Neurocomputing 129 (2014) 25–32

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A theoretical framework for supervised learning from regions

Giorgio Gnecco [a,b], Marco Gori [c], Stefano Melacci [c,*], Marcello Sanguineti [b]

[a] IMT—Institute for Advanced Studies, Piazza S. Ponziano 6, 55100 Lucca, Italy
[b] DIBRIS—University of Genoa, Via Opera Pia 13, 16145 Genova, Italy
[c] DIISM—University of Siena, Via Roma 56, 53100 Siena, Italy

ARTICLE INFO

Article history:
Received 31 January 2012
Received in revised form
24 April 2012
Accepted 5 June 2012
Available online 25 October 2013

Keywords:
Supervised learning
Kernel machines
Propositional rules
Variational calculus
Infinite-dimensional optimization
Representer theorems

ABSTRACT

Supervised learning is investigated, when the data are represented not only by labeled points but also labeled regions of the input space. In the limit case, such regions degenerate to single points and the proposed approach changes back to the classical learning context. The adopted framework entails the minimization of a functional obtained by introducing a loss function that involves such regions. An additive regularization term is expressed via differential operators that model the smoothness properties of the desired input/output relationship. Representer theorems are given, proving that the optimization problem associated to learning from labeled regions has a unique solution, which takes on the form of a linear combination of kernel functions determined by the differential operators together with the regions themselves. As a relevant situation, the case of regions given by multi-dimensional intervals (i.e., "boxes") is investigated, which models prior knowledge expressed by logical propositions.

1. Introduction

The classical supervised learning framework is based on a collection of ℓ labeled points, $\mathcal{L} = \{(x_\kappa, y_\kappa),\ \kappa = 1, \ldots, \ell\}$, where $x_\kappa \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_\kappa \in \{-1, 1\}$. We consider the situation in which supervised learning exploits not only labeled points but also $\ell_{\mathcal{X}}$ labeled regions $\mathcal{L}_{\mathcal{X}} = \{(\mathcal{X}_\kappa, y_\kappa),\ \kappa = 1, \ldots, \ell_{\mathcal{X}}\}$ of the input space, where $\mathcal{X}_\kappa \in 2^{\mathcal{X}}$ and $y_\kappa \in \{-1, 1\}$. In the limit case such regions degenerate to single points, so we focus on a fairly general context in which there is no distinction between the supervised entities and we deal with $\ell_t := \ell + \ell_{\mathcal{X}}$ labeled pairs. This framework and its potential impact in real-world applications has been investigated in different contexts (see [1] and the references therein).

A seminal work in this respect is [2], where it was proposed to embed labeled polyhedral sets into Support Vector Machines (SVMs). The corresponding model, called Knowledge-based SVM (KSVM), has been the subject of a number of various investigations [3–6]. A particularly relevant situation corresponds to regions given by multi-dimensional intervals (i.e., "boxes") $\mathcal{X}_\kappa = \{x \in \mathbb{R}^d : x^i \in [a_\kappa^i, b_\kappa^i],\ i = 1, \ldots, d\}$, where $a_\kappa, b_\kappa \in \mathbb{R}^d$ collect the lower and upper bounds, respectively. The pair $(\mathcal{X}_\kappa, y_\kappa)$ formalizes the knowledge provided by a supervisor in terms of logical propositions of the form $\forall x \in \mathbb{R}^d$, $\bigwedge_{i=1}^d ((x^i \geq a_\kappa^i) \wedge (x^i \leq b_\kappa^i)) \Rightarrow \text{class}(y_\kappa)$.

In [7], the problem of learning was extended by taking into account the supervision on multi-dimensional intervals of the input space, which model prior knowledge expressed by logical propositions. The effectiveness of such an approach was evaluated therein via simulations on real-world problems of medical diagnosis and image categorization. Taking the hint from the numerical experiments presented in [7], in this paper we give theoretical insights into the learning paradigm proposed therein.

