Maximal margin classification for metric spaces

Matthias Hein, Olivier Bousquet and Bernhard Schölkopf

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

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.* [7]). Interestingly enough, the embedding approach, when applied to a metric which can be embedded into a Hilbert space, yields the 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.

Key words: Classification, maximum margin, metric spaces, embedding, pattern recognition.
PACS:

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

Email address: matthias.hein, olivier.bousquet, bernhard.schoelkopf@tuebingen.mpg.de (Matthias Hein, Olivier Bousquet and Bernhard Schölkopf).

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 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.* [7]. 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 which the data is embedded has more structure. A natural choice is to use a Hilbert space (i.e. a Banach space where the norm is derived from an inner product). However, we recall a result of Schoenberg which states that only a certain class of metric spaces can be isometrically embedded into a Hilbert space. Hence, we gain structure at the price of loosing generality. Moreover, we give a characterization of metric spaces that can be embedded into a Hilbert space with some distortion of the large distances. If the metric has the appropriate properties, we thus also derive an embedding into a Hilbert space and the corresponding maximal margin algorithm.
It turns out that the obtained algorithm is equivalent to the well-known Support Vector Machine (e.g. [15]). We thus obtain a new point of view on this algorithm which is based on an isometric embedding of the input space as a metric space, where the metric is induced by a kernel. However, the main distinction between our point of view and the more classical one, is that we show that the solution only depends on the metric induced by the kernel and not on the kernel itself. And given this metric, the effect of the algorithm is to perform maximal margin separation after an *isometric* embedding into a Hilbert space.

Finally we investigate the properties of the class of functions that are associated with these embeddings. In particular we want to measure their capacity. For that we use a (by now) standard measure of the size in learning theory, the Rademacher averages. These can be directly related to the generalization error of the algorithm. Our computations show that in the case of the Banach space embedding of an arbitrary metric space, the size of the obtained class of hypotheses is the same as the size of $(\mathcal{X}, d)$ itself as a metric space, where the size is measured by the covering numbers. For the second embedding into a Hilbert space, we get results similar to the previously known ones for SVM, but we express them in terms of the induced (semi)-metric. Finally, in the case where $\mathcal{X}$ can be embedded isometrically both in a Banach and a Hilbert space, we compare the capacities of both obtained hypotheses classes and show that the SVM algorithm corresponds to a more "parsimonious" space of functions.

The paper is organized as follows. Section 2 introduces the general approach of embedding into a Banach space and performing maximum margin classification in this space. In particular, several possible embeddings with their effects on the metric are discussed. In Section 3 this approach is applied to an arbitrary metric space and we give the resulting general algorithm. Then, section 4 deals with the special case of metric spaces that can be isometrically embedded into a Hilbert space. These metrics are characterized and we derive, with our general approach, an algorithm which turns out to be equivalent to the SVM algorithm. Finally in section 5, we compute Rademacher averages corresponding to the previously mentioned algorithms and compare them.

2 The general approach

We are working in the following setting. We are given a set $\mathcal{X}$, together with a (semi-)metric defined on it, which makes it a (semi-)metric space $(\mathcal{X}, d)$. 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 $d(x, x) = 0$ for all $x \in \mathcal{X}$ (it is a metric if $d(x, y) = 0$ implies $x = y$).

Remark 1 *In the following we will consider only metric spaces. But all the results remain true for semi-metric spaces. The reason why we restrict ourselves to metric spaces is on the one hand simplicity but on the other hand the in general undesired implications of a semi-metric, see Appendix A for this issue.*

Our basic assumption is that this metric is consistent with the classification problem to be solved in the sense that when two points are close, they are likely to belong to the same class. Of course, there are many algorithms that can take into account such an assumption to build a classifier (e.g. nearest neighbors classifiers). Moreover if one has more structure than the pure metric space e.g. when $\mathcal{X}$ is a differentiable manifold, then this knowledge should be used in the classifier. In the sense that one should build functions which satisfy stronger smoothness requirements. One could argue that then the approach presented here is too general since at first sight we only use the metric structure of the input space. However as we will show later the functions generated used by the general maximal margin algorithm are always Lipschitz functions which can be regarded as the lowest level of smoothness. Moreover if the metric has stronger smoothness properties e.g. in the case of a Riemannian manifold then these smoothness properties are also transferred to the associated function space used by the maximal margin classifier. This will become obvious from the form of embedding we use. In that sense the maximal margin algorithm adapts to the smoothness of $\mathcal{X}$.

One of the cornerstones of the algorithm we use is the large margin heuristic. Thus we work with hyperplanes in a linear space. Since $\mathcal{X}$ need not be a linear space, we have to *transform* it into one, which can be done by *embedding* it into a linear space (with a norm defined on it). Since the metric information is the only information available to us to perform classification and we assume that the local structure is correlated to the class affiliations, we should not distort it too much in the embedding process. Or in other words the minimal requirement for our embedding is that it preserves neighborhoods, so it should at least be a homeomorphism of $(\mathcal{X}, d)$ onto a subset of a linear space.
The following diagram summarizes this procedure:

$$(\mathcal{X}, d) \xrightarrow{embedding} (\mathcal{B}, \|\cdot\|) \rightarrow \text{maximal margin classification}$$

2.1 First step: embedding into a normed space

Maximal margin hyperplane classification requires that we work in a linear normed space. We thus have to map $\mathcal{X}$ to a subset of a normed space B (chosen to be complete, hence a Banach space).

Formally, we define a *feature map* $\Phi : \mathcal{X} \rightarrow \mathcal{B}$, $x \rightarrow \Phi(x)$, and denote by $d_{\mathcal{B}}$ the induced metric on $\mathcal{X}$.

$$d_{\mathcal{B}}(x, y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{B}} \, .$$

We require that d and $d_{\mathcal{B}}$ are not too different since we want to preserve the metric information, which we assume to be relevant for classification. In other words we want that the map Φ seen as the identity map id between the metric spaces $(\mathcal{X}, d)$ and $(\mathcal{X}, d_{\mathcal{B}})$ to have one of the properties in the following list. We give the embeddings in the order of increasing requirements and each embedding is a special case of the previous one.

(1) Φ is an embedding if and only if Φ is a homeomorphism, that is

$$\begin{aligned} &\forall\, x, y \in \mathcal{X}, \forall\, \epsilon > 0, \exists\, \delta_1, \delta_2 \text{ such that:} \\ &d(x, y) < \delta_1 \Rightarrow d_{\mathcal{B}}(x, y) < \epsilon, \quad d_{\mathcal{B}}(x, y) < \delta_2 \Rightarrow d(x, y) < \epsilon. \end{aligned}$$

(2) Φ is a uniform embedding if and only if $\mathsf{id} : (\mathcal{X}, d) \rightarrow (\mathcal{X}, d_{\mathcal{B}})$ is a uniform homeomorphism, that is

$$\begin{aligned} &\forall\, \epsilon > 0, \exists\, \delta_1, \delta_2 \text{ such that } \forall\, x, y \in \mathcal{X} : \\ &d(x, y) < \delta_1 \Rightarrow d_{\mathcal{B}}(x, y) < \epsilon, \quad d_{\mathcal{B}}(x, y) < \delta_2 \Rightarrow d(x, y) < \epsilon. \end{aligned}$$

(3) Φ is a Bi-Lipschitz embedding, that is

$$\exists \lambda > 0, \ \forall\, x, y \in \mathcal{X}, \ \frac{1}{\lambda}\, d(x, y) \leq d_{\mathcal{B}}(x, y) \leq \lambda\, d(x, y) \, .$$

(4) Φ is an isometric embedding,

$$\forall\, x, y \in \mathcal{X},\ d_{\mathcal{B}}(x, y) = \|\Phi(x) - \Phi(y)\| = d(x, y)\,.$$

In this paper we will consider two cases.

