Maximal Margin Classification for Metric Spaces

Matthias Hein and Olivier Bousquet

Max Planck Institute for Biological Cybernetics
Spemannstr. 38
72076 Tuebingen, Germany
{matthias.hein, olvier.bousquet}@tuebingen.mpg.de

Abstract. In this article we construct a maximal margin classification algorithm for arbitrary metric spaces. At first we show that the Support Vector Machine (SVM) is a maximal margin algorithm for the class of metric spaces where the negative squared distance is conditionally positive definite (CPD). This means that the metric space can be isometrically embedded into a Hilbert space, where one performs linear maximal margin separation. We will show that the solution only depends on the metric, but not on the kernel. Following the framework we develop for the SVM, we construct an algorithm for maximal margin classification in arbitrary metric spaces. The main difference compared with SVM is that we no longer embed isometrically into a Hilbert space, but a Banach space. We further give an estimate of the capacity of the function class involved in this algorithm via Rademacher averages. We recover an algorithm of Graepel *et al.* [6].

1 Introduction

It often occurs that real-world data does not have a natural vector space structure. It is rather common, however that a natural (semi)-metric exists on this data that measures pairwise dissimilarities. For the task of classification in a (semi)-metric space $(\mathcal{X}, d)$, where $\mathcal{X}$ is a set and d a (semi)-metric on $\mathcal{X}$, in the absence of other information or prior knowledge, we can only use the metric. Therefore all algorithms for classification on metric spaces assume that the metric is somehow adapted to the classification task. This means heuristically that the inner class distances should be low compared with the distance between the two classes. If the metric fulfills these conditions, then it reveals valuable information for the classification problem. Therefore any kind of transformation of $\mathcal{X}$ that distorts this distance structure (the only information we have on the data) should be avoided.
On the other hand the idea of maximal margin separation of two sets, which is equivalent to finding the distance between the convex hulls of the two sets, is a very appealing geometric concept. Obviously we cannot do this for all (semi)-metric spaces, because in general no linear structure is available. Therefore we employ isometric embeddings into linear spaces which, on the one hand, preserve the distance structure of the input space, which can be seen as our prior

knowledge on the data, and on the other hand, provide us with the linearity to do maximal margin separation.
In the first section we start by reviewing the formulation of the SVM. Then we turn around the normal viewpoint on SVM, and show that the SVM is actually a maximal-margin algorithm for (semi)-metric spaces. This means that one can start with a (semi)-metric space, where the negative squared metric is CPD, then embed this (semi)-metric space isometrically into a Hilbert space, and then use the linear structure of the Hilbert space to do maximal margin separation. We emphasize this point by showing that any CPD kernel, which includes any positive definite (PD) kernel, can be expressed as a sum of the squared (semi)-metric and some function, and only the (semi)-metric enters the solution of the SVM. We also show that the optimization problem and the solution can be written in terms of the (semi)-metric of the input space only.
Unfortunately only the class of metric spaces where the negative square of the distance is CPD can be used in the SVM. We thus provide a generalization of the maximum margin principle to arbitrary metric spaces. The main idea is that any (semi)-metric space can be embedded isometrically into a Banach space. Since a Banach space is linear and the concept of maximal margin separation between convex sets can be extended to Banach spaces [2, 12], it is then possible to write down a maximal margin classification algorithm, which can be applied to arbitrary (semi)-metric spaces. However the solution of this algorithm differs from the SVM solution if applied to the same metric space.
Next, we compare semi-metric spaces to metric spaces with respect to classification. It turns out that a semi-metric space can be seen as a space where certain invariances are realized. Therefore using a semi-metric means that one implicitly uses prior knowledge about invariances of the data. If the data does not share this invariance property, the use of semi-metrics may lead to a bad classifier.
In the end we compare both algorithms in terms of their generalization ability and other properties. In particular, we show that the capacity of the class of functions generated by the proposed embedding is directly related to the metric entropy of the input space.

2 SVM as a Maximal Margin Algorithm for Metric Spaces

2.1 The RKHS and the Formulation of the SVM

In this section we construct the Reproducing Kernel Hilbert Space (RKHS) and state the problem of the SVM; see [11] for an overview on kernel methods. We first need the definition of the two classes of kernels that are used in the SVM:

Definition 1. *A real valued function k on $\mathcal{X} \times \mathcal{X}$ is positive definite (resp. conditionally positive definite) if and only if k is symmetric and*

$$\sum_{i,j}^{n} c_i c_j k(x_i, x_j) \geq 0, \tag{1}$$

for all $n \in \mathbb{N}$, $x_i \in \mathcal{X}, i = 1, ..., n$, *and for all* $c_i \in \mathbb{R}, i = 1, ..., n$, *(resp. for all* $c_i \in \mathbb{R}, i = 1, ..., n$, *with* $\sum_i^n c_i = 0$*).*

Notice that a PD kernel is always CPD.

A PD kernel allows the construction of a RKHS $\mathcal{H}$ in the following way:

1. Define a feature map $\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$, $x \to \Phi_x = k(x, .)$
2. Turn it into a vector space by considering all finite linear combinations of Φ_{x_i}: $f = \sum_{i=1}^n \alpha_i \Phi_{x_i}$
3. Turn it into a pre-Hilbert space $\tilde{\mathcal{H}}$ by introducing the dot product: $< \Phi_x, \Phi_y >= k(x, y)$
4. Turn it into a Hilbert space $\mathcal{H}$ by completing $\tilde{\mathcal{H}}$

With these definitions, we can describe the SVM algorithm as follows. The input space $\mathcal{X}$ is mapped into a Hilbert space $\mathcal{H}$ via the feature map Φ, and a maximal margin hyperplane is searched for in this space. Hyperplanes correspond to linear continuous functionals on the Hilbert space. The margin of such a hyperplane is defined as twice the distance from the hyperplane to the closest data point. The margin of the optimal hyperplane is equal to the distance between the convex hulls of the two classes [2, 12]. Due to Riesz theorem, each continuous linear functional can be considered as a vector of the Hilbert space (the normal vector of the corresponding hyperplane).

