Support Vector Machine: Classification

S. Sumitra
Department of Mathematics
Indian Institute of Space Science and Technology

Advanced Machine Learning

Support Vector Machines

- Vapnik and A. Chervonenkis in 1962 at AT& T Bell Laboratories
- Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik in 1992
- Corinna Cortes and Vapnik in 1993

Hypothesis Function

- $\tilde{f}(x_i) = f(x_i) + b$
- f lies in Reproducing Kernel Hilbert Space $\mathcal{F}$, $b \in \mathbb{R}$

Optimisation

$$\min_{f \in \mathcal{F}} \quad \frac{1}{2}\|f\|^2$$

$$\text{subject to} \frac{1}{N}\sum_{i=1}^{N} V(y_i, \tilde{f}(x_i)) \leq k$$

Classification

Data $\{(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)\}$, $x_i \in \mathbb{R}^n$, $y_i \in \{1, -1\}$ be the given data.
The SVM uses regularized cost function

$$J(\tilde{f}, b) = C \sum_{i=1}^{N} V(y_i, \tilde{f}(x_i)) + \frac{1}{2}||f||^2 \tag{1}$$

where $C > 0$ is the regularization parameter.

Prediction with Confidence

- $\tilde{f}(x) \geq 1, y = 1$, Correct Prediction, high confidence
- $\tilde{f}(x) < -1, y = -1$, Correct Prediction, high confidence
- $0 < \tilde{f}(x) < 1, y = 1$, Correct Prediction, low confidence
- $-1 < \tilde{f}(x) < 0, y = -1$, Correct Prediction, low confidence

SVM Loss Function: Hinge Loss Function

$\{x_i, y_i\}_{i=1}^N, x_i \in \mathbb{R}^n, y_i = \{-1, 1\}$ be the data points

- For correct prediction with high confidence
  - $\tilde{f}(x_i) \geq 1, y_i = 1$
  - $\tilde{f}(x_i) \leq -1, y_i = -1$
  - $y_i \tilde{f}(x_i) \geq 1$
- Hinge Loss Function $V(y, \tilde{f}(x)) = \max(0, 1 - y\tilde{f}(x))$.
- In this case, $V = 0$ for correct prediction with high confidence. Else, $V > 0$.

SVM: Signum Function

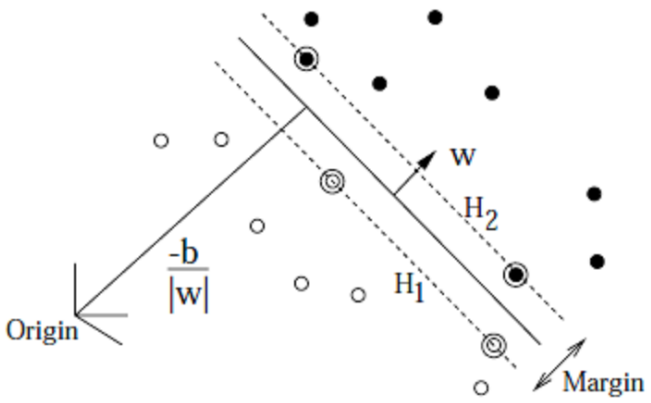
- Signum function is used to determine the class

$\tilde{f}(x) \geq 0$, x is assigned to positive class $\tilde{f}(x) < 0$, x is assigned to negative class

SVM Development

- Linear and loss equal to 0
- Linear and loss not equal to 0
- Nonlinear

Case 1: Hyperplane and $V \equiv 0$



- $\tilde{f}(x) = f(x) + b = \langle w, x \rangle + b$
- $V(y, \tilde{f}(x)) = \max(0, 1 - y\tilde{f}(x))$.
- $V \equiv 0$

Case 1: Problem Formulation

$$\begin{aligned} \underset{f \in \mathcal{F}}{\text{minimize}} \quad & ||f||^2 \\ \text{subject to} \quad & 1 - y_i \tilde{f}(x_i) \leq 0, \ i = 1, \ldots, N. \end{aligned}$$

As $f(x) = \langle w, x \rangle$, $||f|| = ||w||$, the above optimization problem can be formulated as

$$\begin{aligned} &\underset{w \in \mathbb{R}^n}{\text{minimize}} && ||w||^2 \\ &\text{subject to} && y_i(\langle w, x_i \rangle + b - 1) \geq 1, \ i = 1, \ldots, N. \end{aligned}$$

It can be proved that in this case finding a function which is smooth (minimum $||w||$) is equivalent to finding a hyperplane with largest margin.