We formulate the problem of learning via supervision on input regions by introducing a loss function that involves them and adopting the regularization framework proposed in [8]. Each region $\mathcal{X}_\kappa$ is associated with its characteristic function $1_{\mathcal{X}_\kappa}(x)$ and its normalized form $\hat{1}_{\mathcal{X}_\kappa}(x) := 1_{\mathcal{X}_\kappa}(x) / \int_{\mathcal{X}} 1_{\mathcal{X}_\kappa}(x)\, dx$ degenerates to the Dirac distribution $\delta(x - x_\kappa)$ in the case in which $\mathcal{X}_\kappa = \{x_\kappa\}$. We model the corresponding learning problem as the minimization, over an *infinite-dimensional space* (whose elements are the admissible solutions to the supervised learning task) of a functional, called *regularized functional risk*, that consists of two terms. The first term enforces closeness to the labeled data (regions and points), whereas the second one, called *regularization term*, expresses requirements on the global behavior of the desired input/output functional relationship. The trade-off between such terms is achieved by a weight parameter, as typically done in Tikhonov's regularization [9].

We express the regularization term via differential operators, following the line of research proposed in [8]. The loss term results from the following two contributions: one from regions with non-null Lebesgue measure and another one that originates from

* Corresponding author.
E-mail addresses: giorgio.gnecco@imtlucca.it, giorgio.gnecco@unige.it (G. Gnecco), marco@diism.unisi.it (M. Gori), mela@diism.unisi.it (S. Melacci), marcello.sanguineti@unige.it (M. Sanguineti).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.06.065

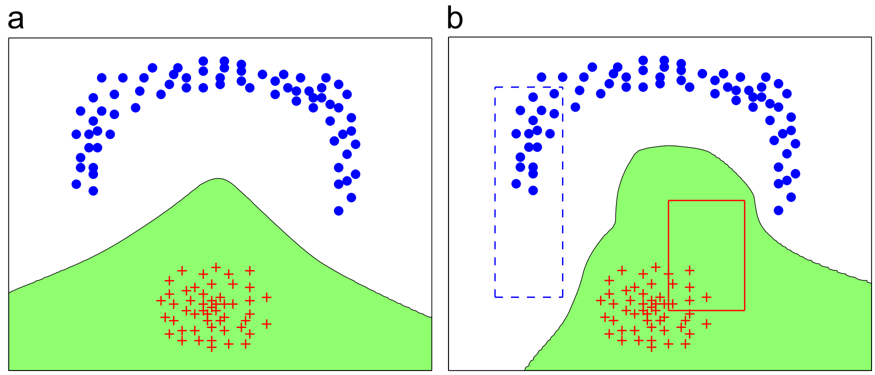

Fig. 1. (a) A two-class data set and the separating surface learned by a non-linear kernel-machine classifier. Examples of class 1 are represented by red crosses, whereas examples of class 2 are drawn with blue dots. (b) The data collection is augmented with two labeled space regions (bounded by the blue-dotted rectangle in the case of class 1, by the red-solid box for class 2). A kernel machine is trained using a box kernel, which allows the classifier to learn from the whole data collection (points and regions). The resulting class-separation boundary is shown. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

points. As the minimization of such a functional entails a difficult infinite-dimensional problem, we also consider learning modeled as a more affordable variational task obtained by replacing the functional risk with an *average risk*. We show that in this case the infinite-dimensional optimization collapses to a finite-dimensional one. As ambient spaces we consider Sobolev spaces of orders guaranteeing that they are made up of continuous functions, in such a way that the learning functionals are well-defined when the regions degenerate to points.

Under the hypothesis that the Green's function of the regularization operator is the kernel of a Hilbert spaces of a special type, called *reproducing kernel Hilbert space*[1] (*RKHS*) [14], we prove new *representer theorems* (see, e.g., [15, p. 42], [16–19]), showing that the minimization problem has a unique solution, which takes on the form of a linear combination of kernel functions determined by the differential operators together with the labeled regions. So, the solution to the regularized problem of learning from regions does not lead to the kernel expansion on the available data points and the kernel is no longer the Green's function of the associated operator, as instead it happens from classical results of this kind (see [20] and [21, p. 94]).

As a meaningful learning case, we investigate regions given by multidimensional intervals (i.e., "boxes"), which originates the *box kernels*. Fig. 1 shows an example of box kernels in the case of a non-linear kernel-machine classifier. First, the classifier is trained on a 2-class data set of points, and the separating hyperplane is depicted in Fig. 1(a). Then, a supervision is given on two space regions bounded by multi-dimensional intervals; Fig. 1(b) shows the resulting separation boundary. The box kernel allows the classifier to embed knowledge on labeled space regions, whereas classical kernels are designed to operate merely on points.