Given a training set $\{(x_i, y_i)\}_{i=1..n}, x_i \in \mathcal{X}, y_i \in \{-1, +1\}$, the optimization problem corresponding to the maximum margin hyperplane can be written as

$$\min_{\alpha} \left\| \sum_{i:y_i=+1} \alpha_i \Phi_{x_i} - \sum_{i:y_i=-1} \alpha_i \Phi_{x_i} \right\|_{\mathcal{H}}^2$$
$$\text{s.th} : \sum_{i:y_i=+1} \alpha_i = \sum_{i:y_i=-1} \alpha_i = 1, \quad \alpha_i \geq 0,$$

which is equivalent to

$$\min_{\alpha} \left\| \sum_i y_i \alpha_i \Phi_{x_i} \right\|_{\mathcal{H}}^2 = \sum_{i,j} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$
$$\text{s.th} : \sum_i y_i \alpha_i = 0, \quad \sum_i \alpha_i = 2, \quad \alpha_i \geq 0,$$

where the normal vector to the hyperplane is given by

$$w = \sum_{i=1}^n \alpha_i y_i \Phi_{x_i} .$$

2.2 The Input Space as a (Semi)-Metric Space

Let us first put the standard point of view on SVM like this:

$$\mathcal{X} \xrightarrow{kernel\, k} \mathcal{H} \longrightarrow \text{maximal margin separation} \tag{2}$$

In this section we show by using results which date back to Schoenberg that there exists an equivalent point of view, which allows us to generalize later on to arbitrary metric spaces. It can be summarized with the following scheme:

$$(\mathcal{X}, d) \xrightarrow{isometric} \mathcal{H} \longrightarrow \text{maximal margin separation} \tag{3}$$

Recall that a semi-metric is a non-negative symmetric function, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which satisfies the triangle inequality and $d(x, x) = 0$ (it is a metric if $d(x, y) = 0 \Rightarrow x = y$).
First we note that through the previous construction of the RKHS, we can induce a semi-metric on $\mathcal{X}$ by the following definition:

$$d^2(x, y) := \|\Phi_x - \Phi_y\|^2_{\mathcal{H}} = k(x, x) + k(y, y) - 2\,k(x, y)\,. \tag{4}$$

Note that d will be a metric if Φ_x is injective (and a semi-metric otherwise). A simple example of a kernel whose feature map is not injective is $k(x, y) = \langle x, y\rangle^2$. We will consider the difference between a metric and semi-metric with respect to classification in a later section.

The next proposition can be found in a different form in Berg et al. (see Proposition 3.2 of [3]). We have rewritten it in order to stress the relevant parts for the SVM.

Proposition 1. *Let $\mathcal{X}$ be the input space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a CPD kernel. Then the function d, defined as*

$$d(x, y) = \sqrt{k(x, x) + k(y, y) - 2\,k(x, y)}\,, \tag{5}$$

is a semi-metric on $\mathcal{X}$ such that $-d^2$ is CPD. All CPD kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are generated by a (semi)-metric d (with $-d^2$ CPD) in the sense that there exists a function $g : \mathcal{X} \to \mathbb{R}$ such that

$$k(x, y) = -\frac{1}{2}d^2(x, y) + g(x) + g(y)\,, \tag{6}$$

and any kernel of this form induces the semi-metric d via Equation (5).

This proposition states that semi-metrics d, where $-d^2$ is CPD, are up to a function f equivalent to the whole class of CPD kernels. Next, one can show that the obtained metric space can be isometrically embedded into a Hilbert space (see also Proposition 3.2 of [3]).

Proposition 2. *Let $(\mathcal{X}, d)$ be the semi-metric space defined in Proposition 1.*

(i) It can be isometrically embedded into a Hilbert space $\mathcal{H}$;
(ii) if k is bounded, $\mathcal{H}$ can be continuously embedded into $(C_b(\mathcal{X}), \|\cdot\|_\infty)$.

Moreover, the class of semi-metric spaces defined in Proposition 1 consists of all metric spaces that can be embedded isometrically into a Hilbert space, which is a result of Schoenberg [9]. Schoenberg proved this theorem already in 1938 and introduced the notion of PD and CPD functions. We are getting back to the roots of the kernel industry.

Theorem 1. *A necessary and sufficient condition for a (semi)-metric space $(\mathcal{X}, d)$ to be isometrically embeddable into a Hilbert space is that $\tilde{k}(x,y) = -\frac{1}{2}d(x,y)^2$ is CPD.*

We now try to show the relevance of these results for the SVM. This theorem together with Proposition 1 gives the full equivalence of the standard (2) and our (3) point of view on SVM. This can be summarized as follows: defining a CPD kernel on the input space $\mathcal{X}$ is equivalent to defining a unique (semi)-metric d on the input space $\mathcal{X}$ via (5); and in the other direction any (semi)-metric d on $\mathcal{X}$, where $-d^2$ is CPD, defines a non-unique PD kernel via (6) and (7), such that $(\mathcal{X}, d)$ can be embedded isometrically into the corresponding RKHS.
We will also use these results to show in the next section that the SVM classifier only depends on the metric, so that all the kernels of the form (6) are equivalent from the SVM point of view.
In the rest of this section we give the proofs of the propositions.

