Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Computer and System Sciences 71 (2005) 333–359
JOURNAL OF COMPUTER AND SYSTEM SCIENCES
www.elsevier.com/locate/jcss

Maximal margin classification for metric spaces

Matthias Hein*, Olivier Bousquet, Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tuebingen, Germany

Received 9 February 2004; received in revised form 30 July 2004

Available online 8 December 2004

Abstract

In order to apply the maximum margin method in arbitrary metric spaces, we suggest to embed the metric space into a Banach or Hilbert space and to perform linear classification in this space. We propose several embeddings and recall that an isometric embedding in a Banach space is always possible while an isometric embedding in a Hilbert space is only possible for certain metric spaces. As a result, we obtain a general maximum margin classification algorithm for arbitrary metric spaces (whose solution is approximated by an algorithm of Graepel et al. (International Conference on Artificial Neural Networks 1999, pp. 304–309)). Interestingly enough, the embedding approach, when applied to a metric which can be embedded into a Hilbert space, yields the support vector machine (SVM) algorithm, which emphasizes the fact that its solution depends on the metric and not on the kernel. Furthermore, we give upper bounds of the capacity of the function classes corresponding to both embeddings in terms of Rademacher averages. Finally, we compare the capacities of these function classes directly.

Keywords: Classification; Maximum margin; Metric spaces; Embedding; Pattern recognition

1. Introduction

Often, the data in real-world problems cannot be expressed naturally as vectors in a Euclidean space. However, it is common to have a more or less natural notion of distance between data points. This distance can often be quantified by a semi-metric (i.e. a symmetric non-negative function which satisfies

* Corresponding author.
E-mail addresses: matthias.hein@tuebingen.mpg.de (M. Hein), olivier.bousquet@tuebingen.mpg.de (O. Bousquet), bernhard.schoelkopf@tuebingen.mpg.de (B. Schölkopf)..

doi:10.1016/j.jcss.2004.10.013

the triangle inequality) or, even better, a metric (a semi-metric which is zero only when the two points are the same).

If the only knowledge available to the statistician is that the data comes from a semi-metric space $(\mathcal{X}, d)$, where $\mathcal{X}$ is the input space and d is the corresponding semi-metric, it is reasonable to assume, for a classification task, that the class labels are somewhat related to the semi-metric. More precisely, since one has to make assumptions about the structure of the data (otherwise no generalization is possible), it is natural to assume that two points that are close (as measured by d) are likely to belong to the same class, while points that are far away may belong to different classes. Another way to express this assumed relationship between class membership and distances is to say that intra-class distances are on average smaller than inter-class distances.

Most classical classification algorithms rely, implicitly or explicitly, on such an assumption. On the other hand, it is not always possible to work directly in the space $\mathcal{X}$ where the data lies. In particular, some algorithms require a vector space structure (e.g. linear algorithms) or at least a feature representation (e.g. decision trees). So, if $\mathcal{X}$ does not have such a structure (e.g. if the elements of $\mathcal{X}$ are DNA sequences of variable length, or descriptions of the structure of proteins), it is typical to construct a new representation (usually as vectors) of the data. In this process, the distance between the data, that is the (semi)-metric, is usually altered. But with the above assumptions on the classification task this change means that information is lost or at least distorted.

It is thus desirable to avoid any distortion of the (semi)-metric in the process of constructing a new representation of the data. Or at least, the distortion should be consistent with the assumptions. For example, a transformation which leaves the small distances unchanged and alters the large distances, is likely to preserve the relationship between distances and class membership. We later propose a precise formulation of this type of transformation.

Once the data is mapped into a vector space, there are several possible algorithms that can be used. However, there is one heuristic which has proven valuable both in terms of computational expense and in terms of generalization performance, it is the maximum margin heuristic. The idea of maximum margin algorithms is to look for a linear hyperplane as the decision function which separates the data with maximum margin, i.e. such that the hyperplane is as far as possible from the data of the two classes. This is sometimes called the hard margin case. It assumes that the classes are well separated. In general one can one always deal with the inseparable case by introducing slack variables, which corresponds to the soft margin case.

Our goal is to apply this heuristic to $(\mathcal{X}, d)$, the (semi)-metric input space directly. To do so, we proceed in two steps: we first embed $\mathcal{X}$ into a Banach space (i.e. a normed vector space which is complete with respect to its norm) and look for a maximum margin hyperplane in this space. The important part being that the embedding we apply is isometric, that is, all distances are preserved.

We explain how to construct such an embedding and show that the resulting algorithm can be approximated by the Linear Programming Machine proposed by Graepel et al. [8]. We also propose to use as a "pre-processing'' step, a transformation of the metric which has the properties mentioned above (i.e. leaving the small distances unaltered and affecting the large ones) which may remove the unnecessary information contained in large distances and hence give a better result when combined with the above-mentioned algorithm.

Embedding the data isometrically into a Banach space is convenient since it is possible for any metric space. But as we will show it has also the disadvantage that the obtained maximum margin algorithm cannot be directly implemented and has to be approximated. It may thus be desirable that the space into