8 On Kernel Target Alignment

Nello Cristianini¹, Jaz Kandola², Andre Elisseeff³, and John Shawe-Taylor⁴

1. Department of Statistics, University of California at Davis, nello@support-vector.net
2. Merrill Lynch, Quantitative Analytics Division, 2 King Edward Street, London EC1A 1HQ, Jasvinder_Kandola@ml.com
3. IBM Research Lab, Saumerstrasse 4, CH-8803 Rueschlikon, Switzerland, AEL@zurich.ibm.com
4. School of Electronics and Computer Science, University of Southampton, UK, jst@ecs.soton.ac.uk

Abstract

Kernel based methods are increasingly being used for data modelling because of their conceptual simplicity and outstanding performance on many tasks. However, in practice the kernel function is often chosen using trial-and-error heuristics. In this paper we address the problem of measuring the degree of agreement between a kernel and a learning task. We propose a quantity to capture this notion, which we call alignment. We study its theoretical properties, and derive a series of simple algorithms for adapting a kernel to the targets. This produces a series of novel methods for both transductive and inductive inference, kernel combination and kernel selection for both classification and regression problems that are computationally feasible for large problems. The algorithms are tested on publicly available datasets and are shown to exhibit good performance.

8.1 Introduction

The development of methods that learn from data has been motivated by the need to discover implicit and non-trivial relationships existing in datasets. Kernel machines [9,19,13] work by first embedding the data into a feature space, and then by using linear algorithms to detect patterns in the images of the data. It is of crucial importance in such algorithms that the correct choice of kernel be made, since different kernels will generate different structures in this embedding space. As such being able to assess the quality of an embedding is an important problem from both a
theoretical and practical point of view, and one that has so far been treated only marginally.

In this paper we attempt to address this issue by proposing a measure of similarity between different embeddings of a set of data points. This quantity is called the alignment and can be used not only to assess the relationship between the embeddings generated by two different kernels, but also to assess the similarity between the embeddings of a labelled dataset induced by a kernel and that induced by the labels themselves.

The main theoretical contributions of this paper are in the definition of the problem of kernel-target similarity and in the proposal of the ‘alignment’ as a possible way to address it. We show that this quantity has several convenient properties: it can be efficiently computed before any training of the kernel machine takes place, and based only on training data information. A proof is included that shows alignment to be sharply concentrated around its expected value and hence its empirical value to be stable with respect to different splits of the data. This allows us to estimate it from finite samples, and to establish a connection between high alignment and good generalization performance. We also give a novel way to characterize a kernel function in terms of the alignment, and discuss some geometric interpretations of this quantity.

The main practical contribution is in giving a series of novel transductive and inductive learning algorithms in a number of settings. These include classification, regression, clustering and for the task of kernel combination. Particularly promising is the transductive approach, that provides a first step in the direction of ‘nonparametric’ kernel selection, in the sense that instead of choosing a parameterized kernel family and tuning the parameters so as to maximize the alignment (or other measures), we show how to directly adapt the entries of the kernel matrix to obtain a better one.

This paper is structured as follows. In Section 8.2 we review some basic concepts about kernel methods and provide a novel alternative characterization of kernel matrices. In Section 8.3 we propose the alignment as a measure of kernel similarity, and we derive some of its theoretical properties including its sharp concentration. In Section 8.5 and 8.6 we present a series of algorithms for kernel combination in both transductive and inductive settings for classification and regression problems.

**Notations** Vectors are generally written in lowercase letters: $y$ stands for the target and is considered either as a finite dimensional vector or as a function from the input space $\mathcal{X}$ into $\mathbb{R}$. In the latter case, we write $y(x)$. 