Proof (Proposition 1). If k is CPD but not PD, we consider for an arbitrary $x_0 \in \mathcal{X}$,

$$\tilde{k}(x,y) := k(x,y) - k(x,x_0) - k(x_0,y) + k(x_0,x_0)\,. \tag{7}$$

This kernel is PD if and only if k is CPD (see [3]) and $\tilde{k}(x,x)+\tilde{k}(y,y)-2\tilde{k}(x,y) = k(x,x) + k(y,y) - 2k(x,y)$, so that $\tilde{k}$ defines the same semi-metric d as k (via Equation (5)). Note that $k(x,y) = \hat{k}(x,y) + g(x) + g(y)$ is CPD, if $\hat{k}$ is CPD:

$$\begin{aligned}\sum_{i,j} c_i c_j k(x_i,x_j) &= \sum_{i,j} c_i c_j \hat{k}(x_i,x_j) + 2\sum_j c_j \sum_i c_i g(x_i) \\ &= \sum_{i,j} c_i c_j \hat{k}(x_i,x_j) \geq 0,\end{aligned}$$

where the second term vanishes because $\sum_i c_i = 0$.
Thus from (5) we get that $-d^2$ is CPD with $f(x) = -\frac{1}{2}k(x,x)$. On the other hand, if we start with a semi-metric d, where $-\frac{1}{2}d^2(x,y)$ is CPD, then k defined by (6) is CPD and k induces d as a semi-metric via (5). Now if two CPD kernels k and $\hat{k}$ induce the same (semi)-metric, then they fulfill $k(x,y) = \hat{k}(x,y) + \frac{1}{2}[(k(x,x)-\hat{k}(x,x))+(k(y,y)-\hat{k}(y,y))]$. Thus they differ by a function $g : \mathcal{X} \to \mathbb{R}$ with $g(.) = \frac{1}{2}(k(.,.) - \hat{k}(.,.))$. □

Proof (Proposition 2). We have shown in the proof of Proposition 1 that each CPD kernel k defines a PD kernel $\tilde{k}$ via (7), which induces the same (semi)-metric. With the PD kernel $\tilde{k}$ we define a reproducing kernel Hilbert space $\mathcal{H}$ as above, with associated feature map Φ ($\Phi_x = \tilde{k}(x,\cdot)$). It trivially defines an isometry.
We note that the kernel is always continuous with respect to the (semi)-metric it induces:

$$\begin{aligned}|k(x,y) - k(x',y')| &= |< k_x, k_y - k_{y'} > + < k_x - k_{x'}, k_{y'} >| \\ &\leq \|k_x\|\|k_y - k_{y'}\| + \|k_x - k_{x'}\|\|k_{y'}\| \\ &= \sqrt{k(x,x)}\sqrt{d(y,y')} + \sqrt{k(y',y')}\sqrt{d(x,x')}.\end{aligned}$$

Furthermore, if the kernel is bounded, then for any $f \in \mathcal{H}$,

$$|f(x)| = |\langle f, k(x,\cdot)\rangle| \leq \|f\|_{\mathcal{H}} \sqrt{k(x,x)}$$

so that f is bounded, and similarly

$$|f(x) - f(y)| \leq \|f\| \, \|k(x,\cdot) - k(y,\cdot)\| = \|f\| \, d(x,y),$$

hence f is continuous. □

2.3 Formulation of the SVM in Terms of the (Semi)-Metric

It was already recognized by Schölkopf [10] that the SVM relies only on distances in the RKHS. This can be seen directly from the optimization problem (2), where we minimize the euclidean distance of the convex hulls in $\mathcal{H}$, which is translation invariant. Schölkopf showed that this implies one can use the bigger class of CPD kernels in SVM. One can show this by directly plugging in the expression of the PD kernel in terms of a CPD kernel from (7) into the optimization problem. All terms except the CPD kernel $k(x,y)$ part cancel out because of the constraints.

We have shown in the last section that a (semi)-metric lies at the core of every CPD kernel, and that there exists a whole class of CPD kernels which induce the same (semi)-metric on $\mathcal{X}$. Applying the results of the last section we go one step further and show that the SVM is a maximal-margin algorithm for a certain class of (semi)-metric spaces.

Theorem 2. *The SVM method can be applied to the class of (semi)-metric spaces $(\mathcal{X}, d)$, where $-d^2$ is CPD. The (semi)-metric space $(\mathcal{X}, d)$ is embedded isometrically via the corresponding positive definite kernel into a Hilbert space. Using the linear structure of the Hilbert space, the two sets of points, corresponding to the two classes, are linearly separated so that the margin between the two sets is maximized. The distance between the convex hulls of the two classes is twice the margin. The solution of the SVM does not depend on the specific isometric embedding Φ, nor on the corresponding choice of the kernel. The optimization problem and the solution can be completely expressed in terms of the (semi)-metric d of the input space,*

$$\min_{\alpha} \left\| \sum_i y_i \alpha_i \Phi_{x_i} \right\|_{\mathcal{H}}^2 = -\frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j d^2(x_i, x_j)$$
$$s.th: \sum_i y_i \alpha_i = 0, \quad \sum_i \alpha_i = 2, \quad \alpha_i \geq 0.$$

