Support Vector Machine

Chunyan Li
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA
Email: chunyan@email.sc.edu.

April 30, 2022

1 Introduction

In this note, we will focus on the support vector machine and explore the KKT conditions and duality.

2 Unconstrained optimization problem

3 Constrained optimization problem

For constrained optimization problem, one way to solve it is convert it into a unconstrained one and then all the tools of unconstrained optimization problem could be used. Otherwise, one can convert it into another constrained problem but with much easier constrains which makes the problem is easier to solve.

Assume f, g_i, h_j are continuous and differentiable in Ω. A general constrained optimization problem could be formulated as follows:

$$min_{\mathbf{x}\in\Omega}\ f(\mathbf{x}) \tag{3.1}$$

Where $\Omega = \{\mathbf{x} \in \mathbf{R}^n | g_i(\mathbf{x}) = 0, i = 1..., m, h_j(\mathbf{x}) \leq 0, j = 1..., l\}$.

We could also write it as the following form:

$$\begin{aligned} min_{\mathbf{x}\in\mathbf{R}^n}\ &f(\mathbf{x}) \\ s.t.\ &\mathbf{g}(\mathbf{x}) = 0 \\ &\mathbf{h}(\mathbf{x}) \leq 0 \end{aligned} \tag{3.2}$$

where $\mathbf{g} = (g_1, ..., g_m)$ and $\mathbf{h} = (h_1, ..., h_l)$.

We could extend f into the whole domain by add the constrains through Lagrangian multipliers. First, we write down the Lagrangian function

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{l} \mu_j h_j(\mathbf{x}) \tag{3.3}$$

where $\mu \geq 0$ elementwise and $\boldsymbol{\alpha}, \boldsymbol{\mu}$ are Lagrangian multipliers or dual variables. Then, one can show that the original constrained optimization problem is equivalent to the unconstrained optimization problem on the extend f.

$$f_{extend} = max_{\boldsymbol{\alpha}} max_{\boldsymbol{\mu}\geq 0}\ L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{3.4}$$

Note

$$f_{extend} = \begin{cases} f(\mathbf{x}), & \mathbf{x} \in \Omega \\ +\infty, & \mathbf{x} \in \Omega^c \end{cases} \tag{3.5}$$

If $\mathbf{x} \in \Omega$, then, $max_{\boldsymbol{\alpha}} max_{\boldsymbol{\mu}\geq 0}\ L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = f(\mathbf{x})$ since $\mathbf{g} = 0$ and $\boldsymbol{\mu}\mathbf{h}(\mathbf{x}) \leq 0$. Otherwise, there exists an $\mathbf{x}$ such that $g_i(\mathbf{x}) \neq 0$ or $h_j(\mathbf{x}) > 0$. Then, L will approach $+\infty$ as corresponding $\boldsymbol{\alpha}$ or $\boldsymbol{\mu}$ goes to ∞. After the discussion above, obsequiously, we know

$$min_{\mathbf{x}\in\Omega}\ f(\mathbf{x}) = min_{\mathbf{x}\in\mathbf{R}^n}\ f_{extend} = min_{\mathbf{x}\in\mathbf{R}^n}\ max_{\boldsymbol{\alpha}} max_{\boldsymbol{\mu}\geq 0}\ L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{3.6}$$

Hence, the key is to extend the function beyond Ω using a hard wall potential so that the extended function is infinite beyond Ω. Now, we are ready to move further for solving this unconstrained optimization problem.

4 Primal and Dual problems

Problem (3.2) is referred to as primal problem and there is the corresponding dual problem which is searching for the best lower bound of the primal optimal denoted $p*$. The dual problem is obtained through the Lagrangian (3.3) and the constrains on dual variables. Note that $\mathbf{x}$ is the primal variable, $\boldsymbol{\alpha}, \boldsymbol{\mu}$ are the dual variables.