Geometrical Interpretation of Cost Function

Let the shortest distance from the separating hyperplane $w^T x + b = 0$ to the positive class be d^+, while to the negative class be d^-.
The *margin* of a separating hyperplane is defined to be $\gamma = d^+ + d^-$.

[*The distance of a point* (x', y') *to the plane* $ax + by = 0$ *is given by* $\frac{|ax' + by'|}{\sqrt{a^2 + b^2}}$]

The distance of a point $x_i = (x_{i1}, x_{i2}, \ldots x_{in}\}$ to the hyperplane $w^T x + b = 0$ is given by

$$\begin{aligned} dist(x_i, \tilde{f}(x)) &= \frac{|w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in} + b|}{\sqrt{w_1^2 + w_2^2 + \ldots w_n^2}} \\ &= \frac{|w^T x_i + b|}{||w||} \end{aligned}$$

$w^T x + b \geq 1$ for all x in the positive class. Therefore $dist(x, \tilde{f}(x)) \geq \frac{1}{||w||}$, for all x in the positive class. The points in the positive class, that lie closest to the separating hyperplane $w^T x + b = 0$ lies in the hyperplane

$$H_1 : w^T x + b = 1$$

Therefore $d+ = \frac{1}{||w||}$.

In the negative class, for all points $w^T x + b \leq -1$. Therefore $dist(x, \tilde{f}(x)) \geq \frac{1}{||w||}$ for al x in the negative class. The points in the negative class, that lie closest to the separating hyperplane $w^T x + b = 0$ lies in the hyperplane

$$H_2 : w^T x + b = -1$$

Therefore $d^- = \frac{1}{||w||}$.

Hence margin $\gamma = d^+ + d^- = \frac{2}{||w||}$

There will be more than one hyperplane that separates the given data points. Among all such hyperplanes, the support vector machine chooses that one, which has the largest margin. That is *for the linearly separable case, the support vector algorithm chooses for the separating hyperplane with largest margin.* That is to $\max \frac{2}{||w||}$, which is equivalent to $\min \frac{||w||}{2}$, which is equivalent to $\min \frac{||w||^2}{2}$.
Now

$$w^T x_i + b \geq +1 \text{ for } y_i = +1 \tag{2}$$

$$w^T x_i + b \leq -1 \text{ for } y_i = -1 \tag{3}$$

The above equations can be can be combined into one set of inequalities:

$$y_i(w^T x_i + b) \geq 1 \ \forall \ i \tag{4}$$

Here the hyperplanes H_1 and H_2 are parallel and that no training points fall between them. Thus we can find the pair of hyperplanes which gives the maximum margin by minimizing $||w||^2$, subject to constraints (4).

Optimisation Problem

The general optimisation problem can be stated as follows:

Definition

(Primal optimisation problem) Given functions f, g_i, $i = 1, 2 \ldots, k$, and $hi, i = 1, 2, \ldots, m$, defined on a domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} \textit{minimise} \quad & f(w), w \in \Omega, \\ \textit{subject to} \quad & g_i(w) \leq 0, i = 1, \ldots k, \\ & h_i(w) = 0, i = 1, \ldots, m \end{aligned}$$

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 optimisation problem.

Maximisation problems can be converted to minimisation ones by changing the the sign of $f(w)$.

Feasible Region

[Refer: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods by Nello Cristianini and John Shawe-Taylor]

Definition

(feasible region) The region of the domain where the objective function is defined and where all the constraints are satisfied is called the feasible region, which can be denoted by

$$S = \{w \in \Omega : g(w) \leq 0, h(w) = 0\}$$

A solution of the optimisation problem is a point $w^* \in S$ such that there exists no other point $w \in S$ for which $f(w) < f(w^*)$. Such a point is also known as a global minimum.

Active Constraint

An inequality constraint $g_i(w) \leq 0$ is said to be active (or tight) if the solution w^* satisfies $g_i(w^*) = 0$, otherwise it is said to be inactive. Hence , equality constraints are always active.