The solution can be written as

$$f(x) = -\frac{1}{2} \sum_i y_i \alpha_i d^2(x_i, x) + c.$$

Proof. By combining the Proposition 1 and the theorem of Schoenberg, we showed the equivalence of the standard view on SVM and the view of an isometric embedding of the (semi)-metric space $(\mathcal{X}, d)$ into a Hilbert space $\mathcal{H}$. Therefore the SVM is restricted to metric spaces $(\mathcal{X}, d)$, where $-d^2$ is CPD. The statement about the equivalence of maximal-margin separation and the distance between the convex hulls of the two classes can be found in [2, 12]. Now the expression of the optimization problem of the SVM in terms of the (semi)-metric follows from (6);

$$\begin{aligned}\left\|\sum_i y_i\alpha_i\Phi_{x_i}\right\|^2_{\mathcal{H}} &= \sum_{i,j} y_iy_j\alpha_i\alpha_j k(x_i,x_j)\\ &= \sum_{i,j} y_iy_j\alpha_i\alpha_j[-\frac{1}{2}d^2(x_i,x_j)+g(x_i)+g(x_j)]\\ &= -\frac{1}{2}\sum_{i,j} y_iy_j\alpha_i\alpha_j d^2(x_i,x_j),\end{aligned}$$

where f drops out due to the constraint $\sum_i y_i\alpha_i = 0$.
The solution expressed in terms of a CPD kernel k can also be expressed in terms of the (semi)-metric by using (6):

$$\begin{aligned}f(x) &= \sum_i y_i\alpha_i k(x_i,x)+b = \sum_i y_i\alpha_i[-\frac{1}{2}d(x_i,x)^2+g(x_i)+g(x)]\\ &= -\frac{1}{2}\sum_i y_i\alpha_i d^2(x_i,x)+c,\end{aligned}$$

where again $\sum_i y_i\alpha_i g(x)$ drops out and $c = b + \sum_i y_i\alpha_i g(x_i)$, but c can also be directly calculated with the average value of $b = y_j + \frac{1}{2}\sum_i y_i\alpha_i d^2(x_i,x_j)$, where j runs over all indices with $\alpha_j > 0$. Since neither the specific isometric embedding Φ nor a corresponding kernel k enter the optimization problem or the solution, the SVM only depends on the (semi)-metric. □

The kernel is sometimes seen as a similarity measure. The last theorem, however, shows that this property of the kernel does not enter the algorithm. On the contrary the (semi)-metric as a dissimilarity measure of the input space only enters the algorithm. Nevertheless it seems to be easier to construct a CPD kernel than a function $d(x, y)$, where d is a (semi)-metric and $-d^2$ is CPD, but one should remain aware that only the induced (semi)-metric has an influence on the solution, and therefore compare two different kernels through their induced (semi)-metrics.
One can use the high ambiguity in the kernel to chose from the whole class of kernels which induce the same (semi)-metric (6) that which is computationally the cheapest, because the solution does not change as is obvious from the last theorem. As a final note we would like to add that the whole argumentation on the isometric embedding of the (semi)-metric space into a Hilbert space also applies to the soft-margin-formulation of the SVM. The reformulation in terms of reduced convex hulls is a little bit tricky, and we refer to [2, 12] for this issue.

3 Maximal Margin Algorithm for Arbitrary (Semi)-Metric Spaces

The maximal margin algorithm where the space one embeds the data isometrically is a Hilbert space, which is equivalent to the SVM, is limited to a subclass of all metric spaces. In this section we will treat arbitrary metric spaces trying to follow the same steps described at the end of the last section. We first define an isometric embedding of an arbitrary metric space into a Banach space. We then use the fact that in Banach spaces the problem of a maximal margin hyperplane is equivalent to finding the distance between the convex hulls. With this property we are able to formulate the problem and discuss the algorithm. The scheme we use can be stated as follows

$$(\mathcal{X}, d) \xrightarrow{isometric} (\bar{D}, \|.\|_\infty) \subset (C_b(\mathcal{X}), \|.\|_\infty) \longrightarrow \text{maximal margin separation}$$

where $\bar{D}$ is a Banach space of (continuous and bounded) functions defined on $\mathcal{X}$ (see definitions below).

3.1 Isometric embedding of a general metric space into a Banach space

In this section we construct a pair of dual Banach spaces. The metric space $\mathcal{X}$ will be isometrically embedded into the first one, and the second one will be used to define continuous linear functionals (i.e. hyperplanes).

Let $(\mathcal{X}, d)$ be a compact[1] metric space and denote by $C_b(\mathcal{X})$ the Banach space of continuous and bounded functions on $\mathcal{X}$ endowed with the supremum norm. The topological dual of $C_b(\mathcal{X})$ is the space of Baire measures $\mathcal{M}(\mathcal{X})$ with the measure norm $\|\mu\| = \int_{\mathcal{X}} d\mu_+ - \int_{\mathcal{X}} d\mu_-$ (where μ_+ and μ_- are respectively the positive and negative parts of μ).
Consider an arbitrary $x_0 \in \mathcal{X}$ and define the following maps

$$\begin{aligned} \Phi : \mathcal{X} &\to \mathbb{R}^{\mathcal{X}} \\ x &\mapsto \Phi_x := d(x, \cdot) - d(x_0, \cdot) \end{aligned} \qquad \text{and} \qquad \begin{aligned} \Psi : \mathcal{X} &\to \mathbb{R}^{\mathcal{X}} \\ x &\mapsto \Psi_x := d(\cdot, x) - d(x_0, x)\,. \end{aligned}$$

Let $D = \operatorname{span}\{\Phi_x : x \in \mathcal{X}\}$ and $E = \operatorname{span}\{\Psi_x : x \in \mathcal{X}\}$ be the linear spans of the images of the maps Φ and Ψ.
We will show that Φ defines an isometric embedding of the metric space $\mathcal{X}$ into the closure $\bar{D}$ of D (with respect to the infinity norm). Moreover, $\bar{D}$ is a Banach space whose dual is isometrically isomorphic to (hence can be identified with) the completion $\bar{E}$ of E with respect to the norm

$$\|e\|_E = \inf \left\{ \sum_{i \in I} |\beta_i| : e = \sum_{i \in I} \beta_i \psi_{x_i},\, x_i \in \mathcal{X},\, |I| < \infty \right\}.$$

The following results formalize the above statements.

