Using Reproducing Kernel Hilbert Spaces to Improve Support Vector Machines

Abhishek Devarajan

November 10, 2021

Abstract

The sharp rise in the availability of data and the demand for said data to be analyzed, has lead to a myriad of different machine learning algorithms. One such algorithm, praised for being simplistic yet highly effective, is the Support Vector Machine (SVM). The secret behind SVMs success is the so called "Kernel Trick"; a clever use of inner products and Hilbert Spaces that turns complicated data into linearly separable sets. In this paper, we will cover the properties of Hilbert Spaces in general and Reproducing Kernel Hilbert Spaces (RKHS) in specific. We will then cover the Kernel Trick and display its power by using a SVM model on a collection of data sets.

1 Introduction

This section should explain the motivation behind RKHS and the kernel trick. Give a high level overview of SVM (maybe an image) and explain why it is useful compared to NN or other algorithms. Introduce the complications faced when using SVM to classify non linearly separable data.

Mention that RKHS and kernels have a history that predates SVM and computing in general (1950s) They are used in fields such as quantum mechanics and functional analysis due to their nice properties.

Give an overview of the following sections.

2 Background

In this section, we will cover the prerequisite knowledge needed to understand the kernel trick and its application for SVM. We start with a description of SVM and its loss function and then continue by discussing a number of definitions leading up to that of a RKHS.

Definition 1 (SVM)**.** *Consider a dataset of n points, $D = \{(x_i, y_i) \mid i = 1, \dots, n\}$ where each x_i is a point in an arbitrary, non-empty input space $\mathcal{X}$ and each label y_i is an element of $\mathcal{Y} = \{-1, 1\}$. Define the hyperplane $h(x)$ as*

$$h(x) = \langle w, x \rangle + b$$

for some $w \in \mathcal{X}$ and $b \in \mathbb{R}$. the Support Vector Machine algorithm involves choosing an optimal w and b such that the sign of $h(x_i)$ is equal to the sign of y_i. To classify a point, $z \notin D$, we take the sign of $h(z)$ and assign z the label in $\mathcal{Y}$ with the same sign.

Definition 2 (SVM Loss Function)**.** *Since complete linear separation is not always possible, we loosen the restriction of having every data point accurately classified. We formulate the loss function of SVM as the primal optimization problem*

$$\begin{aligned} \min \ & \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^{n} \xi_i \\ s.t. \ & y_i \left(\langle w, x_i \rangle + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

where C is a fixed constant and each ξ_i is a slack variable. For the kernel trick, we formulate the corresponding dual problem

$$\begin{aligned} \max \ & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \\ s.t. \ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{1}$$

For now, it suffices to say that this dual problem is equivalent to its primal counterpart [1, 2]. In Section 4 we will revisit Equation (1) *and see its relationship with the kernel trick in more detail.*

Definition 3 (Hilbert Space)**.** *A Hilbert Space $\mathcal{H}$ is a real or complex inner product space such that $\mathcal{H}$ is also a complete metric space with respect to the distance function induced by its inner product.*

Remark 4. *Completeness is a property whose importance is not always obvious when constructing theorems around Hilbert Spaces. Being complete means that every Hilbert Space has an orthonormal basis (**Insert Citation**), which is a property often taken for granted when working in finite dimensional spaces. Crucially for this paper, completeness ensures that we can decompose a Hilbert Space into a direct sum between a closed subspace and its orthogonal complement (**Insert Citation**).*

Example 5 (Types of Hilbert Spaces)**.** *The following a common examples of Hilbert Spaces used in various applications.*

- *Every finite dimensional inner product space*
- $\ell_2 = \{(a_1, a_2, \dots) \mid a_i \in \mathbb{C}, \sum_{i=1}^{\infty} |a_i|^2 < \infty\}$ *with* $\langle a, b \rangle = \sum_{i=1}^{\infty} \overline{b_i} a_i$
- $L^2(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; \int_{-\infty}^{\infty} |f(x)|^2 dx < \infty \right\}$ *with* $\langle f, g \rangle = \int_{-\infty}^{\infty} \overline{g(x)} f(x) dx$

Hilbert Spaces provide us with a set of very nice and intuitive properties thanks to their completeness and inner products. For our purposes, we will build on the structure of a general Hilbert Space and construct an RKHS. From this point on we will limit our discussion to real Hilbert Spaces for simplicity but all of the topics discussed can be generalized to work with complex Hilbert Spaces as well.

Definition 6 (Linear Evaluation Functional)**.** *Let $\mathcal{H}$ be the Hilbert Space of functions and consider the linear function $\mathcal{L}_x : \mathcal{H} \to \mathbb{R}$ given by*

$$\mathcal{L}_x(f) = f(x).$$

$\mathcal{L}_x$ is known as the Linear Evaluation Functional (LEF).

Definition 7 (Reproducing Kernel Hilbert Space)**.** *Suppose $\mathcal{H}$ is a Hilbert Space of real-valued functions with domain $\mathcal{X}$, where $\mathcal{X}$ is any arbitrary set. If $\mathcal{H}$ is such that its linear evaluation functional $\mathcal{L}_x$ is bounded for all $x \in X$ then $\mathcal{H}$ is a RKHS.*

3 Representation Theorem

Theorem 8 (Continuity of Bounded Linear Functionals)**.** *Let $\mathcal{H}$ and $\mathcal{L}_x$ be defined as they were in Definition 6. $\mathcal{L}_x$ is bounded if an only if $\mathcal{L}_x$ is continous.*

Proof. (**Insert Citation**) □

Theorem 9 (Riesz Representation Theorem)**.** *Let $\mathcal{H}$ be a Hilbert Space with inner product $\langle \cdot, \cdot \rangle$. For every continuous functional $\varphi : \mathcal{H} \to \mathbb{R}$, there exists a unique $f_\varphi \in \mathcal{H}$ such that $\varphi(g) = \langle g, f_\varphi \rangle$ for all $g \in \mathcal{H}$*

