thesis

Anh Van Giang (V)

May 26, 2022

Abstract

Contents

Introduction

Dimensionality Reduction

Optimization Theory

Optimization theory is a branch of mathematics that concerns itself with characterizing the solutions of finding a set of parameters that minimizes (or maximizes) a certain cost function, usually with respect to some constraints. This section aims to provide a brief introduction of the field of optimization and the methods used to solve the Support Vector Machine (SVM) problem.

3.1 Problem Formulation

In the context of real-world applications not restricted to SVM, one may encounter problems that are of the form of maximizing or minimizing a function w.r.t some constraints. These problems can be formulated as:

Definition 3.1.1. (Primal optimization problem) Given functions f, g_i, $i = 1, \dots k$, and h_i, $i = 1, \dots, m$, defined on a domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} \min \quad & f(w), \quad w \in \Omega \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \dots, k, , \\ & h_i(w) = 0, \quad i = 1, \dots, m, \end{aligned} \tag{3.1}$$

where $f(w)$ is called the *objective* function, and the remaining relations are called, respectively, the inequality and equality constraints. The optimal value of the objective function is called the value of the optimization problem.

Notice that we can easily convert a maximization problem to a minimization one by reverting the sign of $f(w)$. Using the optimization problem in Definition 3.1.1, define

$$R = \{w \in \Omega \mid g_i(w) \leq 0, h_i(w) = 0\}$$

as the feasible region i.e the region of the domain where the objective function is defined and all the constraints are satisfied.
A solution of the optimization problem is a point $w^* \in R$ such that for any $w \in R$, $f(w) > f(w^*)$, also known as the global minimum (or maximum).

Definition 3.1.2. A point $w^* \in \Omega$ is called a *local minimum* of $f(w)$ if there exists some $\epsilon > 0$ such that for all w that satisfies $\|w - w^*\| < \epsilon$, $f(w^*) \leq f(w)$.

An inequality constraint $g_i(w) \leq 0$ is said to be *active* if the solution w^* satisfies $g_i(w^*) = 0$, i.e, on the boundary, and it is said to be *inactive* otherwise. To transform an inequality constraint into an equality one, *slack variables* denoted as ξ can be applied as follow:

$$g_i(w) \leq 0 \iff g_i(w) + \xi_i = 0, \text{ with } \xi_i \geq 0.$$

These variables will be used extensively when we introduce Soft Margin SVM, to indicate a certain amount of "looseness" in the constraint. Most if not all the loss functions used in this paper will be convex so we will restrict the content accordingly.

Definition 3.1.3. A set $\Omega \subseteq \mathbb{R}^n$ is convex, if for all $x, y \in \Omega$, and for all $\theta \in [0, 1]$

$$\theta x + (1 - \theta)y \in \Omega.$$

A function $f : \Omega \to \mathbb{R}$ is called *convex* if for all $w, u \in \Omega$, and for $\theta \in [0, 1]$,

$$f(\theta w + (1 - \theta)u) \leq \theta f(w) + (1 - \theta)f(u).$$

If a strict inequality holds then the function is said to be *strictly convex.*

The reason why we mainly use convex functions in this paper will be clear from the following proposition.

Proposition 3.1.1. If a function $f : \Omega \to \mathbb{R}$ is convex, then any local minimum of the unconstrained optimization with the objective function f is also a global minimum.

Proof. Let $w^* \in \Omega$ be a local minimum then there exists $\epsilon > 0$ s.t

$$f(w^*) \leq f(w),\ \forall w \in B(w^*, \epsilon) = \{w : \|w - w^*\| < \epsilon\}.$$

Suppose that there is a $u \in \Omega$ with

$$f(u) < f(w^*)$$

then because Ω is convex, we have

$$\theta w^* + (1-\theta)u \in \Omega, \quad \forall \theta \in [0,1].$$

By the convexity of f,

$$\begin{aligned} f(\theta w^* + (1-\theta)u) &\leq \theta f(w^*) + (1-\theta) f(u) \\ &< \theta f(w^*) + (1-\theta) f(w^*) \\ &= f(w^*). \end{aligned}$$

Choose a θ sufficiently close to 1 such that $1 \leftarrow \theta w^* + (1-\theta)u \in B(w^*, \epsilon)$ but then $f(w^*) < f(w^*)$ which is a contradiction. □

Thus, by imposing the convexity condition on our optimization problem, we can guarantee that a unique solution exists. The next part will present the Lagrange multipliers technique to solve convex quadratic optimization problem.

3.2 Lagrangian Theory

The methods of Lagrange multipliers and the Lagrangian function were first developed by Lagrange in 1797 [1] and extended by Kuhn, Tucker [2] in 1951 to allow for inequality constraints. These theories will provide the sufficient solutions for the Support Vector Machine problem.

