Neural Networks 73 (2016) 26–35

Contents lists available at ScienceDirect

Neural Networks

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

A linear functional strategy for regularized ranking

Galyna Kriukova, Oleksandra Panasiuk, Sergei V. Pereverzyev, Pavlo Tkachenko*

Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria

ARTICLE INFO

Article history:
Received 5 June 2015
Received in revised form 20 August 2015
Accepted 28 August 2015
Available online 23 October 2015

Keywords:
Regularization
Ill-posed problem
Ranking
Linear functional strategy
Diabetes technology

ABSTRACT

Regularization schemes are frequently used for performing ranking tasks. This topic has been intensively studied in recent years. However, to be effective a regularization scheme should be equipped with a suitable strategy for choosing a regularization parameter. In the present study we discuss an approach, which is based on the idea of a linear combination of regularized rankers corresponding to different values of the regularization parameter. The coefficients of the linear combination are estimated by means of the so-called linear functional strategy. We provide a theoretical justification of the proposed approach and illustrate them by numerical experiments. Some of them are related with ranking the risk of nocturnal hypoglycemia of diabetes patients.

1. Introduction

In supervised learning one is given a set of examples $\{x_1, x_2, \ldots, x_m\} \subset X \subset \mathbb{R}^d$ labeled with the corresponding values $\{y_1, y_2, \ldots, y_m\} \subset Y \subset \mathbb{R}$ of the dependent variable y. Then the learning task is to use this data as a training set $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m \subset Z = X \times Y$ for assigning a proper label y to a previously unseen $x \in X$.

If the labels y_i are treated as the values of some function at given points x_i, then the above mentioned learning task is referred to as regression, or regression learning, and is one of the most well-studied problems in learning theory. In recent years another problem called ranking has gained attention in this theory.

Ranking is relatively new learning problem that is parallel to regression. After the first paper (Cohen, Schapire, & Singer, 1999) was published in 1999, ranking has been intensively investigated in the literature. Here we refer to Agarwal and Niyogi (2009), Chen (2012), Cossock and Zhang (2006), Crammer and Singer (2001), Freund, Iyer, Schapire, and Singer (2003), Herbrich, Graepel, and Obermayer (2000) and Mukherjee and Zhou (2006), just to mention a few publications.

In ranking one also learns a real-valued function $f : X \to Y$ that assigns a label y to $x \in X$, but the value $y = f(x)$ itself is not so important. What do matter are the relative ranks of instances $x, x' \in X$ induced by the labels $f(x), f(x')$. Namely, item x with higher rank has larger value $f(x)$.

Thus, the task of learning ranking is different from regression, but if we are looking for labeling functions $f : X \to Y$ in some Reproducing Kernel Hilbert Space (RKHS) on X then, in spite of the difference, both learning problems can be formulated as ill-posed linear integral operator equations of the first kind in the chosen RKHS (Chen, 2012; Kurkova, 2012; Smale & Zhou, 2007). The ill-posedness of such formulations calls for the employment of the regularization theory in the construction of regression and ranking algorithms. In this theory, the performance of algorithms is usually estimated under the source conditions expressed in terms of the so-called index functions. It is known (see, e.g. Lu and Pereverzev (2013) and Mathé and Pereverzev (2003)) that ill-posed equations may involve entirely different operators but nevertheless allow the same performance of regularization algorithms, if the solutions of these equations satisfy the source conditions for the same index function.

On the other hand, recently, the authors of Ying and Zhou (2015) have noticed a suboptimality of known ranking performance estimates compared to the corresponding regression ones. The same observation can be made from the comparison of Chen (2012) with Smale and Zhou (2007) and Xu, Fang, and Wang (2014) with Bauer, Pereverzev, and Rosasco (2007), when the same regularization schemes are compared under the source conditions generated by the same index functions.