Dual function is defined as the infimum of Lagrangian function over primal variable $\mathbf{x} \in \mathbf{R}^n$:

$$G(\boldsymbol{\alpha}, \boldsymbol{\mu}) = min_{\mathbf{x} \in \mathbf{R}^n}\ L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = L(\mathbf{x}^*, \boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{4.1}$$

Note that, since G is defined as a point-wise minimum, it is a concave function. We could solve this unconstrained optimization problem by finding the stationary point. Clearly, $\mathbf{x}^*$ satisfies

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{m} \alpha_i \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) + \sum_{j=1}^{l} \mu_j \nabla_{\mathbf{x}} h_j(\mathbf{x}^*) = 0 \tag{4.2}$$

Hence, one can write $x^* = x(\boldsymbol{\alpha}, \boldsymbol{\mu})$.

One can easily find the relationship of the objective value and the dual function value, denote the primal optimal p^*, then, we have

$$\forall \mathbf{x} \text{ feasible},\ f(\mathbf{x}) \geq L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) \geq G(\boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{4.3}$$

That is, the dual function provides a lower bound on the objective value $p*$ in the feasible set. The right-hand side of the above inequality is independent of $\mathbf{x}$. Taking the minimum over $\mathbf{x}$ in the above, we obtain

$$p^* \geq G(\boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{4.4}$$

Since this lower bound is valid for every dual variable $\boldsymbol{\mu} \geq 0, \boldsymbol{\alpha}$, We can search for the best one, that is the largest lower bound d^*:

$$p^* \geq d^* := max_{\boldsymbol{\alpha}} max_{\boldsymbol{\mu} \geq 0}\ G(\boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{4.5}$$

we call this weak duality.

Hence, the Lagrangian dual problem is defined as

$$\begin{aligned} &max_{\boldsymbol{\alpha}} max_{\boldsymbol{\mu}}\ G(\boldsymbol{\alpha}, \boldsymbol{\mu}) \\ &s.t.\ \nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = 0 \\ &\boldsymbol{\mu} \geq 0 \end{aligned} \tag{4.6}$$

Note that the constrain on dual variable of inequality is natural, and the other constrain is determined by solving the unconstrained optimization problem (4.1) whose optima are characterized by setting the derivative with respect to primal variable $\mathbf{x}$ to zero.

Now, it is time to consider when will the primal problem and dual problem are attained. In the next section, we will explore the necessary conditions for primal and dual optimum attainment and the procedure of how to determine the optimal triples $(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\mu}^*)$.

4.1 Strong duality and KKT conditions

Strong duality: The theory of weak duality seen above states that $p^* \geq d^*$. This is true always, even if the original problem is not convex. We say that strong duality holds if $p^* = d^*$.

Slater's sufficient condition for strong duality: Slater's theorem provides a sufficient condition for strong duality to hold. Namely, if

1. The primal problem is convex;
2. It is strictly feasible, that is, there exists $\mathbf{x}_0 \in \mathbf{R}^n$ such that $\mathbf{g}(\mathbf{x}_0) = 0, h_i(\mathbf{x}_0) < 0,\ \forall i = 1, ..., m$

then, strong duality holds: $p^* = d^*$, and the dual problem is attained.

Sufficient condition for dual optimum attainment: Slater condition, namely strict feasibility of the primal, ensures that the dual problem is attained.

Primal optimum attainment: Likewise, if in addition the dual problem is strictly feasible, that is if:

$$\exists \boldsymbol{\mu} > 0,\ \boldsymbol{\alpha} \in R^m\ s.t.\ G(\boldsymbol{\alpha}, \boldsymbol{\mu}) > -\infty,$$

then strong duality holds, and both problems are attained, that is: there exist $(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ such that $\mathbf{x}$ is feasible for the primal problem; $\boldsymbol{\alpha}, \boldsymbol{\mu}$ are feasible for the dual problem: $\boldsymbol{\mu} \geq 0$, and $(\boldsymbol{\alpha}, \boldsymbol{\mu}) \in \mathbf{dom}G$.