Theorem 3.2.1. *(Fermat) A necessary condition for w^* to be a minimum of $f(w)$, $f \in C^1$, is*

$$\frac{\partial f(w^*)}{\partial w} = 0.$$

This condition, together with convexity of f, is also a sufficient condition.

Definition 3.2.1. Given an optimization problem with objective function $f(w)$, and equality constraints $h_i(w) = 0$, $i = 1, \ldots, m$, we define the *Lagrangian function* as

$$L(w, \alpha) = f(w) + \sum_{i=1}^{m} \alpha_i h_i(w)$$

where the coefficients α_i are called the *Lagrange multipliers.*

The Lagrangian function incorporates information about both the objective function and the constraints, whose stationary points can be used to find solutions.

Theorem 3.2.2. *(Lagrange) A necessary condition for a normal point w^* to be a minimum of $f(w)$ subject to $h_i(w) = 0$, $i = 1, \ldots, m$ with $f, h_i \in C^1$, is*

$$\frac{\partial L(w^*, \alpha^*)}{\partial w} = 0,$$
$$\frac{\partial L(w^*, \alpha^*)}{\partial \alpha} = 0,$$

for some values α^*. The above conditions are also sufficient provided that $L(w, \beta^*)$ is a convex function of w.

Proof. The exact proof of this theorem is beyond the scope of this paper, interested readers may find the proof here [3]. □

A simple geometric interpretation of this theorem with only 1 equality constraint can be given as trying to "climb" a surface defined by $f(w)$ subjected to a path $h(w) = 0$. Then, the solution would be the point where the gradient of the surface is parallel to the gradient of the curve, $\nabla f = \lambda \nabla h$, $h \in \mathbb{R}$. It can be easily shown that Theorem 3.2.2 is equivalent to the preceding formulation in conjunction with $h(w) = 0$.

More often than not, one will encounter optimization problems with inequality constraints mixed in such as during a SVM problem. In light of this, we will introduce the generalized Lagrangian.

Definition 3.2.2. Given an optimization with domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} \min_{w \in \Omega} \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k, \\ & h_i(w) = 0, \quad i = 1, \ldots, m, \end{aligned} \tag{3.2}$$

we define the *generalised Lagrangian function* as

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w).$$

Definition 3.2.3. The *Lagrangian dual problem* of the primal problem of Definition 3.1.1 is the following problem:

$$\begin{aligned} \max \quad & \theta(\alpha, \beta) \\ \text{s.t.} \quad & \alpha \geq 0 \end{aligned} \tag{3.3}$$

where $\theta(\alpha, \beta) = \inf_{w \in \Omega} L(w, \alpha, \beta)$.

The relationship between the dual and primal problems will be clear after the following theorems.

Theorem 3.2.3. *(Weak duality theorem) Let $w \in \Omega$ be a feasible solution of the primal problem of 3.1.1 and (α, β) a feasible solution of the dual problem of 3.3. Then $f(w) \geq \theta(\alpha, \beta)$.*

Proof. By definition,

$$\begin{aligned} \theta(\alpha, \beta) &= \inf_{u \in \Omega} L(u, \alpha, \beta) \\ &\leq L(w, \alpha, \beta) \\ &= f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w) \leq f(w), \end{aligned}$$

since $g_i(w) \leq 0$ and $h_i(w) = 0$, $\alpha \geq 0$ by the feasibility of w and (α, β). □

Hence, it is not difficult to see that the value of the dual formation is upper bounded by the value of the primal,

$$\sup\{\theta(\alpha, \beta) : \alpha \geq 0\} \leq \inf\{f(w) : g(w) \leq 0, h(w) = 0\}.$$

Corollary 3.2.1. If $f(w^*) = \theta(\alpha^*, \beta^*)$, where $(\alpha^*, \beta^*), w^*$ are feasible then they solve the dual and primal problems respectively. In this case, $\alpha_i^* g_i(w^*) = 0$, for $i = 1, \ldots, k$.

Proof. Since $f(w^*) = \theta(\alpha^*, \beta^*)$ and $\alpha_i^* g_i(w^*) = 0$, for $i = 1, \ldots, k$,

$$\begin{aligned} \theta(\alpha^*, \beta^*) &= f(w^*) \\ &= f(w^*) + \sum_{i=1}^{k} \alpha_i^* g_i(w) + \sum_{i=1}^{m} \beta_i^* h_i(w) \\ &= L(w^*, \alpha^*, \beta^*) \\ &= \inf_{u \in \Omega} L(u, \alpha^*, \beta^*). \end{aligned}$$

□

By comparing the primal and dual values in parallel and check the *duality gap*i.e their difference, one may be able to find the optimal solution if this reduces to zero. However, it is not generally guaranteed that the primal and dual problems will have the same values as solution.