Convex Function

Definition

A real-valued function $f(w)$ is called convex for $w \in \mathbb{R}^n$ if, $\forall w, u \in \mathbb{R}^n$, and for any $\theta \in (0, 1)$,

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

If a strict inequality holds, the function is said to be strictly convex. A function that is twice differentiable will be convex provided its Hessian matrix is positive semi-definite. An affine function is one that can be expressed in the form

$$f(w) = Aw + b$$

for some matrix A and vector b. Note that affine functions are convex as they have zero Hessian.

Convex Set

Definition

A set $\Omega \subseteq \mathbb{R}^n$ is called convex if, $\forall w, u \in \Omega$, and for any $\theta \in (0, 1)$, the point $(\theta w + (1 - \theta)u) \in \Omega$.

If a function f is convex, any local minimum w^* of the unconstrained optimisation problem with objective function f is also a global minimum. To prove this take $u \neq w^*$. Then by the definition of a local minimum there exists θ sufficiently close to 1 such that

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

Therefore $f(w^*) \leq f(u)$.

Convex Quadratic Programmes

Definition

An optimisation problem in which the set Ω, the objective function and all of the constraints are convex is said to be convex.

For SVM, here we are considering the case where the constraints are linear, the objective function is convex and quadratic and $\Omega = \mathbb{R}^n$, that is, convex quadratic programmes.

Lagrangian Theory

The Lagrangian theory was developed by Lagrange in 1797 for determining the solution of an optimisation problem which there are no inequality constraints. In 1951 Kuhn and Tucker extended the theory by incorporating inequality constraints which is known as Kuhn-Tucker theory.

Definition

(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.*

The Lagrangian is defined as the objective function plus a linear combination of the constraints, where the coefficients of the combination are called the Lagrange multipliers.

Lagrangian Theory

Definition

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

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

where the coefficients β_i are called the Lagrangian multipliers.

Theorem

(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 $f, hi \in C^1$, is

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

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

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

Lagrangian Function

Now let us consider generalized Lagrangian. The definition of generalised Lagrangian function is:

Definition

Given an optimisation problem with domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} \text{minimise} \quad & f(w) \quad , w \in \Omega \\ \text{subject to} \quad & g_i(w) \quad \leq 0, i = 1, 2, \ldots k \\ & h_i(w) \quad = 0, i = 1, 2 \ldots m \end{aligned}$$

The generalized lagrangian function is defined as

$$L(w, \alpha, \beta) \quad = \quad f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w), \alpha_i \geq 0$$

[For constraints of the form $c_i \geq 0$, the constraint equations are multiplied by positive Lagrange multipliers and subtracted from the objective function, to form the Lagrangian. For equality constraints, the Lagrange multipliers are unconstrained].

Dual Problem

Definition

The Lagrangian dual problem of the above given primal problem is the following problem:

$$\begin{aligned} \text{maximize} \quad & \theta(\alpha, \beta) \\ \text{subject to} \quad & \alpha \geq 0 \end{aligned}$$

where $\theta(\alpha, \beta) = \inf_{w \in \Omega} L(w, \alpha, \beta)$. The value of the objective function at the optimal solution is called the value of the problem.

Theorem

Let $w \in \Omega$ be a feasible solution of the primal problem and (α, β) a feasible solution of the dual problem. Then $f(w) \geq \theta(\alpha, \beta)$.

$$\begin{aligned} \theta(\alpha,\beta) &= \inf_u L(u,\alpha,\beta) \\ &\leq L(w,\alpha,\beta) \\ &= f(w) + \alpha' g(w) + \beta' h(w) \\ &\leq f(w) \end{aligned}$$

- If $f(w^*) = \theta(\alpha^*, \beta^*)$, where $\alpha^* \geq 0$, and $g(w^*) \leq 0$, $h(w^*) = 0$, then w^* and (α^*, β^*) solve the primal and dual problems respectively. In this case $\alpha_i^* gi(w^*) = 0$, for $i = 1, \ldots, k$.
- The difference between the values of the primal and dual problems is known as the duality gap.
- The triple (w^*, α^*, β^*) is a saddle point of the Lagrangian function for the primal problem, if and only if its components are optimal solutions of the primal and dual problems and there is no duality gap, that is, $f(w^*) = \theta(\alpha^*, \beta^*)$.