Optimality conditions: The following conditions are called the Karush-Kuhn-Tucker (KKT) conditions

1. Primal feasibility: $\mathbf{g}(\mathbf{x}) = 0$ ($\nabla_{\boldsymbol{\alpha}} L = 0$), and $\mathbf{h}(\mathbf{x}) \leq 0$
2. Dual feasibility: $\boldsymbol{\mu} \geq 0$
3. Lagrangian stationarity: (in the case when every function involved is differentiable) $\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) + \sum_{j=1}^{l} \mu_j \nabla_{\mathbf{x}} h_i(\mathbf{x}) = 0$
4. Complementary slackness:
$$\boldsymbol{\mu}_j h_j(\mathbf{x}) = 0, \quad j = 1, ..., l. \tag{4.7}$$

Note the complementary slackness is derived from the strong duality

$$f(\mathbf{x}^*) = G(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*) \leq f(\mathbf{x}^*) + \sum_{i=1}^{m} \alpha_i^* g_i(\mathbf{x}^*) + \sum_{j=1}^{l} \mu_j^* h_j(\mathbf{x}^*) \leq f(\mathbf{x}^*) \tag{4.8}$$

The first equality is the strong duality, the first inequality follows from the definition of G and the second inequality follows from the primal feasibility and dual feasibility. Hence, the complementary slackness follows.

If the problem is convex, and satisfies Slater's condition, then a primal point is optimal if and only if there exist $(\boldsymbol{\alpha}, \boldsymbol{\mu})$ such that the KKT conditions are satisfied. Conversely, the above conditions guarantee that strong duality holds, and $(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ are optimal. In a summary, under the Slater's condition, KKT conditions are the necessary and sufficient conditions for $(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ to be optimal.

4.2 How to find the primal optimal

Now, the question goes back to searching for the primal optimal. One can solve the dual problem first to get $(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*)$ if the dual problem is easier to solve. Then, The procedure is:

1. write down the Lagrangian function according to the primal problem after rewrite it into the standard form,
2. define the dual function as the minimum of Lagrangian over primal variable,
3. determine the constrains of dual problem using the $2^{nd}, 3^{rd}$ KKT conditions since they are the constrains on dual variables. Usually, one can simplify the dual function using the 3^{rd} condition (after this step one usually end up with a convex optimization/quadratic programming problem),
4. solve for $(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*)$ using the convex optimization/QP algorithm,
5. determine $\mathbf{x}^*$ using the 4^{th} condition (complementary slackness).

5 SVM with Hard margin

5.1 Formulate the model as an optimization problem/primal problem

As show in the Figure.5.1, they are two groups, w.l.o.g. we label the green ones by -1, and the blue ones by 1, and assume that these two groups are completely linear spreadable. then, there is a border-line ('support vectors') for each group, $y = \mathbf{w}^T\mathbf{x} + b = -1$ and $y = \mathbf{w}^T\mathbf{x} + b = 1$. All hyperplanes lie in these two planes could separate these two groups. So how to determine the optimal one? The goal is to maximize the margin/gap between these two border-line and choose the hyperplane that has the same distance to two groups. The points lie on the margins are called support vectors since the classifier/hyper-plane is uniquely determine by these support vectors. The classifier is define as: $f(x) = sign(\mathbf{w}^T\mathbf{x} + b)$

The distance between margins is $d = \frac{|-1-b-(1-b)|}{\sqrt{(\mathbf{w}^T\mathbf{w})}} = \frac{2}{||\mathbf{w}||_2}$. The green points lie below or on the margin $y = \mathbf{w}^T\mathbf{x} + b = -1$ satisfy $y(\mathbf{w}^T\mathbf{x} + b) \geq 1$. The blue points lie above or on the margin $y = \mathbf{w}^T\mathbf{x} + b = 1$ satisfy $y(\mathbf{w}^T\mathbf{x} + b) \geq 1$. Then, We could formulate the problem into the constrained optimization problem as following

$$\begin{aligned} &min_{\mathbf{w},b} \ \frac{1}{2}||\mathbf{w}||^2 \\ &s.t. \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1..., l \end{aligned} \tag{5.1}$$

Now, we solving this problem by primal dual arguments and KKT conditions. first, we write down the Lagrangian function

$$L(\mathbf{w}, b, \boldsymbol{\mu}) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{l} \mu_i[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)] \tag{5.2}$$