Theorem 3.2.4. *(Strong duality theorem) Given an optimization problem with convex domain $\Omega \subset \mathbb{R}^n$,*

$$\begin{aligned} \min_{w \in \Omega} \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k, \\ & h_i(w) = 0, \quad i = 1, \ldots, m, \end{aligned} \tag{3.4}$$

where g_i, h_i are affine functions and f convex then the duality gap is zero.

Proof. See [4]. □

The following theorem given by Kuhn-Tucker states the conditions for an optimal solution to a general optimization problem.

Theorem 3.2.5. *(Kuhn-Tucker) Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^n$,*

$$\begin{aligned} \min_{w \in \Omega} \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k, \\ & h_i(w) = 0, \quad i = 1, \ldots, m, \end{aligned} \tag{3.5}$$

with $f \in C^1$ convex and g_i, h_i affine. The necessary and sufficient conditions for a normal point w^* to be an optimum are the existence of α^*, β^* such that

$$\begin{aligned} \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial w} &= 0, \\ \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial \beta} &= 0, \\ g_i(w^*) &\leq 0, \ i = 1, \ldots, k, \\ \alpha_i^* g_i(w^*) &= 0, \ i = 1, \ldots, k, \\ \alpha_i^* &\geq 0, \ i = 1 \ldots, k. \end{aligned}$$

Proof. See [4]. □

The second and third conditions are necessary to satisfy the primal feasibility while the last condition satisfies the dual's. The fourth condition is usually known as Karush-Kuhn-Tucker complementary condition, it implies that for active constraints, $\alpha_i^* \geq 0$, whereas for inactive constraints $\alpha_i^* = 0$. This also follow from 3.2.1 since we are assuming zero duality gap.

The dual representation of a primal problem often turns out to be easier to solve since handling inequality constraints directly is difficult. The primal can be transformed into a dual by setting the derivatives of the Lagrangian w.r.t the primal variables and substituting the obtained relations back into the Lagrangian therefore removing the dependence on said variables. This corresponds to computing the function

$$\theta(\alpha, \beta) = \inf_{w \in \Omega} L(w, \alpha, \beta).$$

Which leaves us with the step of maximizing the resulting function under a (much) simpler constraint.

Support Vector Machine

Given a set of (linearly separable) data points and labels like in Figure 1 . There are infinitely many hyperplanes that can be used to separate the two classes but optimally, we would want a hyperplane that can generalize well for unseen data. If a hyperplane that is too close to the data points is chosen then the permitted margin of error would be to small to predict unknown data accurately because it is easier for them to fall on either sides of the hyperplanes. Thus, an ideal separator would be the one that has the largest margin i.e the distance between the nearest data points to the plane is greatest [5].

4.1 Hard Margin SVM

We start off with the assumption that there exists a hyperplane that can perfectly separate or classifies the dataset.

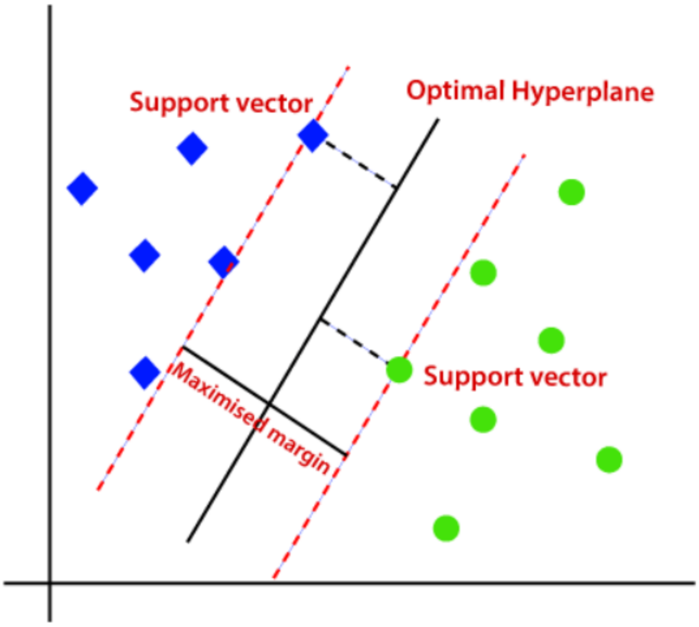

Figure 1: Support vectors [6]

