_s^T YS; \quad s = 1, \ldots, m,$$

which are equivalent to the equalities

$$\Phi_s^T KA + c\Phi_s^T 1_\ell = \Phi_s^T Y; \quad s = 1, \ldots, m. \tag{134}$$

Therefore, in order to estimate the conditional probability function that maintains invariants (134), one has to minimize the functional

$$R(A) = (KA + c1_\ell)^T V_r (KA + c1_\ell) - 2(KA + c1_\ell)^T V_r Y + \gamma_\ell A^T KA$$

subject to constrains (134). This problem has a closed-form solution

$$f(x) = A_*^T \mathcal{K}(x) + c,$$

where

$$
\begin{aligned}
A_* &= A_b^* - cA_c^* - \sum_{k=0}^{m} \mu_s A_s^* \\
A_b^* &= (V_r K + \gamma I)^{-1} V_r Y; \\
A_c &= (V_r K + I)^{-1} V_r 1_\ell \\
A_s^* &= (V_r K + I)^{-1} \Phi_s
\end{aligned}
$$

Coefficients b, c and μ_s are the solutions of the system of linear equations

$$
c[1_\ell^T V_r K A_c^* - 1_\ell^T V_r 1_\ell] + \sum_{s=1}^{m} \mu_s [1_\ell^T V_r K A_s^* - 1_\ell^T \Phi_s] = [1_\ell^T V_r K A_b^* - 1_\ell^T V_r Y]
$$

$$
c[1_\ell^T K A_c^* - 1_\ell^T \Phi_s] + \sum_{s=1}^{m} \mu_s A_s^T K \Phi_s = [\Phi_s^T K A_b^* - \Phi_s^T Y], \quad s = 1, \ldots, m.
$$

References

Corfield, D., Schölkopf, B., & Vapnik, V. (2005). Popper, falsification and the VC-dimension. Technical Report 145, Max Planck Institute for Biological Cybernetics.

Corfield, D., Schölkopf, B., & Vapnik, V. (2009). Falsificationism and statistical learning theory: Comparing the Popper and Vapnik–Chervonenkis dimensions. *Journal for General Philosophy of Science*, *40*(1), 51–58.

Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed 1 Feb 2018.

Popper, K. (1934). *The logic of scientific discovery*. London: Hutchinson.

Tikhonov, A., & Arsenin, V. (1977). *Solutions of Ill-posed problems*. Washington: W.H. Winston.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V., & Chervonenkis, A. (1974). *Theory of pattern recognition*. Moscow: Nauka. (in Russian).

Vapnik, V., & Izmailov, R. (2017). Knowledge transfer in SVM and neural networks. *Annals of Mathematics and Artificial Intelligence*, *81*(1–2), 3–19.

Vapnik, V., & Stefanyuk, A. (1978). Nonparametric methods for estimating probability densities. *Automation and Remote Control*, *8*, 38–52.

Wapnik, W., & Tscherwonenkis, A. (1979). *Theorie der Zeichenerkennung*. Berlin: Akademie-Verlag.

Wigner, E. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics*, *13*(1), 1–14.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.