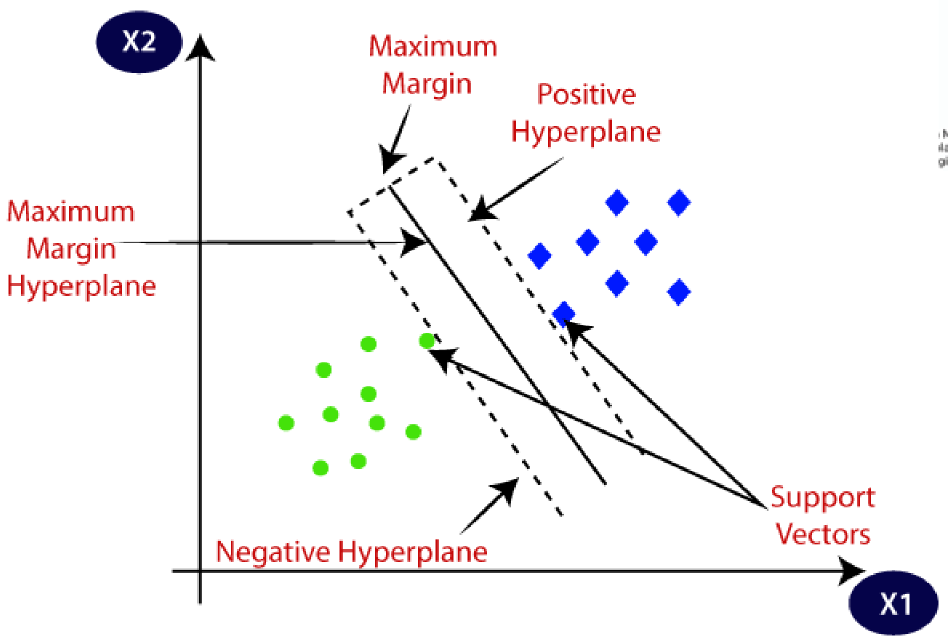

Figure 5.1: schematic of svm download from online source

Then, the Lagrangian dual function is

$$G(\boldsymbol{\mu}) = min_{\mathbf{w}\in\mathbf{R}^n, b\in\mathbf{b}}\ L(\mathbf{w}, b, \boldsymbol{\mu}) = L(\mathbf{w}^*, b^*, \boldsymbol{\mu}) \tag{5.3}$$

where $\mathbf{w}^*, b^*$ satisfies

$$\nabla_{\mathbf{w}} L(\mathbf{w}^*, b, \boldsymbol{\mu}) = \mathbf{w}^* - \sum_{i=1}^{l} \mu_i y_i \mathbf{x}_i = 0 \tag{5.4}$$

$$\nabla_b L(\mathbf{w}^*, b, \boldsymbol{\mu}) = \sum_{i=1}^{l} \mu_i y_i = 0 \tag{5.5}$$

hence,

$$\mathbf{w}^* = \sum_{i=1}^{l} \mu_i y_i \mathbf{x}_i \tag{5.6}$$

The Lagrangian dual problem is

$$\begin{aligned} max_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) \\ s.t.\ \nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\mu}) = 0 \\ \nabla_b L(\mathbf{w}, b, \boldsymbol{\mu}) = 0 \\ \boldsymbol{\mu} \geq 0 \end{aligned} \tag{5.7}$$

We plug (5.5) and (5.6)into it, we could simplify the dual problem as

$$\begin{aligned} & max_{\boldsymbol{\mu}} \sum_{i=1}^{l} \mu_i - \frac{1}{2} \sum_{j=1}^{l} \sum_{i=1}^{l} \mu_j y_j (\mathbf{x}_j^T \mathbf{x}_i) \mu_i y_i \\ & \textbf{s.t.}\ \ \boldsymbol{\mu} \geq 0 \\ & \qquad \sum_{j=1}^{l} \mu_j y_j = 0 \end{aligned} \tag{5.8}$$

The objective function of dual problem is a quadratic function of dual variable $\boldsymbol{\mu}$, hence, this is a quadratic programming problem of $\boldsymbol{\mu}$ which can be solved using well-developed techniques.

5.2 Determine $\mathbf{w}^*, b^*$

As the above stated, one could solve for $\boldsymbol{\mu}^*$ by QP programming. Then, how to get the solution/optimal of primal problem? i.e. how to determine $\mathbf{w}^*, b^*$? One can determine $\mathbf{w}^*, b^*$ by the 3rd and 4th KKT conditions which are the relations between primal variables and dual variables.

Find $\mathbf{w}^*$ by $\nabla_{\mathbf{w}} L = 0$ and complementary slackness condition:

$$\mathbf{w}^* = \sum_{i=1,\mu_i>0} \mu_i y_i \mathbf{x}_i \tag{5.9}$$

such x_i are called support vectors, where $\mu_i > 0$, the constrains are activated.