This observation is not in agreement with the general fact of the regularization theory mentioned above, and it hints at a gap in the analysis of the regularized ranking algorithms. In the present paper we refine this analysis and show that, at least for the so-called offline learning, the performance of the regularized ranking is similar to that of the regularized regression learning.

* Corresponding author.
E-mail address: pavlo.tkachenko@oeaw.ac.at (P. Tkachenko).

http://dx.doi.org/10.1016/j.neunet.2015.08.012
0893-6080/© 2015 Elsevier Ltd. All rights reserved.

The above mentioned refinement is obtained as a by-product of the study of a new a posteriori regularization scheme in the context of learning. Note that usually a posteriori regularization means an adaptive choice of the parameter for single-parameter regularization methods such as Tikhonov, Lavrentiev or Landweber regularization and others like that. In the existing literature on the regularization theory it is suggested to make the above choice by using one of the known rules such as quasi-optimality criterion, cross-validation, the discrepancy principle, the balancing principle. In the context of learning these rules have been discussed in Caponnetto and Yao (2010) and , Pereverzyev, and Rosasco (2010). But these and similar rules select only one element from a family of approximants, calculated according to an employed regularization method, and leave others aside. Of course, the other approximants are used in the selection process, but then they are rejected, in spite of the numerical expenses made for their construction. At the same time, the rejected approximants may also contain important information on the approximated quantity of interest and can contribute to the improvement of the accuracy of its reconstruction (see Fig. 1).

In the present study we explore the idea to use the calculated approximants in the construction of a new one. More precisely, the idea is to use linear combinations of the approximants calculated for different values of the regularization parameter. It is clear that the best Hilbert-space approximation by such a linear combination requires the knowledge of inner products between the calculated approximants and the approximated element, which is of course unknown.

At the same time, the regularization theory tells us (see, e.g. Lu and Pereverzev (2013), Proposition 2.17) that the values of linear bounded functionals (e.g., inner products) at the approximated elements can be estimated more accurately than the elements themselves. The idea is to use the estimated values of the corresponding inner products for simulating the best linear combination of the calculated regularized approximants.

In the regularization theory the above-mentioned accurate estimation of linear functionals is often called as linear functional strategy (LFS). It was proposed in Anderssen (1986) and then further developed in Bauer, Mathé, and Pereverzev (2007), Goldenshluger and Pereverzev (2000) and Mathé and Pereverzev (2002). The previous results on LFS have been obtained under the assumption that the operators from the considered ill-posed equations are directly accessible, but that is not the case in the learning context. Therefore in the present study we at first perform adaptation–extension of LFS to that context.

The paper is organized as follows. In the next section we recall the setting of least squares ranking and its formulation as an ill-posed linear operator equation. Moreover, we describe a general regularization scheme for solving this equation. At the end of the section we specify the idea of a linear combination of regularized rankers. In Section 3 we present the above-mentioned extension of LFS and use it for simulating the best approximation by linear combinations of given rankers. The section also contains new bounds on the excess risk of the regularized ranking. In Section 4 we illustrate our theoretical results by numerical tests and discuss an application of the proposed ranking scheme in diabetes technology.

2. Problem setting

Let the inputs x be taken from a compact domain or a manifold X in the Euclidean space $\mathbb{R}^d$ and the ranking output space is $Y = [-M, M] \subset \mathbb{R}$. The input x and the output y are assumed to be related by a conditional probability distribution $\rho(y|x)$ of y given x. Moreover, the input x is also assumed to be random and governed by an unknown marginal probability ρ_X on X so that there is an unknown probability distribution $\rho(x, y) = \rho_X(x)\rho(y|x)$ on the sample space $Z = X \times Y$ from which the data forming the training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ are drawn independently. We are interested in synthesizing a function $y = f(x)$ that will simulate the relation between the inputs x and the corresponding outputs y. More precisely, the ranking problem is to learn from $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ a ranking function $f = f_{\mathbf{z}} : X \to Y$.