To put it concretely, let $S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ be the set of data points with $x_i \in \mathbb{R}^m$ and $y_i \in \{1, -1\}$ as the labels. Let $h = w \cdot x + b$ be an arbitrary hyperplane then define the positive and negative support vectors, x^+ and x^-, of this hyperplane as the points closest to h. Note that h, ch, for $c \in \mathbb{R}$, define the same hyperplane so we can choose our normalization factor such that

$$\langle w, x^+\rangle + b = 1$$
$$\langle w, x^-\rangle + b = -1.$$

Then the set S is called linear separable if there exists a hyperplane defined by (w, b) such that

$$\begin{cases} \langle w, x_i\rangle + b \geq 1 & \text{for } y_i = 1 \\ \langle w, x_i\rangle + b \leq -1 & \text{for } y_i = -1 \end{cases}$$

or

$$y_i(\langle w, x_i\rangle + b) \geq 1 \quad \forall i \in \{1, 2, \ldots, n\}.$$

The (geometric) margin is then given by

$$\gamma = \frac{|w \cdot x^+ + b|}{\|w\|} + \frac{|w \cdot x^- + b|}{\|w\|} = \frac{2}{\|w\|}$$

where $\|.\|$ is the standard Euclidean norm. Recall that the objective is to find a hyperplane such that γ is maximized while being subjected to $y_i(w \cdot x_i + b) \geq 1$ i.e

$$\begin{aligned} \max_{w,b} \quad & \gamma \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1, \\ & i = 1, 2, \ldots, n \end{aligned}$$

which is equivalent to

$$\begin{aligned} \min_{w,b} \quad & \frac{\|w\|^2}{2} \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1, \\ & i = 1, 2, \ldots, n. \end{aligned} \tag{4.1}$$

i.e a quadratic optimization problem. The formulations stated above belong to the Hard Margin SVM problem because the hyperplane (w, b) must classify each point correctly. In practice, more often than not, this is unachievable and impractical for a variety of reasons. Thus, ideally, it is desirable to have a classifier that allow for "minor" mistakes while retaining its generalization. Note that there are many versions to the Hard Margin SVM problem [7].

Following the steps outlined in Section 3, the primal Lagrangian is

$$L(w, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} \alpha_i [y_i(\langle w, x_i \rangle + b) - 1]$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

Differentiating the Lagrangian w.r.t to w, b to obtain

$$\begin{aligned}
\frac{\partial L(w,b,\alpha)}{\partial w} &= w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \implies w = \sum_{i=1}^{n} \alpha_i y_i x_i, \\
\frac{\partial L(w,b,\alpha)}{\partial b} &= \sum_{i=1}^{n} y_i \alpha_i = 0.
\end{aligned}$$

Resubstituting the above relations into the primal Lagrangian to obtain

$$\begin{aligned}
L(w, b, \alpha) &= \frac{\langle w, w \rangle}{2} - \sum_{i=1}^{n} \alpha_i [y_i(\langle w, x_i \rangle + b) - 1] \\
&= \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle.
\end{aligned}$$

We have the resulting dual problem
c

Remark. The relation $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ shows that the weights w can be described as a linear combination of the training data. This is called the dual representation.

Remark. Recall the Kuhn-Tucker complimentary condition 3.2.5 from the previous chapter that for a solution α^*, (w^*, b^*) to be optimal, it must satisfy

$$\alpha_i^* [y_i(\langle w^*, x_i \rangle + b^*) - 1)] = 0, \quad i = 1, \ldots, n.$$

Define $\langle w, x_i \rangle + b$ as the functional margin of x_i w.r.t (w, b). This implies that for points x_i which the functional margin is 1 and lies closest to the hyperplane has non-zero α_i^* and vice versa. Thus, the hyperplane is only determined by the points closest to it i.e, support vectors.

Having solved the dual problem to obtain the Lagrange multipliers, we can easily find the weights w but b is still unknown. In [8], b is calculated by taking the average of $y_i - \langle w, x_i \rangle$ i.e

$$b = \frac{1}{S} \sum_{i=1}^{S} (y_i - \langle w, x_i \rangle)$$

where S is the number of support vectors. Other authors such as [9] only take the average of the nearest positive and negative support vectors

$$b = -\frac{\max_{y_i=-1} \langle w, x_i \rangle + \min_{y_i=1} \langle w, x_i \rangle}{2}.$$

4.2 Soft Margin SVM

Supposed that the data is not linearly separable, then the optimization problem 4.1 can not be solved since the condition $y_i(\langle w, x_i \rangle + b)$ is not satisfiable. Also, if the data is linearly separable but the separating hyperplane does not leave a wide enough margin to account for unseen data then this would indeed be a bad classifier. Thus arise the need to alter the original Hard Margin SVM problem to account for the trade-off between errors and generalizability.