Find b^* by $\nabla_b L = 0$ and complementary slackness condition:

$$\mu_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1] = 0 \tag{5.10}$$

by complementary slackness condition, we know $\mu_{i*} > 0$ for some $i*$, then, for every $i*$, the following equation is true

$$b = \frac{1}{y_{i*}} - \mathbf{w}^T\mathbf{x}_{i*} = y_{i*} - \mathbf{w}^T\mathbf{x}_{i*} \tag{5.11}$$

here we use the fact that $\frac{1}{y} = y$ since $y = \pm 1$ hence, in practice, we could determine b^* by the average over $i*$.

$$b^* = mean(y_{i*} - \mathbf{w}^T\mathbf{x}_{i*}) \tag{5.12}$$

Then, one can predict the class of a new point $\mathbf{x}$ by

$$sign\left(\sum_{i,\mu_i>0} \mu_i y_i \mathbf{x}_i^T\mathbf{x} + b^*\right). \tag{5.13}$$

One can verify that we solved both dual and primal problems and KKT conditions are satisfied by the 3 steps stated below. Hence, $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\alpha}^*)$ are optimal!

- The assumption of regularity conditions (slater's condition) ensures the primal feasible (1st condition in KKT)
- The dual feasible $\boldsymbol{\mu} \geq 0$ (2nd condition in KKT) are guaranteed when you solve the dual problem since they are the constrains of dual problem
- The constrain on dual variable from 3rd condition in KKT and 4th condition are guaranteed when you solve the dual problem since they are the constrains of dual problem (and the relation between primal variables and dual variables is guaranteed when you solve for primal variables by dual variables)

6 SVM with Soft margin

In the soft margin case, the groups are not completely separated or the margin distance is too small if we classify all points correctly. The goal is to make the margin wide enough and has much less wrongly classified points as possible in the mean time. To quantify the target, we will use hinge function:

$$max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} \quad \forall i = 1, ..., l \tag{6.1}$$

is proportional to the distance from $(\mathbf{x}_i, y_i)$ to the hyperplane when the point is in the margin or in the wrong side of the plane and it is zero while the points are outside the margin in the correct side of the plane.

The loss function could be define using hinge function as follows:

$$f(\mathbf{w}, b) = C\sum_{i=1}^{l} max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} + \frac{1}{2}||\mathbf{w}||^2 \tag{6.2}$$

The first term is to penalty the points that lie in the margin or in the wrong margin. The second term is to maximize the margin distance. Since the function $max\{0, y\}$ is not differentiable, hence, one need to convert it into a differentiable problem by imposing new variables and corresponding constrains. Define

$$\xi_i = max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} \tag{6.3}$$

Then, we get the primal problem:

$$\begin{aligned} min_{\mathbf{w},\boldsymbol{b},\boldsymbol{\xi}}\ f(\mathbf{w},b) = C\sum_{i=1}^{l}\xi_i + \frac{1}{2}||\mathbf{w}||^2 \\ s.t.\quad \xi \geq 0 \quad \forall i = 1...,l \\ -y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq \xi_i - 1 \quad \forall i = 1...,l \end{aligned} \tag{6.4}$$

Lagrangian function is

$$L(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\xi_i + \sum_{i=1}^{l}\mu_i[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)] \tag{6.5}$$

The Lagrangian dual function is

$$G(\boldsymbol{\alpha}, \boldsymbol{\mu}) = min_{(\mathbf{w},b,\boldsymbol{\xi})}L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) \tag{6.6}$$