- In the first case we assume that the metric $d(x, y)$ is meaningful and helpful for the classification task on all scales. That means we should preserve the metric in the embedding process, that is Φ should be an isometric embedding.
- In the second case we assume that the metric $d(x, y)$ is only locally meaningful. What do we mean by that? In the construction of a metric on a set $\mathcal{X}$ for a real-world problem one has some intuition about what it means for two elements $x, y \in \mathcal{X}$ to be 'close' and can encode this information in the metric $d(x, y)$. However larger distances are sometimes not very meaningful or even completely arbitrary. Consider for example the edit-distance for sequences. It is fairly clear, what it means to have an edit-distance of one or two, namely the word sequence is roughly the same. However for two completely different sequences the distance will be large without any meaning and will probably have a great influence on the construction of the classifier with the danger of fitting an irrelevant feature.
  Therefore in cases where we trust our metric only locally it makes no difference if we change the global structure as long as we preserve the local structure. Additionally this change of the global structure should fulfill two requirements. First it should be uniform over $\mathcal{X}$, since without further information we have no reason to change it differently in some regions. Second it should eliminate the influence of high distance values.
  In mathematical terms:
  Definition 1 *The local distortion of a map $\phi : (\mathcal{X}, d) \to (\mathcal{X}, d_{\mathcal{B}})$ is given by*

$$\mu(x) = D_+(x)/D_-(x)$$

  where the functions $D_+(x)$ and $D_-(x)$ are defined as

$$D_+(x) = \lim_{y \to x} \sup \frac{d_{\mathcal{B}}(x, y)}{d(x, y)}, \quad D_-(x) = \lim_{y \to x} \inf \frac{d_{\mathcal{B}}(x, y)}{d(x, y)}.$$

  Definition 2 *A uniform local isometry is a uniform homeomorphism $\phi : (\mathcal{X}, d) \to (\mathcal{X}, d_{\mathcal{B}})$ with local distortion $\mu(x) \equiv 1$.*
  A uniform local isometry preserves the local structure up to a global rescaling, which does not matter for the maximal margin classification. Finally our embedding should be a uniform local isometry such that the transformed metric is bounded, i.e. $\sup_{x,y \in \mathcal{X}} d_{\mathcal{B}}(x, y)$ exists.

It is interesting to note here that for all embeddings $\Phi : (\mathcal{X}, d) \to (\mathcal{B}, \|\cdot\|)$ one can adopt two points of view:

- Direct embedding: $\Phi : (\mathcal{X}, d) \to (\mathcal{B}, \|\cdot\|_{\mathcal{B}})$
- Indirect embedding: identity $\mathsf{id} : (\mathcal{X}, d) \to (\mathcal{X}, d_{\mathcal{B}})$ and isometric embedding $\phi : (\mathcal{X}, d_{\mathcal{B}}) \to (\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ with $\Phi = \phi \circ \mathsf{id}$

The above two points of view are completely equivalent (i.e. any embedding Φ can be written as $\phi \circ \mathsf{id}$ where id is the identity and ϕ an isometric embedding and conversely) but the second point of view emphasizes the importance of isometric embeddings. Namely any embedding can be decomposed into a transformation of the initial metric followed by an isometric embedding. This equivalence allows us to treat isometric and uniform locally isometric embeddings in the same framework.

The first question is how to construct such a uniform local isometry. One general way to do this are the so called metric transforms introduced by Blumenthal. (We use here and in the following $\mathbb{R}_+ = \{x \in \mathbb{R} \,|\, x \geq 0\}$.)

Definition 3 *Let $(\mathcal{X}, d)$ be a metric space and let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a function with $F(0) = 0$. Then $(\mathcal{X}, F(d))$ is called a metric transform of $(\mathcal{X}, d)$.*

The following lemma gives sufficient conditions for a metric transform $F(d)$ to be a metric.

Lemma 1 *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotone increasing concave function, such that $F(0) = 0$ and $F(x) > 0$ for all $x > 0$. If d is a metric on $\mathcal{X}$, then $F(d)$ is also a metric on $\mathcal{X}$.*

The proof of this lemma can be found e.g. in [5]. We denote the functions which fulfill the assumptions of the above lemma as true metric transforms. Note that the map $\mathsf{id} : (\mathcal{X}, d) \to (\mathcal{X}, F(d))$ is a uniform homeomorphism for every true metric transform. The next lemma characterizes all true metric transforms which are in addition uniform local isometries.

Lemma 2 *Let F be a true metric transform. If $\lim_{t\to 0} \frac{F(t)}{t}$ exists and is positive, then identity $\mathsf{id} : (\mathcal{X}, d) \to (\mathcal{X}, F(d))$ is a uniform local isometry. Moreover the resulting metric space $(\mathcal{X}, F(d))$ is bounded if F is bounded.*

Proof With the assumptions the functions $D_+(x)$ and $D_-(x)$ defined in Definition 1 exist and $\forall x \in \mathcal{X}$, $D_+(x) = D_-(x) > 0$, so that $\mu \equiv 1$. □

In order to illustrate this lemma, we give two examples of metric transforms F which result in uniform local isometries, where $(\mathcal{X}, F(d))$ is bounded.

$$F(t) = \frac{t}{1+t}, \qquad F(t) = 1 - \exp(-\lambda t),\ \forall\, \lambda > 0 \tag{1}$$

Furthermore an important question is whether there exists for any given metric space $(\mathcal{X}, d)$ a Banach space $\mathcal{B}$ and a map Φ which embeds $(\mathcal{X}, d)$ isometrically

into $\mathcal{B}$. In the following we will answer this positively, namely any metric space $(\mathcal{X}, d)$ can be embedded isometrically via the *Kuratowski* embedding into $(C_b(\mathcal{X}), \|\cdot\|_\infty)$, where $C_b(\mathcal{X})$ denotes the continuous, bounded functions on $\mathcal{X}$. However in the later analysis of the maximal margin algorithm it turns out that an embedding into a Hilbert space provides a simpler structure of the space of solutions. Therefore we will consider after the general case of an isometric embedding into a Banach space the special case of an isometric embedding into a Hilbert space.
Moreover all isometric embeddings we consider have the following minimal property:

Definition 4 (Total isometric embedding) *Given a metric space $(\mathcal{X}, d)$ and an isometric embedding $\Phi : \mathcal{X} \to \mathcal{B}$ where $\mathcal{B}$ is a Banach space, we say that Φ is a* total isometric embedding *if $\Phi(\mathcal{X})$ is total, that is $\mathcal{B}$ is the norm-closure of* $\operatorname{span}\{\Phi(x)|x \in \mathcal{X}\}$.

This definition is in a sense trivial, since if we have an isometric embedding Φ into a Banach space $\mathcal{C}$, then the norm closure of $\operatorname{span}\Phi(\mathcal{X})$ is again a Banach space $\mathcal{B}$ with the same norm. But this 'minimal' isometric embedding allows then to associate to the dual space $\mathcal{B}'$ (the space of continuous linear functionals on $\mathcal{B}$ endowed with the norm $\|w'\| = \sup_{b\in\mathcal{B}, \|b\|\leq 1} |w'(b)|$ [1] an isometrically isomorphic Banach space of functions on $\mathcal{X}$ as we will see now.

Proposition 1 *Let $\Phi : \mathcal{X} \to B$ be a total isometric embedding. Then there exists a Banach space $\mathcal{F}_{\mathcal{B}'}$ of real-valued Lipschitz functions on $\mathcal{X}$ and a map $\Gamma : \mathcal{B}' \to \mathcal{F}_{\mathcal{B}'}$ such that Γ is an isometric isomorphism. The map Γ is given by*

$$\Gamma(w')(\cdot) = \langle w', \Phi(\cdot)\rangle_{\mathcal{B}',\mathcal{B}}$$

and we define $\|\Gamma(w')\|_{\mathcal{F}_{\mathcal{B}'}} = \|w'\|_{\mathcal{B}'}$. The Lipschitz constant of $\Gamma(w')$ is upper bounded by $\|w'\|_{\mathcal{B}'}$.

We need for the proof of the proposition and in the rest of the article the following notions and a theorem relating them.

Definition 5 *Let M, N be subspaces of $\mathcal{B}$ resp. $\mathcal{B}'$. Then the annihilators $M^\perp$ and ${}^\perp N$ are defined as*

$$M^\perp = \{w' \in \mathcal{B}' : \langle w', m\rangle = 0, \forall\, m \in M\},$$
$${}^\perp N = \{b \in \mathcal{B} : \langle n, b\rangle = 0, \forall\, n \in N\}.$$

Theorem 1 *[11]*

[1] Given an element b of a Banach space B and an element w' of its dual B', we write $w'(b) = \langle w', b\rangle_{B',B}$. This should not be confused with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ in a Hilbert space.

- ${}^{\perp}(M^{\perp})$ *is the norm closure of* M *in* $\mathcal{B}$.
- $({}^{\perp}N)^{\perp}$ *is the weak*-closure of* N *in* $\mathcal{B}'$

Now we can prove Proposition 1.