Instead of forcing our classifier to be correct at every point x_i, we introduce the *slack* variables ξ_i to our constraint as

$$\begin{aligned} \min_{w,\, b} \quad & \frac{1}{2}\langle w, w\rangle \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & i = 1, 2, \ldots, n. \end{aligned} \tag{4.2}$$

Clearly, we could let $\xi_i \to \infty$ and the constraint would be satisfied but this has no added-value to the problem so we would want to constraint the values of ξ_i. There are many formulation of this approach but we shall use the one introduced by [9],

$$\begin{aligned} \min_{w,\, b,\, \xi} \quad & \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n} \xi_i^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & i = 1, 2, \ldots, n \end{aligned} \tag{4.3}$$

for some $C \in \mathbb{R}$. This problem is called L_2-norm soft margin. The Lagrangian and dual problem can then be obtained in the same way as the hard margin one.

The constant C was introduced as a way to control the importance of the *slack* variables. If $C \to \infty$ then it is necessary that $\xi_i \to 0$ for $i = 1, \ldots, n$ and thus become a Hard Margin SVM problem. Each different C can lead to a different classifier that may or may not be satisfactory so to find the optimal C, we have to try different values. Some of the recommended approaches can be found in [10] but they are beyond the scope of this paper.

So far, we have only assumed that the data is either linearly separable or linear separable but with a few outliers that can still be somewhat accurately classified using a separating hyperplane. What if, the data is non-linearly separable in such a way that a linear separator would not yield satisfactory results. Hence, it is necessary that we introduce a very important method to transform our feature vectors in such a way that they become linearly separable.

Reproducing Kernel Hilbert Space, Kernel Method and The Representer Theorem

This section aims to introduce the *kernel* method that is used to transform a feature vector into a higher or infinite dimensional space with the aim of making the dataset linearly separable. We will also present related and necessary objects to supplement the kernel method such as Hilbert space, Reproducing Kernel Hilbert Space (RKHS) and the Representer Theorem. Although there are many publications that omit these definitions, we decline to do the same since they are crucial for the goal of this paper, that is Laplacian SVM.

5.1 Positive Definite Kernels

We begin with a few necessary definitions of basic spaces to build up to the Hilbert space.

Definition 5.1.1. Let X be a non-empty set with a distance function (or metric) $d : X \times X \to \mathbb{R}^+$ then X is a metric space if for all $x, y, z \in X$, d satisfies the following conditions:

1. $d(x, y) = 0 \iff x = y$.
2. $d(x, y) = d(y, x)$.
3. $d(x, y) + d(y, z) \geq d(x, z)$.

Example 5.1.1. The set of real numbers $\mathbb{R}$ forms a metric space with $d(x, y) = |x - y|$.

Definition 5.1.2. Let V be a vector space over $\mathbb{R}$. A norm function on V is a function $\|.\| : V \to \mathbb{R}$ such that for all $v, w\ inV$ and $a \in \mathbb{R}$,

1. $\|v\| \geq 0$ with $\|v\| = 0 \iff v = 0$.
2. $\|av\| = |a|\, \|v\|$.
3. $\|v + w\| \leq \|v\| + \|w\|$.

This norm induces a metric $d(v,w) = \|v-w\|$ on V.

Example 5.1.2. The space $\mathbb{R}^n$ is a vector space over $\mathbb{R}$ with the L_2 norm:

$$\|v\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

Definition 5.1.3. A normed vector space V is complete if every Cauchy sequence in V converges to a vector $v \in V$ with respect to the norm.

Example 5.1.3. The normed vector space $(\mathbb{Q}, \|.\|)$ is not complete because if a sequence is recursively defined as

$$x_0 = \frac{4}{3}$$
$$\forall n \in \mathbb{N} : x_{n+1} = \frac{4+3x_n}{3+2x_n}$$

then it can be shown [11] that this sequence is Cauchy and converges to $\sqrt{2}$ which is not in $\mathbb{Q}$.

Definition 5.1.4. Let V be a vector space over $\mathbb{R}$. An inner product on V is a function $\langle .,. \rangle : V \times V \to \mathbb{R}$ such that for any $v, w, u \in V$ and $a, b \in \mathbb{R}$, we have

1. $\langle u, v \rangle = \langle v, u \rangle$.
2. $\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle$.
3. $\langle v, v \rangle \geq 0$ with $\langle v, v \rangle = 0 \iff v = 0$.

A vector space that is equipped witn an inner product is called an inner product space.

Example 5.1.4. The vector space $\mathbb{R}^n$ is an inner product space with the inner product as the dot product.