Strong Duality Theorem

Theorem (strong duality theorem)

Given an optimisation problem with convex domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{array}{lll} \textit{minimise} & f(w) & , w \in \Omega \\ \textit{subject to} & g_i(w) & \leq 0, i = 1, 2, \ldots k \\ & h_i(w) & = 0, i = 1, 2 \ldots m \end{array}$$

with $f \in C_1$ *convex and* g_i, h_i *affine, then the duality gap is zero.*

The Karush-Kuhn-Tucker Conditions (KKT)

(Kuhn-Tucker) Given an optimisation problem with convex domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} \text{minimise} \quad & f(w) \quad , w \in \Omega \\ \text{subject to} \quad & g_i(w) \quad \leq 0, i = 1, 2, \ldots k \\ & h_i(w) \quad = 0, i = 1, 2 \ldots m \end{aligned}$$

with $f \in C_1$ convex and g_i, h_i affine, necessary and sufficient conditions for a normal point w^* to be an optimum are the existence of α^*, β^* such that

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

$$\frac{\partial(Lw^*, \alpha^*, \beta^*)}{\partial \beta} = 0$$

$$\alpha_i^* g_i(w^*) = 0, i = 1, 2, \ldots, k$$

$$g_i(w^*) \leq 0, i = 1, 2, \ldots, k$$

SVM: Case 1: Optimisation Problem

$$L(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{N} \alpha_i [y_i(\langle w, x_i\rangle + b) - 1] \tag{5}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.
The corresponding dual is found by differentiating with respect to w and b, imposing stationarity,

$$\frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{N} y_i \alpha_i x_i = 0 \tag{6}$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0 \tag{7}$$

From (6)

$$w = \sum_{i=1}^{N} y_i \alpha_i x_i \tag{8}$$

$$f(x) = \langle w, x \rangle = \sum_{i=1}^{N} y_i \alpha_i \langle x_i, x \rangle$$

From the above equation it is clear that the hypothesis can be written as a linear combination of the training points [which also follows from the representer theorem].
From (7)

$$0 = \sum_{i=1}^{N} y_i \alpha_i \tag{9}$$

KKT Complimentary Condition

Using Karush-Kuhn-Tucker complementarity conditions,

$$\alpha_i^*[y_i(\langle w, x_i\rangle + b) - 1] = 0, i = 1, 2, \ldots N \tag{10}$$

Therefore two cases

- $\alpha_i^* = 0$
  - $y_i(\langle w, x_i\rangle + b) - 1 \geq 0$
  - x_i lies either above H_1 or on H_1 or above H_2 or on H_2
- $\alpha_i^* \neq 0$
  - $y_i(\langle w, x_i\rangle + b) - 1 = 0$
  - x_i lies either on H_1 or H_2

Support Vectors

- The *xi′s* corresponding to nonzero α_i's are called support-vectors. Thus the only inputs that appear in the expression of *w*, and hence in the expression of hyperplane are the support vectors.
- Support vectors lie closest to the hyperplane.
- In the expression for the weight vector only support vectors are involved. It is for this reason that they are called support vectors.

Dual Formulation

Sub (8) and (9) into (5) to form the dual

$$\begin{aligned} \|w\|^2 &= \langle w, w \rangle \\ &= \langle \sum_i \alpha_i y_i x_i, \sum_i \alpha_i y_i x_i \rangle \\ &= \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{N} \alpha_i [y_i (\langle w, x_i \rangle] &= \sum_{i=1}^{N} \alpha_i [y_i (\langle \sum_{j=1}^{N} y_j \alpha_j x_j, x_i \rangle] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \end{aligned}$$

Collecting the terms of b,

$$b \sum_i \alpha_i y_i = 0$$

Dual Formulation

Sub (8) and (9) into (5) to form the dual,

$$
\begin{aligned}
W(\alpha) &= \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{N} \alpha_i \\
&= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle
\end{aligned}
$$

Hence the dual optimisation problem is

$$\textit{maximise } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\text{subject to} \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$\alpha_i \geq 0, i = 1, 2, \ldots N$$

$L(w, \alpha, b)$ and $W(\alpha)$ arise from the same objective function but with different constraints and the solution is found by minimizing $L(w, \alpha, b)$ or by maximizing $W(\alpha)$.