Proof The only thing we have to prove is that Γ is injective. Let $f, g \in \mathcal{F}_{\mathcal{B}'}$, then

$$f \equiv g \Leftrightarrow \langle w_f - w_g, \Phi(x) \rangle_{\mathcal{B}',\mathcal{B}} = 0, \ \forall\, x \in \mathcal{X}.$$

Since Φ is a total isometric embedding, $\mathcal{B} = \overline{\text{span}\{\Phi(\mathcal{X})\}} = {}^{\perp}(\text{span}\,\{\Phi(\mathcal{X})\}^{\perp})$. In particular, we have $\text{span}\{\Phi(\mathcal{X})\}^{\perp} = \{0\}$. Therefore $w_f - w_g = 0$, that is, Γ is injective. The Lipschitz constant of $\Gamma(w')$ can be computed as follows. For all $x, y \in \mathcal{X}$,

$$\begin{aligned} |\Gamma(w')(x) - \Gamma(w')(y)| &= |\, \langle w', \Phi(x) - \Phi(y) \rangle_{\mathcal{B}',B}\, | \leq \|w'\|_{\mathcal{B}'}\, \|\Phi(x) - \Phi(y)\|_{\mathcal{B}} \\ &= \|w'\|_{\mathcal{B}'}\, d(x, y). \end{aligned}$$

□

The fact that one always obtains Lipschitz functions has been pointed out in [16] where it is shown that any isometric embedding can be obtained via an embedding into the predual of Lipschitz functions.

2.2 Second step: maximal margin classification

2.2.1 Maximal margin and its dual problem

What does maximal margin classification mean? The classifier is a hyperplane in $\mathcal{B}$, which can be identified with an element in the dual of $\mathcal{B}'$ plus an offset, such that the distance, the margin, to the two classes is maximized. This problem is equivalent to the problem of determining the distance between the convex hulls of the two classes of our training data. This duality was proven in the generality of an arbitrary Banach space by Zhou et al. [17]. We define the convex hull of a finite set $T \subset \mathcal{B}$ as

$$co(T) = \left\{ \sum_{i \in I} \alpha_i x_i \,|\, \sum_{i \in I} \alpha_i = 1,\, x_i \in T,\, \alpha_i \geq 0,\, |I| < \infty \right\}.$$

Theorem 2 *[17] Let* T_1 *and* T_2 *be two finite sets in a Banach space* $\mathcal{B}$. *Then if* $co(T_1) \cap co(T_2) = \emptyset$

$$\begin{aligned} d(co(T_1), co(T_2)) &= \inf_{y \in co(T_1), z \in co(T_2)} \|y - z\| \\ &= \sup_{w' \in \mathcal{B}'} \frac{\inf_{y \in T_1, z \in T_2} \langle w', y - z \rangle_{\mathcal{B}',B}}{\|w'\|}. \end{aligned} \tag{2}$$

The condition $co(T_1) \cap co(T_2) = \emptyset$ is equivalent to the condition of separability.

Corollary 1 *The maximal margin problem is translation invariant in the Banach space $\mathcal{B}$.*

Proof This is a trival statement, since we are only interested in distances. □

Later we will use the above dual formulation in order to derive properties of the solution $w' \in \mathcal{B}'$.

2.2.2 Maximal margin formulations

In this section we derive from the dual problem the usual maximal margin formulation. We consider an input sample $x_1, \ldots, x_n \in \mathcal{X}$ with labels $y_1, \ldots, y_n \in \{-1, 1\}$. These samples can be embedded via Φ into a Banach space $\mathcal{B}$. We denote by Φ_x the embedded point $\Phi(x)$ and by T_1 the set $\{\Phi_{x_i} : y_i = +1\}$ of positive examples and by $T_2 = \{\Phi_{x_i} : y_i = -1\}$ the set of negative examples.

First we rewrite the second line of (2) by using the definition of the infimum:

$$\begin{array}{ll} \sup\limits_{x' \in \mathcal{B}', c, d \in \mathbb{R}} & \dfrac{c-d}{\|x'\|} \\ \text{subject to:} & \langle x', y \rangle_{\mathcal{B}',B} \geq c, \quad \forall\, y \in T_1, \quad \langle x', z \rangle_{\mathcal{B}',B} \leq d, \quad \forall\, z \in T_2. \end{array}$$

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 taking the inverse we arrive at the standard hard margin formulation:

$$\begin{array}{ll} \min\limits_{w' \in \mathcal{B}', b} & \|w'\| \\ \text{subject to:} & y_i(\langle w', \Phi_{x_i} \rangle_{\mathcal{B}',B} + b) \geq 1, \quad \forall i = 1, \ldots, n. \end{array} \tag{3}$$

Another equivalent formulation where we use the space of functions $\mathcal{F}_{\mathcal{B}'}$ which we defined in Proposition 1 takes more the point of view of regularization.

$$\min_{f_{w'} \in \mathcal{F}_{\mathcal{B}'}, b} \|f_{w'}\|_{\mathcal{F}_{\mathcal{B}'}} + \sum_{i=1}^{n} \ell\left(y_i(f_{w'}(x_i) + b)\right) \tag{4}$$

where the loss function ℓ is given by $\ell(x) = 0$, $\forall\, x \geq 1$, $\ell(x) = \infty$, $\forall\, x < 1$.

In principle we have two points of view on the hard margin problem. One is based on the geometric interpretation (2), (3) of finding a separating hyperplane with maximal distance to the two classes. The other is based on (4) and regards the problem as the search for a function which classifies correctly and

has minimal norm, where we assume that the norm is some measure of smoothness. In this paper we will switch between these two viewpoints depending on which is better suited to illustrate a certain property.

2.2.3 Form of the solution

Let us now come back to the initial formulation (2). Our goal is to obtain a characterization of the solutions $w' \in \mathcal{B}'$. We consider the following subspace $A = \text{span}\{\Phi_{x_1} - \Phi_{x_2} \,|\, x_1 \in T_1, x_2 \in T_2\} \subset \text{span}\{\Phi_{x_i} \,|\, x_i \in T\}$ which can be equivalently written as

$$A = \left\{ \sum_{i=1}^{n} \alpha_i \Phi_{x_i} : \sum_{i=1}^{n} \alpha_i = 0 \right\}.$$

The following lemma characterizes the space of solution $w' \in \mathcal{B}'$.

Lemma 3 *The quotient space $\mathcal{B}'/A^\perp$, endowed with the quotient norm, is a Banach space. It is isometrically isomorphic to the dual A' of A and has dimension $n-1$. Moreover the problem of maximal margin separation in $\mathcal{B}'$, (3), is equivalent to the following problem in $\mathcal{B}'/A^\perp$:*

$$\begin{aligned} &\min_{w' \in \mathcal{B}'/A^\perp, b} \|w'\|_{\mathcal{B}'/A^\perp} \\ &\textit{subject to:} \quad y_i \left(\langle w', \Phi_{x_i} \rangle_{\mathcal{B}',\mathcal{B}} + b \right) \geq 1, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{5}$$

Proof A is finite dimensional hence closed in $\mathcal{B}$. It is thus a Banach space with the induced norm. It is well known (see e.g. [11]) that then $\mathcal{B}'/A^\perp$ with the quotient norm $\|b'\|_{\mathcal{B}'/A^\perp} = \inf\{\|b' - a'\| : a' \in A^\perp\}$ is a Banach space isometric isomorphic to A', the dual of A. Since A is a normed space of finite dimension $n-1$, its dual has the same dimension.
Since addition of elements of $A^\perp$ does not change the numerator of (2), but will change the norm in the denominator, the problem can be equivalently formulated in the quotient space $\mathcal{B}'/A^\perp$.
In the constraint of (5), w' is an arbitrary representative of its equivalence class $w' + A^\perp$. This is well defined, since if u' is another representative of the equivalence class we have $m' = u' - w' \in A^\perp$ and $\forall m' \in A^\perp, \phi_{x_i}, \phi_{x_j} \in T$,

$$\left\langle m', \phi_{x_i} - \phi_{x_j} \right\rangle_{\mathcal{B}',\mathcal{B}} = 0.$$

That is, m' is constant on the data. Therefore if w' satisfies the constraint with constant b, u' will satisfy the constraint with the constant $c = b - \langle m, \phi_{x_i} \rangle_{\mathcal{B}',\mathcal{B}}$. □

Remarkably, this lemma tells us that the solution of the maximum margin

problem is effectively in a finite dimensional subspace of $\mathcal{B}'$ which is determined by the data. However, it gives no explicit description how this subspace depends on the data, which makes it hard to be effectively used in general.