Definition 5.1.5. Let $(V, \langle .,. \rangle)$ be an inner product space, the norm induced by $\langle .,. \rangle$ can be defined as

$$\|v\| = \sqrt{\langle v, v \rangle} \text{ for } v \in V.$$

Remark. Not every norm is induced by an inner product. In fact, it holds if and only if the parallelogram identity

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

holds.

Definition 5.1.6. A Hilbert space is an inner product space that is complete with respect to the norm induced by the inner product.

Hilbert space is important to our research because of its rich structure and the methods that we are going to introduce, all operate within Hilbert space. We now can define what is a kernel.

Definition 5.1.7. Let X be a non-empty set, let H be a (complex) Hilbert space and $\varphi : X \to H$ be a function called a *feature map.* The function $\kappa : X \times X \to \mathbb{C}$ given by

$$K(x,y) = \langle \varphi(x), \varphi(y) \rangle, \quad x, y \in X,$$

is called a *kernel function.*

Example 5.1.5. Let $X \subset \mathbb{R}^2$, and let $\varphi : X \to \mathbb{R}^3$ be the map given by

$$\varphi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

We have

$$\begin{aligned}\langle \varphi(x_1,x_2), \varphi(y_1,y_2) \rangle &= \langle (x_1^2, x_1^2, \sqrt{2}x_1x_2), (y_1^2, y_1^2, \sqrt{2}y_1y_2) \rangle \\ &= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 \\ &= (x_1y_1 + x_2y_2)^2 = \langle x, y \rangle,\end{aligned}$$

where $\langle x, y\rangle$ is the usual inner product on $\mathbb{R}^2$. Hence,

$$K(x,y) = \langle x, y\rangle^2$$

is a kernel function associated with the feature space $\mathbb{R}^3$.

One of the reason why the kernel function is often used in various contexts of Machine Learning is that, as you have seen, the calculation and derivation of the feature map φ is time-consuming and resource intensive while it could be replaced by a simple inner product. Furthermore, if the Hilbert space H is infinite dimensional then clearly, computing φ is impossible but nevertheless, the expression $K(x, y)$ can be found, avoiding φ altogether.

Using the kernel function, we can then transform the hard margin dual problem into

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ \text{s.t.} \quad & \alpha_i \geq 1 \quad i = 1, \dots, n, \\ & \sum_{i=1}^{n} y_i \alpha_i = 0 \quad i = 1, \dots, n \end{aligned} \tag{5.1}$$

which can then be solved similarly to the old one with the exception that the dataset is now linearly separable, assuming that it is indeed the case in the feature space.

Remark. The same kernel can arise from different maps into different feature spaces. Example: $\varphi_1(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ in $\mathbb{R}^3$ and $\varphi_2(x_1, x_2) = (x_1^2, x_2^2, x_1x_2, x_1x_2)$ in $\mathbb{R}^4$.

One of the important property of the kernel functions is the positive definite-ness and the reason for its importance will become clear by the end of this section.

Definition 5.1.8. Let X be a non-empty set. A function $\kappa : X \times X \to \mathbb{C}$ is a *positive definite kernel* if for every finite $S \subset X$, if K_s is the $p \times p$ matrix

$$K_s = (K(x_j, x_i))_{1 \leq i,j \leq p}$$

then we have

$$u^* K_s u = \sum_{i,j=1}^{p} K(x_i, x_j) u_i \overline{u_j} \geq 0, \quad \text{for all } u \in \mathbb{C}^p$$

where u^* denotes the complex conjugate of u.

Remark. A positive definite kernel that is also symmetric i.e $K(x, y) = K(y, x)$ is called a $MercerKernel$.

Proposition 5.1.1. Let X be any non-empty set, let H be any (complex) Hilbert space, let $\varphi : X \to H$ be any function, and let $K : X \times X \to \mathbb{C}$ be the kernel given by

$$K(x, y) = \langle \varphi(x), \varphi(y)\rangle, \quad x, y \in X.$$

For any finite subset $S = \{x_1, \dots, x_p\}$ of X, if K_s is the $p \times p$ matrix

$$K_s = (K(x_j, x_i))_{1 \leq i,j \leq p}$$

then we have

$$u^* K_s u \geq 0, \quad \textit{for all } u \in \mathbb{C}^p.$$

Proof.

$$\begin{aligned} u^* K_s u &= \sum_{i,j=1}^{p} K(x_i, x_j) u_i \overline{u_j} \\ &= \sum_{i,j=1}^{p} \langle \varphi(x), \varphi(y)\rangle u_i \overline{u_j} \\ &= \left\langle \sum_{i=1}^{p} u_i \varphi(x_i), \sum_{j=1}^{p} u_j \varphi(x_j) \right\rangle \\ &= \left\| \sum_{i=1}^{p} u_i \varphi(x_i) \right\|^2 \geq 0. \end{aligned}$$

□

Proposition 5.1.2. (I. Schur) If $K_1, K_2 : X \times X \to \mathbb{C}$ are two positive definite kernels, then the function $K : X \times X \to \mathbb{C}$ given by $K(x, y) = K_1(x, y)K_2(x, y)$ for all $x, y \in X$ is also a positive definite kernel.