For given true ranks y and y' of the inputs $x, x' \in X$ the value

$$\left(y - y' - \left(f(x) - f(x')\right)\right)^2$$

is interpreted as the *magnitude-preserving least squares loss* of a ranking function f (see Agarwal and Niyogi (2009), Chen (2012), Cortes, Mohri, and Rastogi (2007) and Ying and Zhou (2015)). Then the quality of a ranking function f can be measured by the expected risk

$$\mathcal{E}(f) = \int_Z \int_Z \left(y - y' - \left(f(x) - f(x')\right)\right)^2 d\rho(x, y) d\rho(x', y').$$

Let $\mathcal{F}_\rho$ be a set of functions minimizing the risk $\mathcal{E}(f)$. As it has been noticed in Chen (2012) and Ying and Zhou (2015), $\mathcal{F}_\rho$ contains the target function

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X,$$

also known in learning theory as the *regression function*. It is easy to observe, that the target function is not unique, for instance $f_\rho(x) + c \in \mathcal{F}_\rho$ for each $c \in \mathbb{R}$.

The ideal estimator $f_\rho(x)$ cannot be found in practice, because the conditional probability distribution $\rho(y|x)$ is unknown. Therefore, the goal might be to find f minimizing the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ over some hypothesis space $\mathcal{H} \in L_2(X, \rho_X)$. A widely used choice of such a space is a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H} = \mathcal{H}_K$, associated with a kernel $K : X \times X \to \mathbb{R}$.

Observe that from the very definition of $\mathcal{E}(f)$ and f_ρ it follows that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \int_{X\times X} \left[(f(x) - f(x')) - (f_\rho(x) - f_\rho(x'))\right]^2 \times d\rho_X(x) d\rho_X(x'). \tag{1}$$

Indeed,

$$\begin{aligned}\mathcal{E}(f) - \mathcal{E}(f_\rho) &= \int_{Z\times Z} \left(y - y' - \left(f(x) - f(x')\right)\right)^2 d\rho(x, y) d\rho(x', y') \\ &\quad - \int_{Z\times Z} \left(y - y' - \left(f_\rho(x) - f_\rho(x')\right)\right)^2 d\rho(x, y) d\rho(x', y') \\ &= \int_X \int_Y \int_X \int_Y [2y - 2y' - f_\rho(x) + f_\rho(x') - f(x) + f(x')] \\ &\quad \times [(f_\rho(x) - f_\rho(x')) - (f(x) - f(x'))] \\ &\quad \times d\rho(y'|x') d\rho(y|x) d\rho_X(x') d\rho_X(x) \\ &= \int_{X\times X} [(f_\rho(x) - f_\rho(x')) - (f(x) - f(x'))]^2 d\rho_X(x) d\rho_X(x').\end{aligned}$$

Consider the space $L_2(X^2, \rho_{X^2})$ of square-integrable functions $g(x, x')$ with respect to the product measure $d\rho_{X^2}(x, x') = d\rho_X(x) d\rho_X(x')$ on $X^2 = X \times X$, and the operators $\Delta : L_2(X, \rho_X) \to L_2(X^2, \rho_{X^2})$, $\mathcal{D}_K : \mathcal{H}_K \to L_2(X^2, \rho_{X^2})$ such that

$$(\Delta f)(x, x') = f(x) - f(x'),$$
$$(\mathcal{D}_K f)(x, x') = \langle K_x - K_{x'}, f \rangle_{\mathcal{H}_K} = f(x) - f(x'),$$

where $K_x = K(x, \cdot)$, $K_{x'} = K(x', \cdot)$, and we use the reproducing property $f(x) = \langle K_x, f \rangle_{\mathcal{H}_K}$. Then in view of (1) the minimization of the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ can be written as the least squares problem

$$\|\mathcal{D}_K f - \Delta f_\rho\|^2_{L_2(X^2, \rho_{X^2})} \to \min$$