Moreover, in order to solve the initial problem using the above lemma, one has to first solve the finite dimensional problem in $\mathcal{B}'/A^\perp$ and then to solve the minimum norm interpolation problem in $\mathcal{B}'$. Indeed, if a is a solution in $\mathcal{B}'/A^\perp$, one has to find an element b' in the equivalence class a. For this one has to solve

$$\inf_{b' \in \mathcal{B}' : b'|_A = a|_A} \|b'\|_{\mathcal{B}'} \ ,$$

which corresponds to minimizing the norm provided the values on a finite dimensional subspace are known.

We give an interpretation of this lemma from the point of view of functions which we developed in the previous section. The closed subspace $A^\perp$ of $\mathcal{B}'$ defines a closed subspace of functions $\mathcal{F}_{A^\perp}$ of $\mathcal{F}_{\mathcal{B}'}$ on $(\mathcal{X}, d)$ which are constant on all data points, namely $\forall\, w' \in A^\perp$, $x_1 \in T_1, x_2 \in T_2$

$$f_{w'}(x_1) - f_{w'}(x_2) = \langle w', \Phi(x_1) - \Phi(x_2) \rangle_{\mathcal{B}', B} = 0$$

The proposition then states that the solution is only defined up to a constant function on the data or in other words we are looking for a solution f in $\mathcal{F}_{\mathcal{B}'}/\mathcal{F}_{A^\perp}$ with the usual quotient norm $\|f\|_{\mathcal{F}_{\mathcal{B}'}/\mathcal{F}_{A^\perp}} = \inf_{g \in \mathcal{F}_{A^\perp}} \|f - g\|$. In particular, if there are constant functions (constant functions are constant on the data) in our function class $\mathcal{F}_{\mathcal{B}'}$, they will not be penalized in the norm. This reflects the fact that constant functions are useless for classification and should therefore not be considered in the norm of our solution space. Since we use the threshold 0 for classification we have to compensate for the constant functions on the data with the bias term b in the final solution.

$$f_{w'}(x) = \operatorname{sgn}(\langle w', \Phi(x) \rangle_{\mathcal{B}', B} + b).$$

Later we will consider also isometric embeddings into a Hilbert space $\mathcal{H}$. There we have $(A^\perp)^\perp = A$ and we can actually decompose $\mathcal{H}$ into $\mathcal{H} = A^\perp \oplus A$. Then the solution of the maximal margin problem is an element of A, which is itself a Hilbert space and consists of all functions $f \in \mathcal{H}$, orthogonal to the functions which are constant on the data. This is a stronger statement than the usual representer theorem, which says that the solution lies in the space spanned by the data.

3 Metric based maximal margin classifier in a Banach space

In this section we treat the general case, where we embed isometrically a given metric space $(\mathcal{X}, d)$ into a Banach space $\mathcal{B}$ followed by a maximal margin classification in $\mathcal{B}$. In general there exist for each metric space several Banach spaces, into which it can be embedded isometrically. In this section we use the very simple Kuratowski embedding. After the definition of the Kuratowski embedding Φ and the corresponding Banach space $\mathcal{B}$ we finally formulate the algorithm of maximal margin classification in $\mathcal{B}$. Unfortunately the full problem cannot be solved exactly. We provide a reasonable approximation to the full problem, which is exact if one considers the training set and a possible test point as a finite metric space.
The following diagram illustrates the employed procedure

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

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

3.1 Isometric embedding into a Banach space

Let $(\mathcal{X}, d)$ be a 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. If $\mathcal{X}$ is compact the topological dual of $C_b(\mathcal{X})$ is the space of finite signed Borel 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 map

$$\Phi : \mathcal{X} \rightarrow \mathbb{R}^{\mathcal{X}}, \quad x \mapsto \Phi_x := d(x, \cdot) - d(x_0, \cdot)$$

Let $D = \overline{\text{span}\{\Phi_x : x \in \mathcal{X}\}}$, where the closure is taken in $(C_b(\mathcal{X}), \|\cdot\|_\infty)$.
We will show that Φ defines an isometric embedding of the metric space $\mathcal{X}$ into D.

Lemma 4 *Φ is a total isometric embedding from $(\mathcal{X}, d)$ into the Banach space $(D, \|\cdot\|_\infty) \subset (C_b(\mathcal{X}), \|\cdot\|_\infty)$.*

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 $\|\Phi_x - \Phi_y\|_\infty = \|d(x, \cdot) - d(y, \cdot)\|_\infty = d(x, y)$ 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 closed subspace of $C_b(\mathcal{X})$. Therefore $(D, \|\cdot\|)$ is also Banach space and Φ a total isometric embedding, since by definition $D = \overline{\text{span}\, \Phi(\mathcal{X})}$. □

Note that, as an isometry, Φ is continuous, and x_0 is mapped to the origin of D. The choice of this origin Φ_{x_0} has no influence on the classifier since the maximal margin problem is translation invariant.

3.2 The algorithm

The maximal margin formulation (3) can be directly stated as:

$$\begin{aligned} &\min_{w' \in D', b \in \mathbb{R}} \|w'\| \\ &\text{subject to:} \quad y_j \left(\left\langle w', \Phi_{x_j} \right\rangle_{D',D} + b \right) \geq 1, \quad \forall\, j = 1, \ldots, n. \end{aligned} \tag{6}$$

Note that since we have no explicit description of the dual space D' we cannot solve this directly. If $\mathcal{X}$ is compact it is well-known that the dual of $C_b(\mathcal{X})$ is isometrically isomorphic to the Banach space of finite signed Borel measures $\mathcal{M}(\mathcal{X})$ on $\mathcal{X}$ with the measure norm. Thus we can state the problem explicitly. Note that even though we work in a bigger space than D', we will get the same solution lying in A' isometrically isomorphic to $\mathcal{M}(X)/A^{\perp}$ [2] since we are minimizing the norm:

$$\begin{aligned} &\min_{w' \in \mathcal{M}(\mathcal{X}), b \in \mathbb{R}} \|\mu\|_{\mathcal{M}(\mathcal{X})} \\ &\text{subject to:} \quad y_j \left(\int_{\mathcal{X}} (d(x_j, x) - d(x, x_0)) d\mu(x) + b \right) \geq 1, \quad \forall\, j = 1, \ldots, n. \end{aligned}$$

This problem also cannot be solved directly, since we have no parametrization of $\mathcal{M}(\mathcal{X})$.
Let us now consider again the general problem (6). Since we have neither a description of the dual $A' \simeq D'/A^{\perp}$ nor of D', we develop a reasonable approximation in the bigger space $C_b(X)'$. We introduce the space E defined as the span of evaluation functionals:

$$E := \operatorname{span}\{\delta_x : x \in \mathcal{X}\}\,.$$

First we have the following lemma:

Lemma 5 *The space E defined above is weak*-dense in the dual of $C_b(\mathcal{X})$ and the norm is given by $\|\sum_{i=1}^n \alpha_i \delta_{x_i}\|_{C_b(\mathcal{X})'} = \sum_{i=1}^n |\alpha_i|$.*

Proof The evaluation functionals are in the dual of $C_b(\mathcal{X})$ since

$$|\delta_x(f) - \delta_x(g)| = |f(x) - g(x)| \leq \|f - g\|_{\infty}$$

[2] Let $A^{\perp}_{\mathcal{M}(X)}, A^{\perp}_{D'}$ denote the annihilator of A in $\mathcal{M}(X)$ resp. D'. Then we have $A' \simeq \mathcal{M}(X)/A^{\perp}_{\mathcal{M}(X)} \simeq D'/A^{\perp}_{D'}$.