Proof. Let φ be a continuous linear functional with domain $\mathcal{H}$. We will first prove the existence of a f_φ that satisfies Theorem 9. First, note that if φ is identically zero– i.e. $\varphi(g) = 0\ \forall g \in \mathcal{H}$– then $\varphi(g) = \langle g, 0\rangle\ \ \forall g \in \mathcal{H}$ and we are done. Consider the case where φ is not identically zero. Let $\mathcal{N} = \{h \mid h \in \mathcal{H}, \varphi(h) = 0\} = \varphi^{-1}(\{0\})$ be the null space of φ. Since φ is continuous and $\{0\}$ is a closed subset of $\mathbb{F}$, the Closed Graph Theorem (**insert citation**) states that $\mathcal{N}$ must also be closed. Consequentially, $\mathcal{N}^\perp$– i.e. the orthogonal complement of $\mathcal{N}$– must contain at least one non-zero vector p (**insert citation**). Now define a vector u as follows:

$$\begin{aligned} u &= (\varphi(g))p - (\varphi(p))g \\ \implies \varphi(u) &= \varphi(g)\varphi(p) - \varphi(p)\varphi(p) \\ &= 0 \\ \implies u &\in \mathcal{N}. \end{aligned}$$

By orthogonality,

$$\begin{aligned} 0 &= \langle u, p\rangle \\ &= \langle (\varphi(g))p - (\varphi(p))g, p\rangle \\ &= \varphi(g)\langle p, p\rangle - \varphi(p)\langle g, p\rangle \\ \implies \varphi(g) &= \left\langle g, \frac{p\varphi(p)}{||p||^2}\right\rangle \end{aligned}$$

Setting $f_\varphi = \dfrac{p\varphi(p)}{||p||^2}$ completes the existence proof.

Now, we will show that f_φ is the unique element in $\mathcal{H}$ that satisfies Theorem 9. Suppose to the contrary that there exists two functions f_1, f_2 such that

$$\phi(g) = \langle g, f_1\rangle = \langle g, f_2\rangle\ ,\ \forall g \in \mathcal{H}.$$

$$\begin{aligned} 0 &= \varphi(g) - \varphi(g) \\ &= \langle g, f_1\rangle - \langle g, f_2\rangle \\ &= \langle g, f_1 - f_2\rangle \end{aligned}$$

Since g is an arbitrary element in $\mathcal{H}$, take $g = f_1 - f_2$. This yields

$$\begin{aligned} 0 &= \langle f_1 - f_2, f_1 - f_2\rangle \\ &= ||f_1 - f_2||^2 \\ \implies 0 &= f_1 - f_2 \end{aligned}$$

□

Combining the definition of a RKHS with Theorem 8 and Theorem 9 indicates that we can write any evaluation functional as an inner product between the function being evaluated and a unique function in the RKHS. This is the so-called "reproducing property" that we use to carry out the kernel trick. We formalize this notion using the following notation

Definition 10 (Reproducing Property)**.** *Let $\mathcal{H}$ be a RKHS of functions $g : \mathcal{X} \to \mathbb{R}$ and let $\mathcal{X}$ be a non-empty set. For all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, there exists a unique $K_x \in \mathcal{H}$ such that we can express the evaluation functional $\mathcal{L}_x$ as*

$$\mathcal{L}_x(f) = \langle f, K_x\rangle\ .$$

4 Kernel Trick

In this section, we will present a high-level overview of the kernel trick. From there, we will rigorously outline the kernel trick in terms of the definitions and theorems from previous sections.

Outline 11 (Kernel Trick)**.** *To understand the kernel trick, we need to first consider the loss function of SVM. Typically, the loss function is formulated as an optimization problem. Rather than giving the primal problem, however, we will present the dual problem as it better illustrates the purpose of the kernel trick*

Definition 12 (Reproducing Kernel)**.** *Let $\mathcal{X}$ be an arbitrary, non-empty set. a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if it can be represented as $k(x, z) = \langle \phi(x), \phi(z) \rangle$ where $x, z \in \mathcal{X}$ and $\phi : \mathcal{X} \to \mathcal{H}$.*

*The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** corresponding to a Hilbert Space $\mathcal{H}$ if the following are satisfied*

1. *The function $k(x, \cdot)$, where x is fixed, is an element of $\mathcal{H}$.*
2. *$f(z) = \langle f, k(z, \cdot) \rangle \ \ \forall z \in \mathcal{X}$*

Theorem 13. *If $\mathcal{H}$ is a Reproducing Kernel Hilbert Space, then there exists a corresponding unique reproducing kernel k. Conversely, if k is a reproducing kernel, there exists some RKHS $\mathcal{H}$ corresponding to k.*

Proof. For the proof of the forward direction, we point to □

Consider a RKHS of functions $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{H}$. Remember from Definition 10 that there exists a unique $K_x \in \mathcal{H}$ for each $x \in \mathcal{X}$ such that $\mathcal{L}_x(f) = \langle f, K_x \rangle$ for all $f \in \mathcal{H}$. In particular we have

$$\begin{aligned} & \mathcal{L}_z(f) = \langle f, K_z \rangle \\ \implies & \mathcal{L}_z(K_x) = \langle K_x, K_z \rangle . \end{aligned}$$

Define $k(x, z) = \langle K_x, K_z \rangle$ and notice that k is a kernel. To see this more clearly, consider the function $\phi : \mathcal{X} \to \mathcal{H}$ as $\phi(x) = K_x$. In particular, we define k as a reproducing kernel.