[1] Compactness is needed for the analysis but the algorithm we present in the next section works without this assumption since it performs an approximation on a finite set.

Lemma 1. *Φ is an isometry from $(\mathcal{X}, d)$ into the Banach space $(\bar{D}, \|\cdot\|_\infty) \subset (C_b(\mathcal{X}), \|\cdot\|_\infty)$.*

Proof. We have $\|\Phi_x\|_\infty \leq d(x, x_0) < \infty$ and $|\Phi_x(y) - \Phi_x(y')| \leq |d(x, y) - d(x, y')| + |d(x_0, y) - d(x_0, y')| \leq 2d(y, y')$, so that $\Phi_x \in C_b(\mathcal{X})$. In addition $\|\Phi_x - \Phi_y\|_\infty = \|d(x, \cdot) - d(y, \cdot)\|_\infty \leq d(x, y)$ and the supremum is attained at x and y. Hence, Φ is an isometry from $(\mathcal{X}, d)$ into $(D, \|\cdot\|_\infty)$ which is a subspace of $C_b(\mathcal{X})$. Defining $\bar{D}$ as the closure of D in $C_b(\mathcal{X})$ which is a Banach space yields that $\bar{D}$ is complete. □

Note that, as an isometry, Φ is continuous, and x_0 is mapped to the origin of D.

Lemma 2. *$\|\cdot\|_E$ is a norm on E.*

Proof. It is easy to see that $\|\cdot\|_E$ satisfies the properties of a semi-norm. To prove that it is a norm, consider $e \in E$ such that $\|e\|_E = 0$. Then there exist sequences $(I_n), (\beta_{i,n})$ and $x_{i,n}$ such that $e = \sum_{i \in I_n} \beta_{i,n} \Psi_{x_{i,n}}$ and $\sum_{i \in I_n} |\beta_{i,n}| \to 0$. As a consequence, for any $x \in \mathcal{X}$, $|e(x)| = |\sum_{i \in I_n} \beta_{i,n} \Psi_{x_{i,n}}(x)| \leq d(x, x_0) \sum_{i \in I_n} |\beta_{i,n}|$, so that taking the limit $n \to \infty$ we obtain $e(x)$. This proves $e \equiv 0$ and concludes the proof. □

As a normed space, E can be completed with respect to the norm $\|\cdot\|_E$ into a Banach space $\bar{E}$ with extended norm $\|\cdot\|_{\bar{E}}$. Let $\bar{D}'$ be the topological dual of $\bar{D}$ with dual norm $\|\cdot\|_{\bar{D}'}$.

Theorem 3. *$(\bar{E}, \|\cdot\|_{\bar{E}})$ is isometrically isomorphic to $(\bar{D}', \|\cdot\|_{\bar{D}'})$.*

Proof. Let $\bar{D}^\perp = \{d' \in \bar{D}' : \langle d', d \rangle = 0, \forall\, d \in D\}$ and consider the space $\mathcal{M}(\mathcal{X})/\bar{D}^\perp$ of equivalence classes of measures that are identical on the subspace $\bar{D}$ and endow this space with the quotient norm $\|\tilde{\mu}\| = \inf\{\|\mu\| : \mu \in \tilde{\mu}\}$. Then by theorem 4.9 of [8] $(\bar{D}', \|\cdot\|_{\bar{D}'})$ is isometrically isomorphic to $(\mathcal{M}(\mathcal{X})/\bar{D}^\perp, \|\cdot\|)$. Recall that the span of measures with finite support is dense in $\mathcal{M}(\mathcal{X})$, so the same is true for the quotient space $\mathcal{M}(\mathcal{X})/\bar{D}^\perp$. The linear map $\sigma : E \to \text{span}\{\delta_x : x \in \mathcal{X}\}/\bar{D}^\perp$ defined as $\sigma(\Psi_x) = \delta_x|_D$ induces an isometric isomorphism between E and $\text{span}\{\delta_x : x \in \mathcal{X}\}/\bar{D}^\perp$, which can be extended to the closure of these spaces. □

3.2 Duality of Maximal Margin Hyperplanes and Distance of Convex Hulls in Banach Spaces

We have stated in the beginning that the two problems of finding the distance between two disjoint convex hulls and finding a maximal margin hyperplane are equivalent for Banach spaces. This can be seen by the following theorem (see [12] for a proof), where we define $co(T) = \{\sum_{i \in I} \alpha_i x_i | \sum_{i \in I} \alpha_i = 1,\, x_i \in T,\, \alpha_i \in \mathbb{R}^+,\, |I| < \infty\}$.

Theorem 4. *Let T_1 and T_2 be two finite sets of points in a Banach space B then if $co(T_1) \cap co(T_2) = \emptyset$*

$$d(co(T_1), co(T_2)) = \inf_{y \in co(T_1), z \in co(T_2)} \|y - z\| = \sup_{x' \in B'} \frac{\inf_{y \in T_1, z \in T_2} < x', y - z >}{\|x'\|}.$$

We now rewrite the right term by using the definition of the infimum:

$$\begin{aligned}&\inf_{x' \in B', c, d} \quad \frac{\|x'\|}{c-d} \\ &\text{subject to:} \quad x'(y) \geq c, \quad \forall\, y \in T_1, \quad x'(z) \leq d, \quad \forall\, z \in T_2.\end{aligned}$$

