Appl. Comput. Harmon. Anal. 37 (2014) 185–217

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Learning sets with separating kernels

Ernesto De Vito [a,*], Lorenzo Rosasco [b,c,d], Alessandro Toigo [e,f]

[a] DIMA, Università di Genova, Genova, Italy
[b] DIBRIS, Università di Genova & Istituto Italiano di Tecnologia, Italy
[c] Massachusetts Institute of Technology, USA
[d] Istituto Italiano di Tecnologia, Italy
[e] Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
[f] I.N.F.N., Sezione di Milano, Milano, Italy

ARTICLE INFO

Article history:
Received 11 April 2012
Received in revised form 31 October 2013
Accepted 7 November 2013
Available online 15 November 2013
Communicated by Ding-Xuan Zhou

Keywords:
Set estimation
Kernel methods
Spectral regularization

ABSTRACT

We consider the problem of learning a set from random samples. We show how relevant geometric and topological properties of a set can be studied analytically using concepts from the theory of reproducing kernel Hilbert spaces. A new kind of reproducing kernel, that we call separating kernel, plays a crucial role in our study and is analyzed in detail. We prove a new analytic characterization of the support of a distribution, that naturally leads to a family of regularized learning algorithms which are provably universally consistent and stable with respect to random sampling. Numerical experiments show that the proposed approach is competitive, and often better, than other state of the art techniques.

1. Introduction

In this paper we study the problem of learning from data the set where the data probability distribution is concentrated. Our study is more broadly motivated by questions in unsupervised learning, such as the problem of inferring geometric properties of probability distributions from random samples.

In recent years, there has been great progress in the theory and algorithms for supervised learning, i.e. function approximation problems from random noisy data [10,22,29,55,74]. On the other hand, while there are a number of methods and studies in unsupervised learning, e.g. algorithms for clustering, dimensionality reduction, dictionary learning (see Chapter 14 of [38]), many interesting problems remain largely unexplored.

Our analysis starts with the observation that many studies in unsupervised learning hinge on at least one of the following two assumptions. The first is that the data are distributed according to a probability distribution which is absolutely continuous with respect to a reference measure, such as the Lebesgue measure. In this case it is possible to define a density and the corresponding density level sets. Studies in this scenario include [8,30,44,69] to name a few. Such an assumption prevents considering the case where the data are represented in a high-dimensional Euclidean space but are concentrated on a Lebesgue negligible subset, as a lower-dimensional submanifold. This motivates the second assumption – sometimes called *manifold assumption* – postulating that the data lie on a low-dimensional Riemannian manifold embedded in a Euclidean space. This latter idea has triggered a large number of different algorithmic and theoretical studies (see for example [4,6,20,21,27,59]). Though the manifold assumption has proved useful in some applications, there are many practical scenarios where it might not be satisfied. This observation has motivated considering more general situations such as *manifold plus noise* models [18,52], and models where the data are described by combinations of more than one manifold [46,76].

* Corresponding author.
E-mail addresses: devito@dima.unige.it (E. De Vito), lrosasco@mit.edu (L. Rosasco), alessandro.toigo@polimi.it (A. Toigo).

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.11.003

Here we consider a different point of view and work in a setting where the data are described by an abstract probability space and a *similarity function* induced by a reproducing kernel [65]. In this framework, we consider the basic problem of estimating the set where the data distribution is concentrated (see Section 1.2 for a detailed discussion of related works). A special class of reproducing kernels, that we call separating kernels, plays a special role in our study. First, it allows to define a suitable metric on the probability space and makes the support of the distribution well defined; second, it leads to a new analytical characterization of the support in terms of the null space of the integral operator associated to the reproducing kernel.