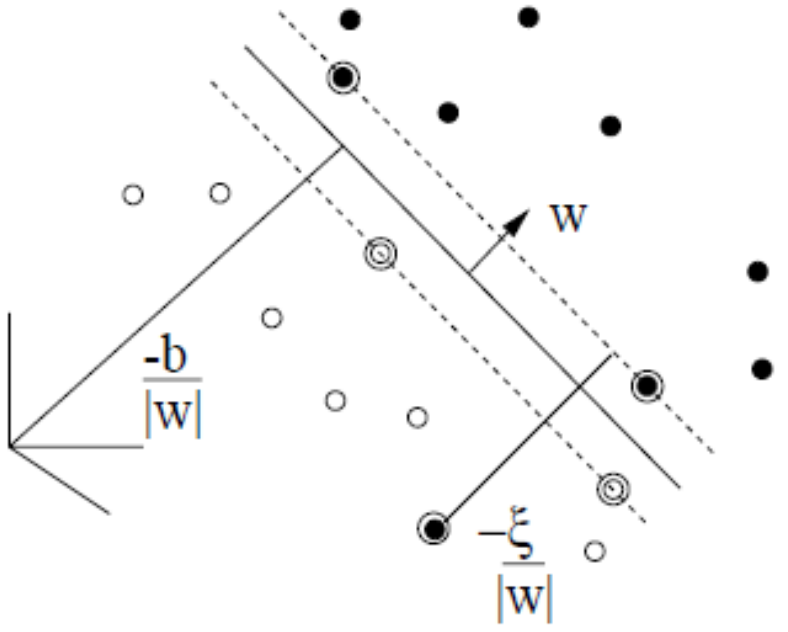
Bias

To find b, choose x_i for which $\alpha_i \neq 0$ and use (10) [take the average of $b's$ over all such x_i's].

Test Phase

- The class of the testing point x is determined by $sgn(w^T x + b)$, that is,
  - x belongs to the positive class if $sgn(w^T x + b) \geq 0$
  - Else to the negative class

Case 2: Hyperplane and $V \not\equiv 0$



$$\text{minimise}_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{||w||^2}{2} + C \sum_{i=1}^{N} max(0, 1 - y_i(w^T x_i + b))$$

$$V(y, \tilde{f}(x_i)) = \max(0, 1 - y_i \tilde{f}(x_i)) = \xi_i, i = 1, 2, \ldots N$$

Optimization

$$\text{minimise}_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{||w||^2}{2} + C \sum_{i=1}^{N} \xi_i$$

where $C > 0$ is the regularization parameter, subject to the constraints

$$y_i(w^T x_i + b) \geq +1 - \xi_i \text{ for } i = 1, 2, \ldots N \tag{11}$$

$$\xi_i \geq 0 \; i = 1, 2, \ldots N \tag{12}$$

Lagrangian Formulation

$$L(w, b, \xi, \alpha) = \frac{||w||^2}{2} + C\sum_{i=1}^{N} \xi_i - \sum_i \alpha_i [y_i(w^T x_i + b) - 1 + \xi_i] - \sum_i \mu_i \xi_i \tag{13}$$

Applying the KKT conditions,
For $i = 1, 2, \ldots N$

$$\frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{N} y_i \alpha_i x_i = 0 \tag{14}$$

Therefore,

$$w = \sum_{i=1}^{N} y_i \alpha_i x_i$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0 \tag{15}$$

For i=1,2,. . . , N,

$$\frac{\partial L(w, b, \alpha)}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{16}$$

$$\alpha_i \geq 0 \tag{17}$$

$$\mu_i \geq 0 \tag{18}$$

KKT Complimentary Conditions

The KKT complimentary conditions are,

$$\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] = 0 \tag{19}$$

$$\mu_i \xi_i = 0 \tag{20}$$

From (16) and(18)

$$C - \alpha_i = \mu_i \geq 0 \tag{21}$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \ldots N \tag{22}$$

Three cases

- $\alpha_i = 0$
- $0 < \alpha_i < C$
- $\alpha_i = C$

- When $\alpha_i = 0$, from (21), $\mu_i = C > 0$, therefore from (20), $\xi_i = 0$. Hence from (19), $y_i(\langle w, x_i \rangle + b) \geq 1$.
- If $0 < \alpha_i < C$, from (21), $\mu_i > 0$. Hence from (20), $\xi_i = 0$. Therefore from (19), $y_i(\langle w, x_i \rangle + b) = 1$. Hence, points for which $0 < \alpha_i < C$ lie at the target distance of $\frac{1}{||w||}$ from the hyperplane.
- If $\alpha_i = C$, from (21), $\mu_i = 0$. Hence from (20), $\xi_i \geq 0$. Hence (19) , $y_i(\langle w, x_i \rangle + b) = 1 - \xi_i$. That is, $y_i(\langle w, x_i \rangle + b) \leq 1$.