Consider now the span of evaluation functionals $\text{span}\{\delta_x : x \in \mathcal{X}\}$. The norm induced by $C_b(\mathcal{X})$ is given as

$$\left\| \sum_{i=1}^{n} \alpha_i \delta_{x_i} \right\|_{C_b(\mathcal{X})'} = \sup_{f \in C_b(\mathcal{X})'} \frac{|\langle \sum_{i=1}^{n} \alpha_i \delta_{x_i}, f \rangle|}{\|f\|_\infty} = \sup_{f \in C_b(\mathcal{X})'} \frac{|\sum_{i=1}^{n} \alpha_i f(x_i)|}{\|f\|_\infty} = \sum_{i=1}^{n} |\alpha_i|$$

Further on we have that ${}^{\perp}\{\delta_x : x \in \mathcal{X}\} = 0$ since $\langle \delta_x, f \rangle = 0, \forall\, x \in \mathcal{X} \Leftrightarrow f \equiv 0$. That implies

$$({}^{\perp}\{\delta_x : x \in \mathcal{X}\})^{\perp} = C_b(\mathcal{X})'$$

Therefore by Theorem 1 the weak*-closure of $\text{span}\{\delta_x : x \in \mathcal{X}\}$ is $C_b(\mathcal{X})'$. $\square$

Let us explain shortly what this result means. The weak*-topology is the topology of pointwise convergence on $C_b(\mathcal{X})$. Therefore the weak*-denseness of E in $C_b(\mathcal{X})'$ can be equivalently formulated as follows: $\forall\, \mu \in C_b(\mathcal{X})'$, $\exists\, \{e_\alpha\}_{\alpha \in I} \in E$ such that $e_\alpha \to \mu$ in the weak*-topology, that is

$$\forall\, f \in C_b(\mathcal{X}), \quad \langle e_\alpha, f \rangle_{C_b(\mathcal{X})', C_b(\mathcal{X})} \longrightarrow \langle \mu, f \rangle_{C_b(\mathcal{X})', C_b(\mathcal{X})}\,.$$

In other words one can approximate in the above sense any element of $C_b(\mathcal{X})'$ arbitrarily well with elements from E. On the other hand weak*-dense does not imply norm-dense.

Our first step is that we formulate the problem in $C_b(\mathcal{X})'$ which seems at first to be an approximation. But according to the same argument as before we have $A' \simeq C_b(X)'/A^{\perp} \simeq D'/A^{\perp}$. Since we are minimizing the norm under the given constraints this implies that the solution will lie in $C_b(\mathcal{X})'/A^{\perp}$ which is isometrically isomorphic to A'. Then as an first approximation we restrict $C_b(\mathcal{X})'$ to E. Since the span of evaluation functionals is not norm dense in $C_b(\mathcal{X})'$, this implies that even in the limit of an infinite number of evaluation functionals we might not get the optimal solution.
This approximation can be formulated as the following optimization problem:

$$\begin{aligned} \inf_{e \in E, b} \|e\| = \inf_{m \in \mathbb{N},\, z_1, \ldots, z_m \in \mathcal{X}^m,\, b} \quad & \sum_{i=1}^{m} |\beta_i| \\ \text{s.t. } y_j \left(\sum_{i=1}^{m} \beta_i \left\langle \delta_{z_i}, \Phi_{x_j} \right\rangle + b \right) = y_j & \left(\sum_{i=1}^{m} \beta_i (d(x_j, z_i) - d(x_0, z_i)) + b \right) \geq 1 \\ \forall j = 1, \ldots, n. & \end{aligned}$$

Unfortunately it is not possible to prove that the solution can be expressed in terms of the data points only (which would be a form of a representer theorem for this algorithm). We could actually construct explicit counterexamples. Note however that in [16] a representer theorem was derived for a similar but different setting. Namely they showed that if one considers all Lipschitz functions together with the Lipschitz constant as a norm, the solution lies in the *vector lattice* spanned by the data. However it is also shown there that this setting is not equivalent to the setting presented here. Moreover as we will

show later the capacity of all Lipschitz functions measured by Rademacher averages is higher than of our approach.
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 $Z \in \mathcal{X}, |Z| = m$, which includes the training set $T \subset Z$. We are free to choose the point x_0 in the embedding, so we choose it as $x_0 = z_1, z_1 \in Z$. Since the problem stated in Theorem 2 is translation invariant, this choice has no influence on the solution. This leads to the following optimization problem:

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

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 the second term in the constraint, $\sum_{i=1}^{m} \beta_i d(z_1, z_i)$, does not depend on j, we can integrate it in a new constant c and solve the equivalent problem:

$$\begin{aligned} &\min_{\beta_i, c} \sum_{i=1}^{m} |\beta_i| \\ &\text{subject to:} \quad y_j \left(\sum_{i=1}^{m} \beta_i \, d(x_j, z_i) + c \right) \geq 1, \quad \forall x_j \in T. \end{aligned} \tag{7}$$

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. [7]. 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 [2,17]. This algorithm was compared to other distance based classifiers by Pekalska et al. in [10] and showed good performance.

The approximation with a finite subset Z, $|Z| = m$, such that $T \subset Z$ can also be seen from another point of view. Namely consider the finite metric space (Z, d). Since the isometric embedding Φ is possible for any metric space, we can use it also in this special case and the Banach space of continuous, bounded functions $(C_b(Z), \|\cdot\|_\infty)$ is actually equal to $l_\infty^m = (\mathbb{R}^m, \|\cdot\|_\infty)$. We note that in the case of finite dimension m the dual of l_∞^m is given by l_1^m. Formulating the maximal margin problem in the Banach space l_∞^m leads then exactly to the

optimization problem (7). Therefore the approximation to the maximal margin problem for $(\mathcal{X}, d)$ using a finite subset of evaluation functionals indexed by Z is equivalent to the maximal margin problem for the finite metric space (Z, d) without any approximation. Moreover one can embed $m+1$ points isometrically into l_∞^m with the embedding Φ (z_1 is mapped to the origin of l_∞^m). Thus the resulting classifier is not only defined on Z but by embedding Z plus a possible test point $x \in \mathcal{X}$ isometrically into l_∞^m we can classify all points $x \in \mathcal{X}$ respecting all the distance relationships of x to Z.

4 Metric based maximal margin classifier in a Hilbert space

In the previous section we constructed a maximal margin classifier in the Banach space $D \subset (C_b(\mathcal{X}), \|\cdot\|_\infty)$ which works for any metric space $(\mathcal{X}, d)$, since any metric space can be embedded isometrically into $(C_b(\mathcal{X}), \|\cdot\|_\infty)$. The problem of the resulting maximal margin classifier is that the space of solutions $D'/A^\perp$ is not easily accessible. However in a Hilbert space the dual space $\mathcal{H}'$ is isometrically isomorphic to $\mathcal{H}$. Therefore we have $\mathcal{H}/A^\perp = (A^\perp)^\perp = A$, that is given n data points we have an explicit description of the at most $(n-1)$-dimensional space of solutions.
Regarding these properties of the space of solutions in $\mathcal{H}$ it seems desirable to rather embed isometrically into a Hilbert space than into a Banach space. It turns out that isometric embeddings into Hilbert spaces are only possible for a subclass of metric spaces. Following the general framework we first treat isometric and uniform locally isometric embeddings. Then the resulting maximal margin classifier is determined. Finally we show the equivalence to the SVM and provide an alternative point of view on kernels regarding SVM.

4.1 Isometric embedding into a Hilbert space

We have seen in the previous part that all metric spaces can be embedded isometrically into a Banach space. Is this true also for isometric embeddings into Hilbert spaces? The answer was given by Schoenberg in 1938 in terms of the following class of functions, by now well-known as positive definite resp. conditionally positive definite kernels.

Definition 6 *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*

$$\sum_{i,j}^{n} c_i c_j k(x_i, x_j) \geq 0, \tag{8}$$

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* $\sum_i^n c_i = 0$*).*

The metric spaces which can be isometrically embedded into a Hilbert space can be characterized as follows:

Theorem 3 (Schoenberg [12]) *A metric space* $(\mathcal{X}, d)$ *can be embedded isometrically into a Hilbert space if and only if* $-d^2(x, y)$ *is conditionally positive definite.*

Based on this characterization, one can introduce the following definition.

Definition 7 *A metric* d *defined on a space* $\mathcal{X}$ *is called a* Hilbertian metric *if* $(\mathcal{X}, d)$ *can be isometrically embedded into a Hilbert space, or equivalently if* $-d^2$ *is conditionally positive definite.*

We notice that isometric embeddings into a Hilbert space are only possible for a restricted subclass of metric spaces. So we achieve the advantage of having a small and easily accessible space of solutions by losing the ability to handle the whole class of metric spaces in this framework.
Let us now construct explicitly the corresponding isometric embedding.

Proposition 2 *Let* $d(x, y)$ *be a Hilbertian metric. Then for every point* $x_0 \in \mathcal{X}$ *there exists a reproducing kernel Hilbert space* $\mathcal{H}_k$ *and a map* $\Psi : \mathcal{X} \to \mathcal{H}_k$ *given by*

$$x \to \Psi_x(\cdot) = \frac{1}{2}\left(-d^2(x, \cdot) + d^2(x, x_0) + d^2(\cdot, x_0)\right)$$

such that

- $\{\Psi_x | x \in X\}$ *is total in* $\mathcal{H}_k$
- $\|\Psi_x - \Psi_y\|_{\mathcal{H}_k} = d(x, y)$.
- $\Psi_{x_0} = 0$

We need the following two lemmata to prove this proposition.