This last result is the key towards a new computational approach to learn the support from data, since the integral operator can be approximated with high probability from random samples [58,65]. Estimation of the null space of the integral operator can be unstable, and regularization techniques can be used to obtain stable estimators. In this paper we study a class of regularization techniques proposed to solve ill-posed problems [34] and already studied in the context of supervised learning [3,48]. Regularization is achieved by *filtering* out the small eigenvalues of the sample empirical matrix defined by the kernel. Different algorithms are defined by different filter functions and have different computational properties. Consistency and stability properties for a large class of spectral filters and of the corresponding algorithms are established in a unified framework. Numerical experiments show that the proposed algorithms are competitive, and often better, than other state of the art techniques.

The paper is divided into two parts. The first part includes Section 2, where we establish several mathematical results relating reproducing kernel Hilbert spaces of functions on a set X and the geometry of the set X itself. In particular, in this section we introduce the concept of separating kernel, which we further explore in Section 3. These results are of interest in their own right, and are at the heart of our approach. In the second part of the paper we discuss the problem of learning the support from data. More precisely, in Section 4 we illustrate some algorithms for learning the support of a distribution from random samples. In Section 5 we establish universal consistency for the proposed methods and discuss stability to random sampling. We conclude in Sections 6 and 7 with some further discussions and some numerical experiments, respectively. A conference version of this paper appeared in [28]. We now start by describing in some more detail our results and discussing some related works.

1.1. Summary of main results

In this section we briefly describe the main ideas and results in the paper.

The setting we consider is described by a probability space (X, ρ) and a measurable reproducing kernel K on the set X [2]. The data are independent and identically distributed (i.i.d.) samples $x_1, \dots, x_n$, each one drawn from X with probability ρ. The reproducing kernel K reflects some prior information on the problem and, as we discuss in the following, will also define the geometry of X. The goal is to use the sample points $x_1, \dots, x_n$ to estimate the region where the probability measure ρ is concentrated.

To fix some ideas, the space X can be thought of as a high-dimensional Euclidean space and the distribution ρ as being concentrated on a region X_ρ, which is a smaller – and potentially lower dimensional – subset of X (e.g. a linear subspace or a manifold). In this example, the goal is to build from data an estimator X_n which is, with high probability, close to X_ρ with respect to a suitable metric.

We first note that a precise definition of X_ρ requires some care. If ρ is assumed to have a continuous density with respect to some fixed reference measure (for example, the Lebesgue measure in the Euclidean space), then the region X_ρ can be easily defined to be the closure of the set of points where the density function is non-zero. Nevertheless, this assumption would prevent considering the situation where the data are concentrated on a "small", possibly lower dimensional, subset of X. Note that, if the set X were endowed with a topological structure and ρ were defined on the corresponding Borel σ-algebra, it would be natural to define X_ρ as the support of the measure ρ, i.e. the smallest *closed* subset of X having measure one. However, since the set X is only assumed to be a measurable space, no a priori given topology is available. Here we also remark that the definition of X_ρ is not the only point where some further structure on X would be useful. Indeed, when defining a learning error, a notion of distance between the set X_ρ and its estimator X_n is also needed and hence some metric structure on X is required.

The idea is to use the properties of the reproducing kernel K to induce a metric structure – and consequently a topology – on X. Indeed, under some mild technical assumptions on K, the function

$$d_K(x, y) = \sqrt{K(x, x) + K(y, y) - 2K(x, y)} \quad \forall x, y \in X$$

defines a metric on X, thus making X a topological space. Then, it is natural to define X_ρ to be the support of ρ with respect to such metric topology. Moreover, the Hausdorff distance d_H induced by the metric d_K provides a notion of distance between closed sets.

The problem we consider can now be restated as follows: we want to learn from data an estimator X_n of X_ρ, such that $\lim_{n\to\infty} d_H(X_n, X_\rho) = 0$ almost surely. While X_ρ is now well defined, it is not clear how to build an estimator from data. A main result in the paper, given in Theorem 3, provides a new analytic characterization of X_ρ, which immediately suggests a new computational solution for the corresponding learning problem. To derive and state this result, we introduce a new notion of reproducing kernels, called separating kernels, that, roughly speaking, captures the sense in which the reproducing