Now subtract $-\frac{c+d}{2}$ from both inequalities, and define the following new quantities: $b = \frac{c+d}{d-c}$, $w' = \frac{2}{c-d}x', T = T_1 \cup T_2$. Then one gets the standard form:

$$\begin{aligned}&\min_{w' \in B', b} \|w'\| \\ &\text{subject to:} \quad y_i(w'(x_i) + b) \geq 1 \quad \forall x_i \in T = T_1 \cup T_2.\end{aligned} \tag{8}$$

3.3 The Algorithm

We now plug our isometric embedding into the equation (8) to get the optimization problem for maximal margin classification in arbitrary (semi)-metric spaces:

$$\begin{aligned}&\min_{w' \in \bar{D}', b \in \mathbb{R}} \|w'\| \\ &\text{subject to:} \quad y_j(w'(\Phi_{x_j}) + b) \geq 1 \quad \forall x_j \in T.\end{aligned}$$

We are using the isometric isomorphism between $\bar{D}'$ and $\bar{E}$ to state it equivalently in $\bar{E}$. By density of E in $\bar{E}$ and by continuity of the norm and of the duality-product, the minimum on $\bar{E}$ can be replaced by an infimum on E:

$$\begin{aligned}&\inf_{e \in E, b} \|e\| = \inf_{m \in \mathbb{N},\, x_1, \ldots, x_m \in \mathcal{X}^m,\, b} \quad \sum_{i=1}^{m} |\beta_i| \\ &\text{s.t. } y_j(\sum_{i=1}^{m} \beta_i \psi_{x_i}(\Phi_{x_j}) + b) = y_j(\sum_{i=1}^{m} \beta_i (d(x_j, x_i) - d(x_0, x_i)) + b) \geq 1 \; \forall x_j \in T.\end{aligned}$$

Notice that the infimum may not be attained in E. Unlike in the SVM case there seems to be no guarantee such as a representer theorem that the solution can be expressed in terms of points in the training set only.
In order to make the problem computationally tractable, we have to restrict the problem to a finite dimensional subspace of E. A simple way to do this is to consider only the subspace of E generated by a finite subset $Z \in \mathcal{X}, |Z| = m$. We are free to choose the point x_0, so we choose it as $x_0 = z_1, z_1 \in Z$. Since the problem stated in Theorem 4 is translation invariant, this choice has no influence on the solution. This leads to the following optimization problem:

$$\begin{aligned}&\min_{\beta_i, b} \sum_{i=1}^{m} |\beta_i| \\ &\text{subject to:} \quad y_j(\sum_{i=1}^{m} \beta_i (d(x_j, z_i) - d(z_1, z_i)) + b) \geq 1, \quad \forall x_j \in T.\end{aligned}$$

In general, a convenient choice for Z is $Z = T$. In a transduction setting one can use for Z the union of labelled and unlabelled data.
As $\sum_{i=1}^{m} \beta_i d(z_1, z_i)$ does not depend on j, due to translation invariance, we can put it in the constant b and solve the equivalent problem:

$$\min_{\beta_i, c} \sum_{i=1}^{m} |\beta_i|$$
$$\text{subject to:} \quad y_j(\sum_{i=1}^{m} \beta_i \, d(x_j, z_i) + c) \geq 1, \quad \forall x_j \in T.$$

The corresponding decision function is given by

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \beta_i \, d(x, z_i) + c\right).$$

The above optimization problem can be transformed into a linear programming problem, and is easily solvable with standard methods. Note that if we take $Z = T$ we recover the algorithm proposed by Graepel et al. [6]. We also note that it is easily possible to obtain a soft-margin version of this algorithm. In this case there still exists the equivalent problem of finding the distance between the reduced convex hulls [2, 12]. This algorithm was compared to other distance based classifiers by Pekalska et al. in [7] and showed good performance.

Using Theorem 4, we can also formulate the problem (in dual form) as follows

$$\min_{\alpha_i \in \mathbb{R}} \sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{n} y_i \alpha_i d(x, x_i) \right|$$
$$\text{subject to:} \quad \sum_{i=1}^{n} y_i \alpha_i = 0, \, \sum_{i=1}^{n} \alpha_i = 2 \, , \alpha_i \geq 0 \, .$$

Unfortunately, there is no simple relationship between primal (β_i) and dual (α_i) variables which allows to compute the decision function from the α_i. However, it is interesting to notice that the approximation of the primal problem which consists in looking for a solution generated by a finite subset Z corresponds, in dual form, to restricting the supremum to Z only. This means for finite metric spaces the problem can be solved without approximation.

4 Semi-Metric Spaces compared to Metric Spaces for Classification

In the last two sections we made no distinction between semi-metric and metric spaces. In fact there is a connection between both of them which we want to clarify in this section.

Theorem 5. *Let $(\mathcal{X}, d)$ be a (semi)-metric space and $\sim$ be the equivalence relation defined by $x \sim y \Leftrightarrow d(x,y) = 0$. Then $(\mathcal{X}/\sim, d)$ is a metric space, and if $-d^2(x,y)$ is a CPD Kernel and k a PD Kernel on $\mathcal{X}$ which induces d on $\mathcal{X}$, then $-d^2$ is also a CPD Kernel and k a PD kernel on $(\mathcal{X}/\sim, d)$.*