Lemma 6 *[3] Let* $\mathcal{X}$ *be a nonempty set,* $x_0 \in \mathcal{X}$*, and let* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a symmetric function. Let* $\tilde{k}(x, y)$ *be given by*

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

Then $\tilde{k}$ *is positive definite if and only if* k *is conditionally positive definite.*

Lemma 7 *[15] Given a positive definite kernel* $k(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *there exists a unique reproducing kernel Hilbert space (RKHS) of functions on* $\mathcal{X}$*, where* $\mathcal{H} = \overline{\text{span}\{k(x, \cdot) | x \in \mathcal{X}\}}$.

Proof [Proposition 2] Define the symmetric kernel function $k(x,y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x,y) = \frac{1}{2}\left(-d^2(x,y) + d^2(x,x_0) + d^2(y,x_0)\right).$$

Using Lemma 6, $k(x,y)$ is a positive definite kernel. Moreover by Lemma 7 there exists a unique reproducing kernel Hilbert space $\mathcal{H}_k$ associated to $k(x,y)$ such that $k(x,y) = \langle \Psi_x, \Psi_y \rangle_{\mathcal{H}_k}$ and $\{k(x,\cdot)|x \in \mathcal{X}\} = \{\Psi_x | x \in \mathcal{X}\}$ is total in $\mathcal{H}_k$. Moreover we have

$$\|\Psi_x - \Psi_y\|^2 = k(x,x) + k(y,y) - 2k(x,y) = d^2(x,y)$$

and $\Psi_{x_0}(\cdot) = \frac{1}{2}(-d^2(x,\cdot) + d^2(x,\cdot)) = 0$. □

4.2 Uniform locally isometric embedding into a Hilbert space

In the previous section we constructed an isometric embedding into a Hilbert space. If one trusts the metric $d(x,y)$ only locally we argued in section 2.1 that one should use a uniform locally isometric embedding.
The following proposition gives necessary and sufficient conditions for a uniform embedding of a metric space into a Hilbert space:

Proposition 3 *[1] A metric space $(\mathcal{X}, d)$ can be uniformly embedded into a Hilbert space if and only if there exists a positive definite kernel $k(x,y)$ on $\mathcal{X}$ such that*

- *For every $x \in \mathcal{X}$, $k(x,x) = 1$*
- *k is uniformly continuous*
- *For every $\varepsilon > 0$, $\inf\{1 - k(x,y) : d(x,y) \geq \varepsilon\} > 0$*
- *$\lim_{\varepsilon \to 0} \sup\{1 - k(x,y) : d(x,y) \leq \varepsilon\} = 0$*

The following corollary extends the previous proposition to uniform local isometries.

Corollary 2 *Let $(\mathcal{X}, d)$ be a metric space and k a positive definite kernel which fulfills the conditions of Proposition 3. If the limits*

$$\limsup_{y \to x} \frac{1 - k(x,y)}{d^2(x,y)}, \qquad \liminf_{y \to x} \frac{1 - k(x,y)}{d^2(x,y)}$$

exist and are non-zero then $\phi_x : x \to k(x,\cdot)$ is a uniform local isometry of $(\mathcal{X}, d)$ onto a subset of the RKHS associated to k.

Proof Simply calculate the metric induced by the positive definite kernel k, $d_k^2(x,y) = 2 - 2k(x,y)$ and use the definition of the functions D_+ and D_- in

Definition 1. The explicit embedding ϕ follows from Lemma 7. $\square$

In principle the above proposition and the corollary are not very satisfying since they provide no explicit construction of a positive definite kernel which fulfills the conditions for a given metric.
In the case where the given metric is a Hilbertian metric we can use a result of Schoenberg. It characterizes the metric transforms F of a given Hilbertian metric d, such that $F(d)$ is also a Hilbertian metric. This implies that the the identity map $\mathsf{id} : (\mathcal{X}, d) \to (\mathcal{X}, F(d))$ is a uniform homeomorphism. Moreover using Lemma 7 we get a uniform embedding into a Hilbert space.

Theorem 4 *[13] Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a function such that $F(0) = 0$ and all derivatives of F exist on $\mathbb{R}_+ \backslash \{0\}$. Then the following assertions are equivalent:*

- *$F(d)$ is a Hilbertian metric, if d is a Hilbertian metric.*
- $$F(t) = \left(\int_0^\infty \frac{1 - e^{-t^2 u}}{u} \, d\gamma(u) \right)^{1/2},$$

  where $\gamma(u)$ is monotone increasing for $u \geq 0$ and satisfying $\int_1^\infty \frac{d\gamma(u)}{u} < \infty$ [3] *.*
- *$(-1)^{n-1} \frac{d^n}{dt^n} F^2(\sqrt{t}) \geq 0$ for all $t > 0$ and $n \geq 1$.*

Moreover F is bounded if and only if

$$\lim_{\epsilon \to 0} \gamma(\epsilon) = \gamma(0), \text{ and } \lim_{\epsilon \to 0} \int_\epsilon^1 \frac{d\gamma(u)}{u} \text{ exists.}$$

For a uniform, local isometric embedding one has to fulfill in addition the requirements of Lemma 2. Combining Theorem 4 and Lemma 2 we get a complete description of all metric transforms for a given Hilbertian metric which induce a uniform local isometry and where the transformed metric is Hilbertian. The examples given in (1) fulfill both the conditions of Theorem 4 and of Lemma 2. Therefore they provide two examples of metric transforms which induce uniform local isometries and produce Hilbertian metrics if one starts with a Hilbertian metric. The drawback of the Theorem 4 is that we have to start with a Hilbertian metric. A more general theorem which characterizes metric transforms of an arbitrary metric space such that the transformed metric is Hilbertian seems not to be available in the literature.

[3] The integrals are Lebesgue-Stieltjes integrals.

In the general considerations we defined the subspace $A \subset \text{span}\{\Psi_x | x \in T\}$ by

$$A := \left\{ \sum_{i=1}^{n} \alpha_i \Psi_{x_i} : \sum_{i=1}^{n} \alpha_i = 0 \right\}.$$

Since in a Hilbert space the dual is isometrically isomorphic to the Hilbert space itself we get the following form of the space of solutions:

Lemma 8 *The space of solutions $\mathcal{H}/A^\perp$ is equal to A.*

Proof We have the simple equalities $\mathcal{H}/A^\perp = (A^\perp)^\perp = A$. □

Following Zhou [17] note that if in (2) the infimum on the left is achieved by $y_0 \in co(T_1)$ and $z_0 \in co(T_2)$ then w' is aligned with $y_0 - z_0$, that is

$$\langle y_0 - z_0, w' \rangle_{\mathcal{H}} = \|y_0 - z_0\|_{\mathcal{H}} \|w'\|_{\mathcal{H}}$$

In a Hilbert space it follows from the Cauchy-Schwarz inequality that in this case $w' = y_0 - z_0$. Therefore in a Hilbert space the problem of maximal margin separation is not only equivalent to the problem of finding the distance of the convex hulls but it has also the same solution. Therefore we can equivalently formulate the problem of maximal margin separation as finding the distance of the convex hulls of the isometrically embedded training data in $\mathcal{H}_k$.
The optimization problem corresponding to the maximum margin hyperplane can be written as

$$\begin{aligned} \min_{\alpha} & \left\| \sum_{i:y_i=+1} \alpha_i \Psi_{x_i} - \sum_{i:y_i=-1} \alpha_i \Psi_{x_i} \right\|^2_{\mathcal{H}_k} \\ \text{subject to: } & \sum_{i:y_i=+1} \alpha_i = \sum_{i:y_i=-1} \alpha_i = 1, \quad \alpha_i \geq 0, \end{aligned}$$