By the 3rd condition in KKT, note that $(\mathbf{w}, b, \boldsymbol{\xi})$ are primal variables, then, we have

$$\nabla_{\mathbf{w}}L = \mathbf{w} - \sum_{i=1}^{l}\mu_i y_i\mathbf{x}_i = 0 \tag{6.7}$$

$$\nabla_b L = \sum_{i=1}^{l}\mu_i y_i = 0 \tag{6.8}$$

$$\nabla_\xi L = C - \alpha_i - \mu_i = 0 \tag{6.9}$$

plug (6.7) into dual function to satisfy it, we get dual function:

$$G(\boldsymbol{\alpha}, \boldsymbol{\mu}) = C\sum_{i=1}^{l}\xi_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mu_i y_i(\mathbf{x}_i\cdot\mathbf{x}_i)\mu_j y_j - \sum_{i=1}^{l}\alpha_i\xi_i + \sum_{i=1}^{l}\mu_i - \sum_{i=1}^{l}\xi_i\mu_i - \sum_{i=1}^{l}\mu_i y_i(\mathbf{w}^T\mathbf{x}_i) - \sum_{i=1}^{l}\mu_i y_i b \tag{6.10}$$

$$= \sum_{i=1}^{l}\mu_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mu_i y_i(\mathbf{x}_i\cdot\mathbf{x}_i)\mu_j y_j \quad by(6.8, 6.9) \tag{6.11}$$

Hence, the dual problem is given by

$$\begin{aligned} max_{\boldsymbol{\mu}}G(\boldsymbol{\mu}) = \sum_{i=1}^{l}\mu_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mu_i y_i(\mathbf{x}_i\cdot\mathbf{x}_i)\mu_j y_j \\ s.t.\quad \sum_{i=1}^{l}\mu_i y_i = 0 \quad \forall i = 1, ..., l \\ \mu_i \geq 0 \quad \forall i = 1, ..., l \\ \mu_i \leq C \quad \forall i = 1, ..., l \end{aligned} \tag{6.12}$$

The last constrain is obtained by $\mu_i + \alpha_i = C$ and $\alpha_i \geq 0$ and then dual function is a quadratic function only of $\boldsymbol{\mu}$ with linear constrains. It is efficiently solvable by quadratic programming algorithms.

6.1 Determine $\mathbf{w}, b, \xi_i, \alpha_i$ via KKT conditions

After find the $\boldsymbol{\mu}^*$ by solving dual problem via QP algorithm, one can determine

$$\mathbf{w}^* = \sum_i \mu_i y_i x_i \tag{6.13}$$

for i such that $\mu_i > 0$ and such x_i are called support vectors. Note that these support vectors are either incorrectly classified or are classified correctly but are on or inside the margin.

$$\alpha_i = C - \mu_i^* \quad \forall i = 1, ..., l \tag{6.14}$$

To determine b^*, ξ_i^*, we need to play with the 4th KKT condition. there are 2 types inequalities in primal problem.

$$\xi_i \geq 0 \quad \forall i = 1, ..., l \tag{6.15}$$

$$\xi_i \geq 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \quad \forall i = 1, ..., l \tag{6.16}$$

Hence, there are 2 types corresponding complementary slackness conditions:

$$\alpha_i \xi_i = 0 \quad \forall i = 1, ..., l \tag{6.17}$$

$$\mu_i[1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)] = 0 \quad \forall i = 1, ..., l \tag{6.18}$$

since $C = \alpha_i + \mu_i$, then,

$$(C - \mu_i)\xi_i = 0 \quad \forall i = 1, ..., l \tag{6.19}$$

$$\mu_i[1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)] = 0 \quad \forall i = 1, ..., l \tag{6.20}$$

We observe that if $0 < \mu_i < C$, then $\xi_i = 0$ by eq (6.19), and then, we could solve for b by eq (6.20).

$$b^* = average(y_i - \mathbf{w}^T \mathbf{x}_i), \tag{6.21}$$

where i is the one such that $0 < \mu_i < C$.

$$\xi^* = \begin{cases} 0, & 0 \leq \mu_i < C \\ 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b), & \mu_i = C \end{cases} \tag{6.22}$$

Then, one can predict the class of a new point $\mathbf{x}$ by

$$sign\left(\sum_{i, \mu_i > 0} \mu_i y_i \mathbf{x}_i^T \mathbf{x} + b^*\right). \tag{6.23}$$

7 Kernel SVM (C-SVC)

We replace $\mathbf{x}$ with $\phi(\mathbf{x})$ and replace the inner product $\mathbf{x}_i^T \mathbf{x}_j$ with $k(\mathbf{x}_i, \mathbf{x}_j)$ in soft margin case, we get the primal problem of kernel SVM with soft margin:

$$\begin{aligned} min_{\mathbf{w}, \mathbf{b}, \boldsymbol{\xi}} \ f(\mathbf{w}, b) &= C \sum_{i=1}^{l} \xi_i + \frac{1}{2} ||\mathbf{w}||^2 \\ s.t. \quad & \xi_i \geq 0 \quad \forall i = 1..., l \\ & - y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \leq \xi_i - 1 \quad \forall i = 1..., l \end{aligned} \tag{7.1}$$