Proof. [7]. □

Here are some ways of obtaining new positive definite kernels from old ones

Proposition 5.1.3. Let $K_1 : X \times X \to \mathbb{C}$ and $K_2 : X \times X \to \mathbb{C}$ be two positive definite kernels, let $f : X \to \mathbb{C}$, $\psi : X \to \mathbb{R}^N$ be functions, $K_3 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ be a positive definite kernel, and $a \in \mathbb{R}^+$, $p(z)$ be a polynomial with nonnegative coefficients. Then the following functions are positive definite kernels:

(1) $K(x, y) = K_1(x, y) + K_2(x, y)$.

(2) $K(x, y) = aK_1(x, y)$.

(3) $K(x, y) = K_3(\psi(x), \psi(y))$.

(4) $K(x, y) = f(x)\overline{f(y)}$.

(5) $K(x, y) = p(K_1(x, y))$.

(6) $K(x, y) = e^{K_1(x,y)}$.

(7) If X is a real Hilbert space with inner product $\langle -, - \rangle_X$ and corresponding norm $\| \ \|_X$,

$$K(x, y) = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

for any $\sigma > 0$.

Proof. $(1), (2)$ and (3) are trivial.

(4) Let $S = \{x_1, \dots, x_p\} \subset X$, if K is the $p \times p$ matrix

$$K = (\overline{f(x_k)}f(x_j))_{1 \le j,k \le p}$$

then we have

$$u^*Ku = \sum_{k,j=1}^{p} u_j f(x_j)\overline{u_k f(x_k)} = \left| \sum_{j=1}^{p} u_j f(x_j) \right|^2 \ge 0.$$

(5) Let $p(z) = \sum_{i=0}^{m} a_i z^i$, then

$$p(K_1(x, y)) = a_m K_1(x, y)^m + \cdots + a_1 K_1(x, y) + a_0.$$

Since $a_i \in \mathbb{R}^+$ for $i = 0, \dots, m$, by Proposition 5.1.2 and (2), each $a_i K_i(x, y)^i$ is a positive definite kernel. By $(4), (1)$ with $f(x) = \sqrt{a_0}$, $p(K_1(x, y))$ is a positive definite kernel.

(6) We first show that if each $K_i : X \times X \to \mathbb{C}$ is a positive definite kernel,

$$\lim_{i \to \infty} K_i(x, y) = K(x, y)$$

is also a positive definite kernel if it exists. We have

$$\begin{aligned} \sum_{j,k=1}^{n} u_j u_k K(x_j, x_k) &= \sum_{j,k=1}^{n} u_j u_k \left(\lim_{i \to \infty} K_i(x_j, x_k) \right) \\ &= \sum_{j=k=1}^{n} \lim_{i \to \infty} (u_j u_k K_i(x_j, x_k)) \\ &= \lim_{i \to \infty} \underbrace{\left(\sum_{j,k=1}^{n} u_j u_k K_i(x_j, x_k) \right)}_{\ge 0} \ge 0. \end{aligned}$$

Thus, K is positive definite. Note that

$$e^{K_1(x,y)} = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{K_1(x,y)^k}{k!}$$

but each partial sum

$$\sum_{k=0}^{n} \frac{K_1(x,y)^k}{k!}$$

is a positive definite kernel so $e^{K_1(x,y)}$ is also a positive definite kernel.

(7) By (2) and since the map $(x,y) \to \langle x,y\rangle_X$ is a positive definite kernel by Proposition 5.1.1 with identity feature map, the function

$$(x,y) \to \frac{\langle x,y\rangle_X}{\sigma^2}$$

is a positive definite kernel so it follows that

$$K_1 = e^{\frac{\langle x,y\rangle_X}{\sigma^2}}$$

is also a positive definite kernel. Let $f : X \to \mathbb{R}$ be defined as

$$f(x) = e^{-\frac{\|x\|_X^2}{2\sigma^2}}$$

then by (4),

$$K_2(x,y) = f(x)\overline{f(y)} = f(x)f(y) = e^{-\frac{\|x\|_X^2}{2\sigma^2}} e^{-\frac{\|y\|_X^2}{2\sigma^2}} = e^{-\frac{\|x\|_X^2+\|y\|_X^2}{2\sigma^2}}$$