The distance $\left\| \sum_{i:y_i=+1} \alpha_i \Psi_{x_i} - \sum_{i:y_i=-1} \alpha_i \Psi_{x_i} \right\|_{\mathcal{H}_k}$ can be calculated explicitly with the expression of the inner product $\langle \Psi_x, \Psi_y \rangle_{\mathcal{H}_k} = k(x, y)$ from the proof of Proposition 2:

$$\begin{aligned} \left\| \sum_i y_i \alpha_i \Phi_{x_i} \right\|^2_{\mathcal{H}_k} &= \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \\ &= \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (-d^2(x_i, x_j) + d^2(x_i, x_0) + d^2(x_0, x_j)) \\ &= -\frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j d^2(x_i, x_j) \end{aligned}$$

where the other terms vanish because of the constraint $\sum_{i=1}^n y_i\alpha_i = 0$. So the final optimization problem becomes

$$\begin{aligned}&\min_{\alpha} -\frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j d^2(x_i, x_j)\\ &\text{subject to: } \sum_i y_i\alpha_i = 0, \quad \sum_i \alpha_i = 2, \quad \alpha_i \geq 0,\end{aligned}$$

and with $w = \sum_{i=1} y_i\alpha_i\Phi_{x_i}$ the final classifier has the form

$$\begin{aligned}f(x) &= \langle w, \Phi_x\rangle_{\mathcal{H}_k} + b = \sum_{i=1}^{n} y_i\alpha_i k(x_i, x) + b\\ &= -\frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i(d^2(x_i, x) - d^2(x_i, x_0)) + b = -\frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i d^2(x_i, x) + c\end{aligned}$$

The constant c is determined in such a way that the hyperplane lies exactly half way between the two closest points of the convex hulls. Following this consideration the point $m = \frac{1}{2}\sum_{i=1}^n \alpha_i\Phi_{x_i}$ lies on the hyperplane. Then c can be calculated by:

$$c = -\langle w, m\rangle_{\mathcal{H}_k} = \frac{1}{2}\sum_{i,j=1}^{n} y_i\alpha_i\alpha_j(d^2(x_i, x_j) - d^2(x_i, x_0))$$

4.4 Equivalence to the Support Vector Machine

The standard point of view on SVM is that we have an input space $\mathcal{X}$ which describes the data. This input space $\mathcal{X}$ is then embedded via Φ into a Hilbert space $\mathcal{H}$ with a positive definite kernel [4] and then maximal margin separation is done. The following diagram summarizes this procedure:

$$\mathcal{X} \xrightarrow{\text{kernel } k} \mathcal{H}_k \longrightarrow \text{maximal margin separation} \tag{9}$$

where the kernel k is positive definite.
We show now that this is equivalent to the point of view in this paper:

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

where d is a Hilbertian metric.
The next proposition is the key to this equivalence. It is a characterization of

[4] Originally the SVM was only formulated with positive definite kernels. Later it was shown in [14] that due to the translation invariance of the maximal margin problem in feature space one can use the class of conditionally positive definite kernels. In this case the kernel $k(x, y)$ is not equal to an inner product $\langle\Phi_x, \Phi_y\rangle$ in a Hilbert space, but it defines an inner product on a subspace which includes A.

the class of all conditionally positive definite kernels in terms of the class of Hilbertian metrics. It can be found in Berg et al. (see Proposition 3.2 of [3]). We have rewritten it in order to stress the relevant result.

Proposition 4 *All conditionally positive definite kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are generated by a Hilbertian metric $d(x, y)$ in the sense that there exists a function $g : \mathcal{X} \to \mathbb{R}$ such that*

$$k(x, y) = -\frac{1}{2} d^2(x, y) + g(x) + g(y)\,, \tag{10}$$

and any kernel of this form induces the Hilbertian metric d via

$$d^2(x, y) = k(x, x) + k(y, y) - 2k(x, y). \tag{11}$$

This proposition establishes a many-to-one correspondence between the set of conditionally positive definite kernels and Hilbertian metrics. This is rather obvious since already any change of the origin in the RKHS corresponds to a new kernel function on $\mathcal{X}$ but the induced metric (11) is invariant. Moreover the following theorem shows that only the Hilbertian metric d matters for classification with the SVM.

Theorem 5 *The SVM is equivalent to the metric based maximal margin classifier in a Hilbert space. The solution of the SVM does not depend on the specific isometric embedding Φ, nor on the corresponding choice of the kernel in a given family determined by a Hilbertian metric, see* (10)*. The optimization problem and the solution can be completely expressed in terms of the (semi)-metric d of the input space,*

$$\begin{aligned} &\min_{\alpha} \left\| \sum_i y_i \alpha_i \Phi_{x_i} \right\|^2_{\mathcal{H}_k} = -\frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j d^2(x_i, x_j) \\ &\textit{subject to} : \sum_i y_i \alpha_i = 0, \quad \sum_i \alpha_i = 2, \quad \alpha_i \geq 0. \end{aligned}$$

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 Proposition 4 all conditionally positive definite kernels are generated by a Hilbertian metric $d(x, y)$. Using (10) one can show now that for each kernel associated to a Hilbertian metric the corresponding optimization problem for maximal margin separation and the corresponding solution are equivalent to the metric maximal margin classification problem in a Hilbert space for the associated Hilbertian metric.

The expression of the optimization problem of the SVM in terms of the (semi)-metric follows from (10);

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

where the terms with g vanish 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 (10):

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

where again $\sum_i y_i\alpha_i g(x)$ vanishes 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 matter for support vector classifiers. On the contrary the (semi)-metric as a dissimilarity measure of the input space only matters for the maximal margin problem. Nevertheless it seems to be easier to construct a conditionally positive definite kernel than a Hilbertian metric, but one should have in mind that only the induced metric has an influence on the solution, and therefore compare two different kernels through their induced metrics. This should also be considered if one uses eigenvalues of the kernel matrix. They depend on the underlying Hilbertian metric and as well on the function $g(x)$ in (10) whereas the solution of the SVM only depends on the Hilbertian metric. In other words properties which are not uniform over the class of kernels induced by a semi-metric are not relevant for the solution of the SVM.
One could use the ambiguity in the kernel to chose from the whole class of kernels which induce the same (semi)-metric (10) the one which is computationally the cheapest, because the solution does not change as is obvious from the last theorem. Furthermore note that Lemma 8 provides a slight refinement of the usual representer theorem of the SVM which states that the solution lies in an at most n dimensional space spanned by the data (see e.g. [15]). This refinement seems to be a marginal effect for large training sets. However

the crucial point here is that the constraint on the subspace implies that the SVM is actually equivalent to the metric based maximal margin classifier in a Hilbert space.
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 [4,2,17] for this issue.

5 Measuring the capacity via Rademacher averages

In this section we compute the Rademacher averages corresponding to the function classes induced by our embeddings. 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 [9]).

5.1 General case

Given a sample of input points $x_1, \ldots, x_n$, we define the empirical Rademacher average $\widehat{R}_n$ of the function class $\mathcal{F}$ as

$$\widehat{R}_n(\mathcal{F}) := E_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)\,, \tag{12}$$

where σ are Rademacher variables, that are independent uniform random variables with values $\{-1, +1\}$, 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 hyperplanes with a given margin. Now hyperplanes correspond to elements of the dual of the Banach space into which the data is embedded and the margin corresponds to the norm in that space. Therefore we have to consider the Rademacher averages of balls in the dual space.
For a function $f_{w'}$ in $\mathcal{F}_{\mathcal{B}'}$, $f_{w'}(x) = \langle w', \Phi_x \rangle_{\mathcal{B}',\mathcal{B}}$ with $\|f_{w'}\|_{\mathcal{F}_{\mathcal{B}'}} = \|w'\|_{\mathcal{B}'}$ so that

$$E_\sigma \sup_{\|w'\|_{\mathcal{B}'} \leq B} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) = \frac{B}{n} E_\sigma \left\| \sum_{i=1}^{n} \sigma_i \Phi_{x_i} \right\|_{\mathcal{B}} .$$

Notice that even if the embedding Φ is isometric, the above quantity depends on how the $\Phi(x_i)$ are located in the embedded linear space. So, a priori, the above quantity depends on the embedding and not only on the geometry of the input space.

More precisely, we consider the following two classes. For a given positive definite kernel k, let $\tilde{k}$ be defined as $\tilde{k}(x,y) = k(x,y) - k(x,x_0) - k(x_0,y) + k(x_0,x_0)$ [5] and $\mathcal{H}$ be the associated RKHS for $\tilde{k}$. We define $\mathcal{F}_1 = \{g \in \mathcal{H}, \|g\| \leq B\}$. Also, with the notations of the previous section, we define $\mathcal{F}_2 = \{e \in D, \|e\| \leq B\}$.