Proof. The property $d(x,y) = 0$ defines an equivalence relation on $\mathcal{X}$, $x \sim y \Longleftrightarrow d(x,y) = 0$. Symmetry follows from the symmetry of d, and transitivity $x \sim y, y \sim z \Rightarrow x \sim z$ follows from the triangle inequality $d(x,z) \leq d(x,y) + d(y,z) = 0$. Then $d(x,y)$ is a metric on the quotient space $\mathcal{X}/\sim$ because all points with zero distance are identified, so

$$d(x,y) = 0 \iff x = y,$$

and obviously symmetry and the triangle inequality are not affected by this operation. d is well-defined because if $x \sim z$ then $|d(x,.) - d(z,.)| \leq d(x,z) = 0$. The fact that $-d^2$ is CPD on $\mathcal{X}/\sim$ follows from the fact that all possible representations of equivalence classes are points in $\mathcal{X}$ and $-d^2$ is CPD on $\mathcal{X}$. It is also well defined because if $x \sim z$ then

$$|d^2(x,.) - d^2(z,.)| \leq d(x,z)|(d(x,.) + d(z.,)| = 0.$$

The argumentation that k is also PD on $\mathcal{X}/\sim$ is the same as above. It is well defined because if $x \sim x'$ then $\|\Phi_x - \Phi_{x'}\| = 0$, so that actually $k(x,\cdot) = k(x',\cdot)$ (since for all $y \in \mathcal{X}$, $|k(x,y) - k(x',y)| \leq \|\Phi_x - \Phi_{x'}\| \, \|\Phi_y\|$). □

The equivalence relation defined in Theorem 5 can be seen as defining a kind of global invariance on $\mathcal{X}$. For example in the SVM setting when we have the kernel $k(x,y) = \langle x, y \rangle^2$, the equivalence relation identifies all points which are the same up to a reflection. This can be understood as one realization of an action of the discrete group $D = \{-e, +e\}$ on $\mathbb{R}^n$, so this kernel can be understood as a kernel on $\mathbb{R}^n/D$.

Assume now that there are no invariances in the data and two different points $x \neq y$ with different labels are such that $d(x,y) = 0$. Then they cannot be separated by any hyperplane. This means that using semi-metrics implicitly assumes invariances in the data, which may not hold.

5 Generalization Bounds using Rademacher Averages

In this section we calculate the Rademacher averages corresponding to the function classes of the two algorithms presented. The Rademacher average is a measure of capacity of a function class with respect to classification, and can be used to derive upper bounds on the error of misclassification (see e.g. Theorems 7 and 11 from [1]).

Let P be a probability distribution on $\mathcal{X} \times \{\pm 1\}$ and consider a training sample $T = \{(X_i, Y_i)_{i=1}^n\}$ drawn according to P^n. Let $\widehat{R}_n$ be the empirical Rademacher average of the function class $\mathcal{F}$, defined as

$$\widehat{R}_n(\mathcal{F}) = E_\sigma \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i)| \,,$$

where σ are Rademacher variables and E_σ denotes the expectation conditional to the sample (i.e. with respect to the σ_i only). The function classes we are interested in are those of continuous linear functionals on Hilbert or Banach spaces. More precisely, we consider the following two classes. For a given PD kernel k, let $\tilde{k}$ be defined as $\tilde{k}(x,\cdot) = k(x,\cdot) - k(x_0,\cdot)$[2] and $\mathcal{H}$ be the associated RKHS for $\tilde{k}$. We define $\mathcal{F}_1 = \{g : g \in \mathcal{H}, \|g\| \leq B\}$. Also, with the notations of the previous section, we define $\mathcal{F}_2 = \{e \in \bar{E}, \|e\| \leq B\}$.

Theorem 6. *With the above notation, we have*

$$\widehat{R}_n(\mathcal{F}_1) \leq \frac{B}{n}\sqrt{\sum_{i=1}^n d(x_i,x_0)^2}\,.$$

where $d(x_i,x_0) = \|k(x_i,\cdot) - k(x_0,\cdot)\|_{\mathcal{H}}$ is the distance induced by the kernel on $\mathcal{X}$. Also, there exists a universal constant C such that

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{CB}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\frac{\varepsilon}{2},\mathcal{X},d)}\,d\varepsilon.$$

Proof. We first compute the Rademacher average for $\mathcal{F}_2$:

$$\begin{aligned}\widehat{R}_n(\mathcal{F}_2) &= E_\sigma \sup_{e\in\bar{E},\|e\|\leq B}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \langle e, \Phi_{x_i}\rangle\right| = E_\sigma \sup_{e\in\bar{E},\|e\|\leq B}\left|\left\langle e, \frac{1}{n}\sum_{i=1}^n \sigma_i \Phi_{x_i}\right\rangle\right| \\ &= \frac{B}{n}E_\sigma\left\|\sum_{i=1}^n \sigma_i\Phi_{x_i}\right\|_\infty = \frac{B}{n}E_\sigma \sup_{x\in\mathcal{X}}\left|\sum_{i=1}^n \sigma_i \Phi_{x_i}(x)\right| \qquad (9)\end{aligned}$$

We will use Dudley's upper bound on the empirical Rademacher average [5] which gives that there exists an absolute constant C for which the following holds: for any integer n, any sample $\{x_i\}_{i=1}^n$ and every class $\mathcal{F}_2$,

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{C}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\varepsilon,\mathcal{F}_2,L_2(\mu_n))}d\varepsilon, \qquad (10)$$

where μ_n is the empirical measure supported on the sample and $N(\varepsilon,\mathcal{F}_2,L_2(\mu_n))$ are the covering numbers of the function class $\mathcal{F}_2$ with respect to $L_2(\mu_n)$.