is a positive definite kernel. Thus,

$$\begin{aligned} K_1(x,y)K_2(x,y) &= e^{\frac{2\langle x,y\rangle_X}{2\sigma^2}} e^{-\frac{\|x\|_X^2+\|y\|_X^2}{2\sigma^2}} \\ &= e^{-\frac{\|x-y\|_X^2}{2\sigma^2}} \end{aligned}$$

is a positive definite kernel. It is usually called *Gaussian kernel.*

□

5.2 Reproducing Kernel Hilbert Space

We are ready to introduce the Reproducing Kernel Hilbert Space (RKHS) which is a Hilbert space of functions with the reproducing property. The reason why this space is important has to do with the *Moore-Aronszajn theorem* which connect positive definite kernels to feature space and allow the kernel trick to work.

Definition 5.2.1. A Reproducing Kernel Hilbert Space is a Hilbert space $\mathcal{H}$ of functions $f : X \to \mathbb{R}$ with a reproducing kernel $K : X \times X \to \mathbb{R}$ where $K(x,.) \in \mathcal{H}$ and $f(x) = \langle K(x,.), f\rangle$.

Remark. The property $f(x) = \langle K(x,.), f\rangle$ is called the *reproducing* property.

Theorem 5.2.1 (Moore-Aronszajn)**.** *If K is a Mercer kernel on a set X. Then there is a unique RKHS space of functions on X for which K is the reproducing kernel.*

Proof. See [12]. □

This theorem combines with the definition of RKHS states that for each RKHS we have a reproducing Mercer kernel and for each Mercer kernel, we can get a unique RKHS.

To put it more concretely, suppose that instead of wanting to compute $\langle x_i, x_j\rangle$ in the dual formulation of the Hard Margin SVM problem

$$L(w,b,a) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle,$$

we want to compute the inner product of the feature map $\varphi : X \to H$, for H a Hilbert space, $\langle \varphi(x_i), \varphi(x_j)\rangle$ in a higher dimensional space. This is potentially problematic because it requires us to explicitly compute the feature map which can be computationally expensive or impossible if the dimension of H is infinite. However, if instead one chooses a Mercer kernel K then by 5.2.1, there exists a unique RKHS $\mathcal{H}$ where K has the reproducing property. Then this becomes much easier by letting $\varphi(x) = K_x \in \mathcal{H}$ then

$$K(x_i, x_j) = \langle K_{x_i}(.), K_{x_j}(.)\rangle = \langle \varphi(x_i), \varphi(x_j)\rangle,$$

thus skipping the step of computing the feature map.

Remark. Note that not every arbitrary feature map correspond to a kernel trick.

Riemannian Manifold

Laplacian Support Vector Machine

Results

Conclusion

Further Research

Bibliography

[1] Dirk Jan Struik. *Joseph-Louis Lagrange, comte de l'Empire — French mathematician — Britannica*. en. URL: https://www.britannica.com/biography/Joseph-Louis-Lagrange-comte-de-lEmpire (visited on 05/11/2022).

[2] H. W. Kuhn and A. W. Tucker. "Nonlinear programming". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*. Berkeley and Los Angeles: University of California Press, 1951, pp. 481–492.

[3] William Trench. "The method of Lagrange multipliers". Trinity University. URL: http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_LAGRANGE_METHOD.PDF.

[4] Amir Beck. *Introduction to Nonlinear Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2014. DOI: 10.1137/1.9781611973655. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611973655. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611973655.

[5] Vladimir N. Vapnik. *An overview of statistical learning theory*. 1999. DOI: 10.1109/72.788640.

[6] Vivek Salunkhe. *Support Vector Machine (SVM)*. 2021. URL: https://medium.com/@viveksalunkhe80/support-vector-machine-svm-88f360ff5f38.

[7] Jocelyn Quaintance Jean Gallier. *Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Machine Learning*. 2022. URL: https://www.cis.upenn.edu/~jean/math-deep.pdf.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[9] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. 1st ed. Cambridge University Press, 2000. ISBN: 0521780195. URL: http://www.amazon.com/Introduction-Support-Machines-Kernel-based-Learning/dp/0521780195/ref=sr_1_1?ie=UTF8&s=books&qid=1280243230&sr=8-1.

[10] Chih-Chung Chang and Chih-Jen Lin Chih-Wei Hsu. "A Practical Guide to Support Vector Classification". In: *BJU international* 101 (1 2008). ISSN: 1464-410X.

[11] *Normed Vector Space of Rational Numbers is not Banach Space.* URL: https://proofwiki.org/wiki/Normed_Vector_Space_of_Rational_Numbers_is_not_Banach_Space.

[12] N. Aronszajn. "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68 (3 1950). ISSN: 00029947. DOI: 10.2307/1990404.

Appendix