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 of $\mathcal{F}_2$:

$$\widehat{R}_n(\mathcal{F}_2) = \frac{B}{n} E_\sigma \left\| \sum_{i=1}^{n} \sigma_i \Phi_{x_i} \right\|_\infty = \frac{B}{n} E_\sigma \sup_{x\in\mathcal{X}} \left| \sum_{i=1}^{n} \sigma_i \Phi_{x_i}(x) \right| \tag{13}$$

We will use Dudley's upper bound on the empirical Rademacher average [6] which states 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 $\mathcal{F}_2$,

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}_2, \ell_2^n)}\, d\varepsilon, \tag{14}$$

where $N(\varepsilon, \mathcal{F}_2, \ell_2^n)$ are the covering numbers of the function class $\mathcal{F}_2$ with respect to the ℓ_2 distance on the data, i.e. $\|f-g\|_{\ell_2^n}^2 := \frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2$. 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 $y \in \mathcal{X}$, 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 (13), we get

$$\widehat{R}_n(\mathcal{F}_2) = B\, E_\sigma \sup_{x\in\mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Phi_{x_i}(x) \right| = B\, \widehat{R}_n(\mathcal{G}). \tag{15}$$

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

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

$$\begin{aligned}\|f_{y_1} - f_{y_2}\|_{\ell_2^n} &\leq \max_{x_i \in T} |\Phi_{x_i}(y_1) - \Phi_{x_i}(y_2)| \\ &= \max_{x_i \in T} |d(x_i, y_1) - d(x_i, y_2) + d(x_0, y_2) - d(x_0, y_1)| \\ &\leq 2d(y_1, y_2). \end{aligned} \tag{16}$$

Combining (14) and (16) we arrive at

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

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 $\sqrt{\cdot}$). □

If we can assume that the data is inside a subset of $\mathcal{X}$ with finite diameter R, then this simplifies to

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

The above theorem gives an upper bound on the Rademacher average directly in terms of the covering numbers of the metric space $(\mathcal{X}, d)$.
In particular, this shows that the Rademacher average corresponding to the Kuratowski embedding are much smaller than those corresponding to the Lipschitz embedding of [16]. Indeed, for a bounded subset of the metric space $\mathbb{R}^d$, the covering numbers behave like ϵ^{-d} so that the Rademacher average in our case is of order $\sqrt{d/n}$ while in the Lipschitz case it is of order $(1/n)^{1/d}$.
Notice that a trivial bound on $\widehat{R}_n(\mathcal{F}_2)$ can be found from (13) 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. This is undesirable as it would mean that the bound on the error does not decrease when the sample size is increased.

More interesting than upper bounds on the Rademacher averages of the individual algorithms is to compare them directly in the cases where both algorithms can be applied (i.e. when $-d^2$ is conditionally positive definite). In this case, one can choose to embed isometrically the input space either into a Hilbert space or into a Banach space. The question is then how different balls of same radius in the dual spaces are.

Theorem 7 *If d is a Hilbertian metric, then*

$$\widehat{R}_n(\mathcal{F}_1) \leq \frac{B}{n}\sqrt{\sum_{i=1}^{n} d(x_i, x_0)^2} \leq \sqrt{2}\widehat{R}_n(\mathcal{F}_2)\,.$$

Proof We have

$$\begin{aligned}\widehat{R}_n(\mathcal{F}_2) &= \frac{B}{n} E_\sigma \sup_{x\in\mathcal{X}} \left|\sum_{i=1}^{n} \sigma_i \Phi_{x_i}(x)\right| \geq \frac{B}{n} E_\sigma \left|\sum_{i=1}^{n} \sigma_i \Phi_{x_i}(x_0)\right| \\ &\geq \frac{B}{\sqrt{2}n}\sqrt{E_\sigma \sum_{i,j=1}^{n} \sigma_i\sigma_j \Phi_{x_j}(x_0)\Phi_{x_i}(x_0)} \\ &= \frac{1}{\sqrt{2}}\frac{B}{n}\sqrt{\sum_{i=1}^{n} d(x_i, x_0)^2} \geq \frac{1}{\sqrt{2}}\widehat{R}_n(\mathcal{F}_1)\end{aligned}$$

The second step follows from the Khintchine-Kahane inequality. The constant $1/\sqrt{2}$ is optimal, see e.g. [8]. □

This result can be seen as an indication that the SVM is as good as the general algorithm for arbitrary metric spaces in terms of complexity of the unit ball. However, this does not directly allow to compare the generalization abilities of both algorithms. Indeed, the obtained margin in each case could be quite different.

6 Conclusion and perspectives

In this article we have built a general framework for the generation of maximal margin algorithms for metric spaces. We considered two general cases. In the first one we trust the metric globally, in the second one we believe only in the local structure of the metric which seems to be often the case for metrics defined on real-world data. In the first case we embed directly isometrically into a Banach space, in the second one we first perform a uniform transformation of the metric such that the local structure is preserved and then embed

isometrically the transformed space into a Banach space.
For each metric space we presented a Banach space into which it can be embedded isometrically. It turned out that the optimization problem of the maximal margin algorithm in this Banach space cannot be solved exactly. We provided an approximation which is exact if one considers the training data plus one test point as a finite metric space. One special approximation is the LP-machine for distances of [7].
Since the space of classifiers has a considerably nicer structure if one embeds in a Hilbert space, we considered in the second part isometric embeddings into a Hilbert space. These are no longer possible for all metric spaces, but are restricted to the subclass of Hilbertian metrics. We showed that the resulting algorithm is equivalent to the SVM classifier, but since the relationship between kernels and Hilbertian metrics is many-to-one, the metric based point of view provides a better insight into the structural properties of the SVM.
For the class of Hilbertian metrics we can compare the two isometric embeddings. They both preserve the metric structure, that is, all available information on the data. Therefore the question arises which norm on the linear extension provides the better results in the sense of generalization error. We provided a first answer to this question by comparing the Rademacher averages of both algorithms. It turned out that the Rademacher average of the SVM are upper bounded by a constant times the Rademacher average of the metric based classifier in the Banach space. This result suggests that the SVM has a better generalization performance. But further work has to be done in that direction.

Acknowledgements

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

References

[1] I. Aharoni, B. Maurey, B. S. Mityagin, *Uniform Embeddings of Metric Spaces and of Banach Spaces into Hilbert Spaces*, Israel Journal of Mathematics, **52**(3), 251-265, (1985).

[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] C. J. C. Burges, D. J. Crisp, *Uniqueness of the SVM Solution*, Neural Information Processing Systems (NIPS), **12**, (1999).

[5] M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, Springer Verlag, New York, (1997).

[6] R. M. Dudley, *Universal Donsker Classes and Metric Entropy*, Ann. Prob., **15**, 1306-1326, (1987).

[7] 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).

[8] R. Latala, K. Oleszkiewicz, *On the best constant in the Khintchine-Kahane inequality*, Studia Math., **109**, 101-104, (1994).

[9] P. L. Bartlett, S. Mendelson, *Rademacher and Gaussian Complexities: Risk Bounds and Structural Results*, Journal of Machine Learning Research, **3**, 463-482, (2002).

[10] 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).

[11] W. Rudin, *Functional Analysis*, McGraw Hill, (1991).

[12] I. J. Schoenberg, *Metric Spaces and Positive Definite Functions*, TAMS, **44**, 522-536, (1938).

[13] I. J. Schoenberg, *Metric Spaces and Completely Monotone Functions*, Ann. Math., **39**, 811-841, (1938).

[14] B. Schölkopf, *The Kernel Trick for Distances*, Neural Information Processing Systems (NIPS), **13**, (2000).

[15] B. Schölkopf, A. J. Smola *Learning with Kernels*, MIT Press, MA, Cambridge, (2002).

[16] U. von Luxburg, O. Bousquet, *Distance-Based Classification with Lipschitz Functions*, Journal of Machine Learning Research, **5**, 669-695, (2004).

[17] 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).

A Semi-metric spaces compared to metric spaces for classification

In this article all results were stated for metric spaces. As the following observations show they can be formulated equivalently for semi-metric spaces. In

fact there is a connection between both of them which we want to clarify in this appendix.

Theorem 8 *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 conditionally positive definite kernel and k a positive definite kernel on $\mathcal{X}$ which induces d on $\mathcal{X}$, then $-d^2$ is also a conditionally positive definite kernel and k a positive definite 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

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

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 conditionally positive definite on $\mathcal{X}/\sim$ follows from the fact that all possible representations of equivalence classes are points in $\mathcal{X}$ and $-d^2$ is conditionally positive definite 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 positive definite 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 8 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.