In order to apply this result of Dudley, we notice that the elements of $\mathcal{X}$ can be considered as functions defined on $\mathcal{X}$. Indeed, for each $x \in \mathcal{X}$, one can define the function $f_y : x \mapsto \Phi_x(y)$. We denote by $\mathcal{G}$ the class of all such functions, i.e. $\mathcal{G} = \{f_y : y \in \mathcal{X}\}$. Then using (9), we get

$$\widehat{R}_n(\mathcal{F}_2) = B\,E_\sigma \sup_{x\in\mathcal{X}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\Phi_{x_i}(x)\right| = B\,\widehat{R}_n(\mathcal{G}). \qquad (11)$$

[2] where $k(x_0,\cdot)$ corresponds to the origin in $\mathcal{H}$ and is introduced to make the comparison with the space $\bar{E}$ easier

We now try to upper bound the empirical L_2-norm of $\mathcal{G}$:

$$\begin{aligned}\|f_{y_1} - f_{y_2}\|_{L_2(\mu_n)} &\leq \|f_{y_1} - f_{y_2}\|_{L_\infty(\mu_n)} = \max_{x_i \in T} |\Phi_{x_i}(y_1) - \Phi_{x_i}(y_2)| \\ &= \max_{x_i \in T} |d(x_i, y_1) - d(x_i, y_2) + d(x_0, y_2) - d(x_0, y_1)| \\ &\leq 2d(y_1, y_2). \end{aligned} \tag{12}$$

Combining (10) and (12) we get

$$\widehat{R}_n(\mathcal{G}) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\frac{\varepsilon}{2}, \mathcal{X}, d)} d\varepsilon$$

This gives the first result. Similarly, we have

$$\widehat{R}_n(\mathcal{F}_1) = \frac{B}{n} E_\sigma \left\| \sum_{i=1}^n \sigma_i (k(x_i, .) - k(x_0, .)) \right\|_{\mathcal{H}} \leq \frac{B}{n} \sqrt{\sum_{i=1}^n d(x_i, x_0)^2}\,,$$

where the second step follows from Jensen's inequality (applied to the concave function $\sqrt{\cdot}$). □

Notice that a trivial bound on $\widehat{R}_n(\mathcal{F}_2)$ can be found from (9) and

$$\left| \sum_{i=1}^n \sigma_i (d(x_i, x) - d(x_0, x)) \right| \leq \sum_{i=1}^n d(x_i, x_0),$$

which gives the upper bound

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{B}{n} \sum_{i=1}^n d(x_i, x_0)\,,$$

which is also an upper bound on $\widehat{R}_n(\mathcal{F}_1)$. However, this upper bound is loose since if all the data is at approximately the same distance from x_0 (e.g. on a sphere), then this quantity does not decrease with n.

6 Conclusion and Perspectives

In this article we have built a general framework for the generation of maximal margin algorithms for metric spaces. We first use an isometric embedding of the metric space into a Banach space followed by a maximal margin separation. It turned out that the SVM uses the same principle, but is restricted to the special class of metric spaces that allow an isometric embedding into a Hilbert space. In the following diagram the structure of both algorithms is shown:

$$\begin{array}{ccc} & & RKHS \xrightarrow{continuous} C(\mathcal{X}) \\ (\mathcal{X}, d) & \xrightarrow{isometric} & \\ & & (\bar{D}, \|.\|_\infty) \xrightarrow{isometric} (C_b(\mathcal{X}), \|.\|_\infty) \end{array}$$

The structural difference between the two algorithms is the space into which they embed. Since there exist several isometric embeddings of metric spaces into normed linear spaces, this raises two questions. First what is their difference in terms of mathematical structure, and second what are the consequences for a learning algorithm, especially its generalization ability ?
Further on in the SVM case we shifted the problem of choosing a kernel on $\mathcal{X}$ to the problem of choosing a metric on $\mathcal{X}$. Maybe one can construct a measure on the space of metrics for a given space $\mathcal{X}$, which can be calculated on the data, that captures the heuristic notion of "small inner class distance and big distance between the classes".

Acknowledgements

We would like to thank Ulrike von Luxburg, Bernhard Schölkopf and Arthur Gretton for helpful discussions and comments during the preparation of this article.

References

1. P. L. Bartlett, S. Mendelson, *Rademacher and Gaussian Complexities: Risk Bounds and Structural Results*, JLMR, **3**, 463-482, (2002).
2. K. P. Bennett,E. J. Bredensteiner, *Duality and Geometry in SVM classifiers*, Proceedings of the Seventeenth International Conference on Machine Learning, 57-64, (2000).
3. C. Berg, J.P.R. Cristensen, P. Ressel, *Harmonic Analysis on Semigroups*, Springer Verlag, New York, (1984).
4. F. Cucker, S. Smale, *On the Mathematical Foundations of Learning*, Bull. Amer. Math. Soc., **39**, 1-49, (2002).
5. R. M. Dudley, *Universal Donsker Classes and Metric Entropy*, Ann. Prob.,**15**,1306-1326, (1987).
6. T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.R. Müller, K. Obermayer and R. Williamson, *Classification on proximity data with LP-machines*, International Conference on Artificial Neural Networks, 304-309, (1999).
7. E. Pekalska, P. Paclik, R.P.W. Duin, *A Generalized Kernel Approach to Dissimilarity-based Classification*, Journal of Machine Learning Research, **2**, 175-211, (2001).
8. W. Rudin, *Functional Analysis*, McGraw Hill, (1991).
9. I. J. Schoenberg, *Metric Spaces and Positive Definite Functions*, TAMS, **44**, 522-536, (1938).
10. B. Schölkopf, *The Kernel Trick for Distances*, Neural Information Processing Systems (NIPS), **13**, (2000).
11. B. Schölkopf, A. J. Smola *Learning with Kernels*, MIT Press, MA, Cambridge, (2002).
12. D. Zhou, B. Xiao, H. Zhou, R. Dai, *Global Geometry of SVM Classifiers*, Technical Report 30-5-02, AI Lab, Institute of Automation, Chinese Academy of Sciences, (2002).