Support Vectors

Support vectors are those points for which $y_i(\langle w, x_i \rangle + b) \leq 1$.

- $y_i(\langle w, x_i \rangle + b) = 1$, x_i on H_1 or on H_2
- $0 \leq y_i(\langle w, x_i \rangle + b) < 1$ Between H_1 and H or H_2 and H or on H
- $y_i(\langle w, x_i \rangle + b) < 0$ Incorrect classification

Dual Formulation

- The terms that involve w : $\frac{1}{2}\|w\|^2 - \sum_i \alpha_i [y_i(w^T x_i]$
- The terms that involve b : $\sum_i \alpha_i y_i b$
- The terms that involve ξ_i : $\sum_{i=1}^{N} \xi_i (C - \alpha_i - \mu_i)$
- The remaining terms are: $\sum_i \alpha_i$

Dual form is

$$\textit{maximise } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\text{subject to} \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$C \geq \alpha_i \geq 0, i = 1, 2, \ldots N.$$

General Form

$$V(y_i, \tilde{f}(x_i)) = \max(0, 1 - y_i(\tilde{f}(x_i)) = \xi_i$$

$$\tilde{f}(x) = f(x) + b$$

$$\begin{aligned} &\min_{f \in \mathcal{F}, b \in \mathbb{R}} \frac{1}{2}||f||^2 + C\sum_{i=1}^{N} \xi_i \\ &y_i(f(x_i) + b) \geq 1 - \xi_i, \ i = 1, \ldots, N \\ &\xi_i \geq 0. \end{aligned} \tag{23}$$

Hyperplane Models: RKHS

If $\tilde{f}$ is the unknown function of a datamodeling problem, then by kernel theory

$$\tilde{f}(x) = f(x) + b = \langle f, k_x \rangle + b, \forall x \in \mathcal{X};\ f, k_x \in \mathcal{F}, b \in \mathbb{R} \quad (24)$$

Consider $\tilde{H}_{f,b} : \mathcal{F} \to \mathbb{R}$ where

$$\tilde{H}_{f,b}(g) = \langle f, g \rangle + b, \forall g \in \mathcal{F} \quad (25)$$

(25) is the equation of a hyperplane in RKHS.
Comparing (24) and (25),

$$\tilde{f}(x) = \tilde{H}_{f,b}(k_x), x \in \mathcal{X}, k_x \in \mathcal{F}$$

Thus finding $\tilde{f}$ in input space is equivalent in finding $\tilde{H}_{f,b}$ in RKHS.

Optimisation Problem

$$\min_{f\in\mathcal{F},b\in\mathbb{R}} \frac{1}{2}||f||^2 + C\sum_{i=1}^{N}\xi_i$$

subject to (26)

$$y_i(\langle f, k_{x_i}\rangle + b) - 1 + \xi_i \geq 0,\ i = 1,\ldots,N$$

$$\xi_i \geq 0.$$

- The objective function is quadratic and all the constraints affine.

Lagrangian Formulation

The primal Lagrangian is,

$$L(f, b, \xi, \alpha) = \frac{||f||^2}{2} + C\sum_{i=1}^{N}\xi_i - \sum_i \alpha_i[y_i(\langle f, k_{x_i}\rangle + b) - 1 + \xi_i] - \sum_i \mu_i\xi_i \tag{27}$$

KKT conditions

$$\frac{\partial L}{\partial f} = f - \sum_{i=1}^{N} \alpha_i y_i k_{x_i} = 0$$

$$f = \sum_{i=1}^{N} \alpha_i y_i k_{x_i} \tag{28}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{29}$$