The paper is organized as follows. In Section 2 we state the problem of learning from labeled regions and/or labeled points as the infinite-dimensional minimization of the functional risk. We investigate existence and uniqueness of its solution on the Sobolev space of functions that are square-integrable together with their partial derivatives up to a suitable order. Section 3 provides representer theorems for such a solution and considers the learning problem modeled via the minimization of the average risk. The particular case of regions given by multi-dimensional intervals (i.e., "boxes") is addressed in Section 4. In Section 5 we compare the two learning problems associated with the minimizations of the functional risk and the average risk, respectively. To make the paper self-contained, two appendices are provided on RKHSs and functionals. Preliminary results were presented in [7].

2. Learning from labeled regions and points

We formulate the problem of learning from labeled regions and/or labeled points in a unique framework, where each point corresponds to a singleton. Given a labeled set $\mathcal{X}_\kappa$, the characteristic function $1_{\mathcal{X}_\kappa}(x)$ associated with it is identically 1 when $x \in \mathcal{X}_\kappa$, otherwise it is identically 0. Denoting by $\mathrm{vol}(\mathcal{X}_\kappa) = \int_{\mathbb{R}^d} 1_{\mathcal{X}_\kappa}(x)\, dx$ the measure of the set, the normalized characteristic function is $\hat{1}_{\mathcal{X}_\kappa}(x) := 1_{\mathcal{X}_\kappa}(x)/\mathrm{vol}(\mathcal{X}_\kappa)$. When the region degenerates to a single point x_κ we denote by $\hat{1}_{\mathcal{X}_\kappa}(x)$ the Dirac delta $\delta(x - x_\kappa)$.

Let $w : \mathcal{X} \to \mathbb{R}^+$ be a continuous weight function (e.g., proportional to the probability density $p : \mathcal{X} \to \mathbb{R}^+$ of the inputs), $V : \mathbb{R}^2 \to \mathbb{R}^+$ a convex and differentiable loss function, $\lambda > 0$ a regularization parameter, and $P := (P^0, \ldots, P^{r-1})$ a vector of r finite-order differential operators of maximum order of derivation k and with constant coefficients, with formal adjoint $P^\star$ [22,23]. Adopting the framework described in [8], we formulate the problem of learning from labeled regions as the minimization on a suitable class of functions $\mathcal{F}$ of the functional

$$\mathcal{R}(f) := \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)\, dx + \frac{\lambda}{2} \|Pf\|^2, \tag{1}$$

where $\mathbb{N}_m$ denotes the set of the first m positive integers,

$$\|Pf\|^2 := (Pf, Pf) = (P^\star P f, f) = (Lf, f),$$

$$(f, g) := \int_{\mathbb{R}^d} f(x) \cdot g(x)\, dx,$$

and $L := P^\star P$ (which has $2k$ as its maximum order of derivation). We call $\mathcal{R}(\cdot)$ in Eq. (1) the *regularized functional risk*. When all the regions degenerate to points, we get $\ell_t = \ell$, $\hat{1}_{\mathcal{X}_\kappa}(x) = \delta(x - x_k)$, and Eq. (1) becomes

$$\mathcal{R}(f) := \sum_{\kappa \in \mathbb{N}_\ell} V(y_\kappa, f(x_\kappa)) \cdot w(x_\kappa) + \frac{\lambda}{2} \|Pf\|^2, \tag{2}$$

which is the classical form of the regularized risk [24] when supervision is performed on labeled points.

We search for the minimizer f° in the Sobolev space $\mathcal{F} = \mathcal{W}^{k,2}$, i.e., the subset of $\mathcal{L}^2$ whose functions have square-integrable weak partial derivatives up to the order k. The loss term in Eq. (1) can be considered as resulting from the following two contributions: $\sum_{\kappa \in \mathbb{N}_{\ell_\mathcal{X}}} \int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)\, dx$, coming from a region with non-null Lebesgue measures $\mathrm{vol}(\mathcal{X}_\kappa)$, and $\sum_{\kappa \in \mathbb{N}_\ell} w(x_\kappa) \cdot V(y_\kappa, f(x_\kappa))$,

[1] Such spaces were introduced into applications closely related to learning by Parzen [10] and Wahba [11], and into learning theory by Cortes and Vapnik [12] and Girosi [13].