Journal of Complexity 37 (2016) 1–33

Contents lists available at ScienceDirect

Journal of Complexity

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

On the robustness of regularized pairwise learning methods based on kernels

Andreas Christmann [a,*], Ding-Xuan Zhou [b]

[a] University of Bayreuth, Germany
[b] City University of Hong Kong, China

ARTICLE INFO

Article history:
Received 14 October 2015
Accepted 1 July 2016
Available online 11 July 2016

Keywords:
Machine learning
Pairwise loss function
Regularized risk
Robustness

ABSTRACT

Regularized empirical risk minimization including support vector machines plays an important role in machine learning theory. In this paper regularized pairwise learning (RPL) methods based on kernels will be investigated. One example is regularized minimization of the error entropy loss which has recently attracted quite some interest from the viewpoint of consistency and learning rates. This paper shows that such RPL methods and also their empirical bootstrap have additionally good statistical robustness properties, if the loss function and the kernel are chosen appropriately. We treat two cases of particular interest: (i) a bounded and non-convex loss function and (ii) an unbounded convex loss function satisfying a certain Lipschitz type condition.

1. Introduction

Regularized empirical risk minimization based on kernels has attracted a lot of interest during the last decades in statistical machine learning. To fix ideas, let $D_n = ((x_1, y_1), \ldots, (x_n, y_n))$ be a given data set, where the value x_i denotes the input value and y_i denotes the output value of the ith data point. Let L be a loss function which is typically of the form $L(x, y, f(x))$, where $f(x)$ denotes the predicted value for y, when x is observed, and the real-valued function f is unknown. Many regularized

* Corresponding author.
E-mail address: andreas.christmann@uni-bayreuth.de (A. Christmann).

http://dx.doi.org/10.1016/j.jco.2016.07.001
0885-064X/

learning methods are then defined as minimizers of the optimization problem

$$\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \operatorname{pen}(\lambda, f), \tag{1.1}$$

where the set $\mathcal{F}$ consists of real-valued functions f, $\lambda > 0$ is a regularization constant, and $\operatorname{pen}(\lambda, f) \geq 0$ is some regularization term to avoid overfitting for the case, that $\mathcal{F}$ is rich. One example is that $\mathcal{F}$ is a reproducing kernel Hilbert space H and $\operatorname{pen}(\lambda, f) = \lambda\|f\|_H^2$, see e.g. Vapnik [48,49], Poggio and Girosi [40], Wahba [50], Schölkopf and Smola [44], Cucker and Zhou [14], Christmann and Steinwart [9,10], Steinwart and Christmann [46] and the references cited there. Regularized empirical risk minimization based on kernels has also been investigated for additive models. We refer to Christmann and Hable [7] for results on consistency and robustness and to Christmann and Zhou [12] for fast learning rates.

In recent years there is quite some interest in related learning methods where a *pairwise loss function* is used, which yields optimization problems like

$$\inf_{f\in H} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L(x_i, y_i, x_j, y_j, f(x_i), f(x_j)) + \lambda\|f\|_H^2 \tag{1.2}$$

or asymptotically equivalent versions of it. In other words, the estimator for f is defined as the minimizer of the sum of a V-statistic of degree 2 and the regularizing term $\lambda\|f\|_H^2$, see e.g. Serfling [45]. An example of this class of learning methods occurs when one is interested in minimizing Renyi's entropy of order 2, see e.g. Hu et al. [32], Fan et al. [23], and Ying and Zhou [54] for consistency and fast learning rates. Another example arises from ranking algorithms, see e.g. Clemencon et al. [13] and Agarwal and Niyogi [1]. Other examples include gradient learning, and metric and similarity learning, see e.g. Mukherjee and Zhou [39], Xing et al. [53], and Cao et al. [5]. However, much less theory is currently known for such regularized learning methods given by (1.2) based on a pairwise loss function than for the more classical problem (1.1) using a standard loss function. This is true in particular for statistical robustness aspects. Statistical robustness is one important facet of a statistical method, especially if the data quality is only moderate or unknown, which is often the case in the so-called big data situation.

The main goal of this paper is to show that such regularized learning methods given by (1.2) have nice statistical robustness properties if a bounded and continuous kernel is used in combination with a convex, smooth, and separately Lipschitz continuous (see Definition 2.5) pairwise loss function. We also establish a representer theorem for such regularized pairwise learning methods, because we need it for our proofs, but the representer theorem may also be helpful to further research.

The rest of the paper has the following structure. In Section 2, we define pairwise loss functions, their corresponding risks, derive some basic properties of pairwise loss functions and their risks, and give some examples. In Section 3 we define regularized pairwise learning (RPL) methods treated in this paper and derive results on existence and uniqueness. We will show that shifted loss functions (defined in (3.9)) are useful to define RPL methods on the set of *all* probability measures without making moment assumptions. This is of course desirable, because the probability measure chosen by nature to generate the data is completely unknown in machine learning theory. Section 4 contains a representer theorem for RPL methods, which is our first main result, see Theorem 4.3. This result is interesting in its own right, but we use it as a tool to prove our statistical robustness results in Section 5. For a bounded kernel in combination with a bounded, but not necessarily convex pairwise loss function, we show that RPL methods have a bounded maxbias, see Theorem 5.1. For a bounded continuous kernel in combination with a convex pairwise loss function, which is separately Lipschitz continuous in the sense of Definition 2.5, we can formulate the two other main results of this paper: Theorem 5.3 shows that the RPL operator has a bounded Gâteaux derivative and hence a bounded influence function, see Corollary 5.4, and Theorem 5.5 shows that RPL methods and even their empirical bootstrap approximations are qualitatively robust, if some *non-stochastic* conditions are satisfied. Hence these statistical robustness properties of RPL methods hold for all probability measures provided that weak conditions on the input space, on the output space, on the kernel, and