For i=1,2,. . . , N,

$$\frac{\partial L(w, b, \alpha)}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{30}$$

$$\alpha_i \geq 0 \tag{31}$$

$$\mu_i \geq 0 \tag{32}$$

KKT Complimentary Conditions

The KKT complimentary conditions are,

$$\alpha_i[y_i(\langle f, k_{x_i}\rangle + b) - 1 + \xi_i] = 0 \tag{33}$$

$$\mu_i \xi_i = 0 \tag{34}$$

From (30) and(32)

$$C - \alpha_i = \mu_i \geq 0 \tag{35}$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \ldots N \tag{36}$$

Three cases

- $\alpha_i = 0$
- $0 < \alpha_i < C$
- $\alpha_i = C$

- When $\alpha_i = 0$, from (35), $\mu_i = C > 0$, therefore from (34), $\xi_i = 0$. Hence from (33), $y_i(\langle f, k_{x_i} \rangle + b) \geq 1$.
- If $0 < \alpha_i < C$, from (35), $\mu_i > 0$. Hence from (34), $\xi_i = 0$. Therefore from (33), $y_i(\langle f, k_{x_i} \rangle + b) = 1$. Hence, points for which $0 < \alpha_i < C$ lie at the target distance of $\frac{1}{||f||}$ from the hyperplane.
- If $\alpha_i = C$, from (35), $\mu_i = 0$. Hence from (34), $\xi_i \geq 0$. Hence (33) , $y_i(\langle f, k_{x_i} \rangle + b) = 1 - \xi_i$. That is, $y_i(\langle f, k_{x_i} \rangle + b) \leq 1$.

Support Vectors

Support vectors are those points for which $y_i(\langle f, k_{x_i} \rangle + b) \leq 1$.

- $y_i(\langle f, k_{x_i} \rangle + b) = 1$, k_{x_i} on H_1 or on H_2
- $0 \leq y_i(\langle f, k_{x_i} \rangle + b) < 1$ Between H_1 and H or H_2 and H or on H
- $y_i(\langle f, k_{x_i} \rangle + b) < 0$ Incorrect classification

Dual Formulation

- The terms that involve f : $\frac{1}{2}\|f\|^2 - \sum_i \alpha_i [y_i(\langle f, k_{x_i} \rangle]$
- The terms that involve b : $\sum_i \alpha_i y_i b$
- The terms that involve ξ_i : $\sum_{i=1}^{N} \xi_i (C - \alpha_i - \mu_i)$
- The remaining terms are: $\sum_i \alpha_i$

$$\|f\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$\|f\|^2 = \beta^T K \beta$$

where K is the kernel matrix and $\beta = (\alpha_1 y_1, \alpha_2 y_2, \ldots \alpha_N y_N)^T$

- $\|f\|^2 \geq 0$
  - k should be positive semi-definite
  - K should be positive semi-definite

Dual Optimisation

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

$$\text{subject to} \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$C \geq \alpha_i \geq 0, i = 1, 2, \ldots N.$$

.

SVM Model

$$f = \sum_{i=1}^{sv} y_i \alpha_i^* k_{x_i} \tag{37}$$

where sv is the number of support vectors. Then

$$\begin{aligned} \tilde{f}(x) &= \langle f, k_x \rangle + b \\ &= \sum_{i=1}^{sv} y_i \alpha_i^* \langle k_{x_i}, k_x \rangle + b \\ &= \sum_{i=1}^{sv} y_i \alpha_i^* k(x_i, x) + b \end{aligned}$$

Characteristics of SVM Model

- Sparse algorithm
- In SVM classification, the optimal hyperplane is found in RKHS by minimizing the training error and maximizing the margin.

RKHS and Reproducing Kernel

- There exists a bijection between the set of all reproducing kernel Hilbert spaces and the set of all positive kernel functions.
- Regularization attempts to provide well-posed solutions to a learning task, specifically ERM, by constraining the capacity of the hypothesis space through the elimination of complex functions that are unlikely to generalize, thereby isolating a unique and stable solution.

Hyperparameters

Training a model requires the choice of few relevant quantities:

- The kernel function, that determines the shape of the decision surface;
- Parameter in the kernel function (eg: for Gaussian kernel:variance of the Gaussian, for polynomial kernel: degree of the polynomial)
- The regularization parameter

Non Parametric Algorithm

In kernel methods, the number of basis functions depends on the number of data and hence should not be fixed a priori. Hence they are non parametric methods.