Similarly, We replace $\mathbf{x}$ with $\phi(\mathbf{x})$ and replace the inner product $\mathbf{x}_i^T \mathbf{x}_j$ with $k(\mathbf{x}_i, \mathbf{x}_j)$, we get the dual problem of kernel SVM with soft margin:

$$\begin{aligned} max_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) &= \sum_{i=1}^{l} \mu_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \mu_i y_i k(\mathbf{x}_i, \mathbf{x}_i) \mu_j y_j \\ s.t. \quad & \sum_{i=1}^{l} \mu_i y_i = 0 \quad \forall i = 1, ..., l \\ & \mu_i \geq 0 \quad \forall i = 1, ..., l \\ & \mu_i \leq C \quad \forall i = 1, ..., l \end{aligned} \tag{7.2}$$

Once the optimization problem is solved as what we did in linear SVM with soft margin case, the output of decision function for a given sample $\mathbf{x}$ becomes:

$$f(\mathbf{x}) = sign\left(\sum_{i \in SV} y_i \mu_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{7.3}$$

We only need to sum over the support vectors because the dual variables $\mu_i = 0$ for other samples.

8 Multi-Classification through SVM

One usually use one vs one to implement multi-classification. For k classes, we will train $C_k^2 = \frac{k(k-1)}{2}$ SVM models and vote for final decision.

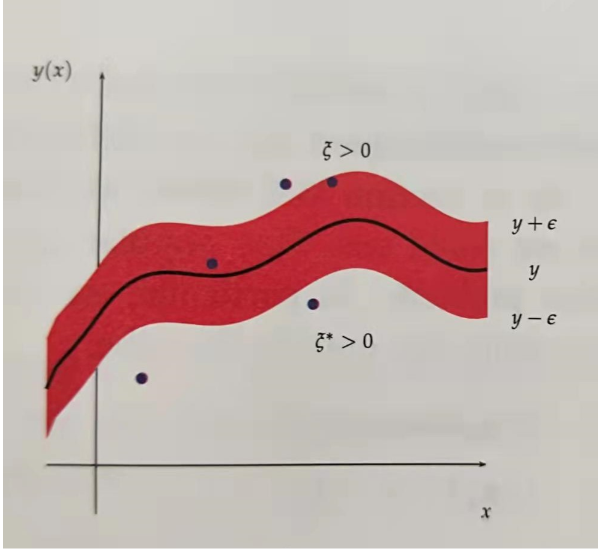

Figure 9.1: ϵ-tube of SVR model

9 Support vector regression (ϵ-SVR and ν-SVR)

Vapnik (Vapnik et al. 1997) proposed a variant of the Huber loss function called the **epsilon insensitive loss function**, defined by

$$L_\epsilon(y, \hat{y}) := \begin{cases} 0, & if\ |y - \hat{y}| \leq \epsilon \\ |y - \hat{y}|, & otherwise \end{cases} \tag{9.1}$$

This means that any point lying inside an ϵ-tube around the predictions is not penalized, as in Fig.9.1. The corresponding objective function is

$$min_{\mathbf{w},b,C}\ f(\mathbf{w}, b) = C \sum_{i=1}^{l} L_\epsilon(y_i, \hat{y}_i) + \frac{1}{2}||\mathbf{w}||^2 \tag{9.2}$$

where $\hat{y}_i = f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b$ and $C = \frac{1}{\lambda}$ is a regularization constant. The objective is convex and unconstrained, but not differentiable, because of the absolute value function in the objective function. One popular approach is to formulate the problem as a constrained optimization problem as we used in soft margin. In particular, we introduce **slack variables** to represent the degree to which each point lies outside the tube:

$$y_i \leq \mathbf{w}^T\mathbf{x}_i + b + \epsilon + \xi_i^+ \tag{9.3}$$

$$y_i \geq \mathbf{w}^T\mathbf{x}_i + b - \epsilon - \xi_i^- \tag{9.4}$$

Then, we get the primal problem:

$$\begin{aligned} min_{\mathbf{w},\boldsymbol{b},\boldsymbol{\xi}^+,\boldsymbol{\xi}^-}\ f(\mathbf{w}, b) = C \sum_{i=1}^{l} (\xi_i^+ + \xi_i^-) + \frac{1}{2}||\mathbf{w}||^2 \\ s.t.\quad \xi_i^+ \geq 0 \quad \forall i = 1..., l \\ \xi_i^- \geq 0 \quad \forall i = 1..., l \\ y_i - (\mathbf{w}^T\mathbf{x}_i + b) \leq \epsilon + \xi_i^+ \quad \forall i = 1..., l \\ y_i - (\mathbf{w}^T\mathbf{x}_i + b) \geq -\epsilon - \xi_i^- \quad \forall i = 1..., l \end{aligned} \tag{9.5}$$

The Lagrangian function is:

$$L(\mathbf{w}, b, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}) = C \sum_{i=1}^{l} (\xi_i^+ + \xi_i^-) + \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i \xi_i^+ - \sum_{i=1}^{l} \beta_i \xi_i^- \tag{9.6}$$

$$+ \sum_{i=1}^{l} \gamma_i [y_i - (\mathbf{w}^T\mathbf{x}_i + b) - \epsilon - \xi_i^+] - \sum_{i=1}^{l} \mu_i [y_i - (\mathbf{w}^T\mathbf{x}_i + b) + \epsilon + \xi_i^-] \tag{9.7}$$

The Lagrangian dual function is

$$G(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu}) = min_{(\mathbf{w},b,\boldsymbol{\xi})} L(\mathbf{w},b,\boldsymbol{\xi}^+,\boldsymbol{\xi}^- \boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\gamma},\boldsymbol{\mu}) \tag{9.8}$$

By the 3rd condition in KKT, note that $(\mathbf{w},b,\boldsymbol{\xi}^+,\boldsymbol{\xi}^-)$ are primal variables, then, we have

$$\begin{align}
\nabla_{\mathbf{w}} L &= \mathbf{w} - \sum_{i=1}^{l} \gamma_i \mathbf{x}_i + \sum_{i=1}^{l} \mu_i \mathbf{x}_i = 0 \tag{9.9}\\
\nabla_b L &= \sum_{i=1}^{l} (-\gamma_i + \mu_i) = 0 \tag{9.10}\\
\nabla_{\xi_i^+} L &= C - \alpha_i - \gamma_i = 0 \tag{9.11}\\
\nabla_{\xi_i^-} L &= C - \beta_i - \mu_i = 0 \tag{9.12}
\end{align}$$

plug them into dual function to satisfy it, we get dual function:

$$\begin{align}
G(\boldsymbol{\gamma},\boldsymbol{\mu}) &= \frac{1}{2}[\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\gamma_j + \sum_{i,j}^{l} \mu_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j - 2\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j] + \sum_{i=1}^{l}(\gamma_i - \mu_i)y_i \tag{9.13}\\
&+ \sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\gamma_j + \sum_{i,j}^{l} \mu_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j - 2\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j \tag{9.14}\\
&= \frac{3}{2}\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\gamma_j + \frac{3}{2}\sum_{i,j}^{l} \mu_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j - 3\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j + \sum_{i=1}^{l}(\gamma_i - \mu_i)y_i. \tag{9.15}
\end{align}$$

Hence, the dual problem is given by

$$\begin{align}
max_{\boldsymbol{\mu}} G(\boldsymbol{\gamma},\boldsymbol{\mu}) = \frac{3}{2}\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\gamma_j + \frac{3}{2}\sum_{i,j}^{l} \mu_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j - 3\sum_{i,j}^{l} \gamma_i(\mathbf{x}_i^T\mathbf{x}_j)\mu_j + \sum_{i=1}^{l}(\gamma_i - \mu_i)y_i \\
s.t. \quad \sum_{i=1}^{l}(\mu_i - \gamma_i) = 0 \quad \forall i = 1,...,l \tag{9.16}\\
0 \le \gamma_i \le C \quad \forall i = 1,...,l \\
0 \le \mu_i \le C \quad \forall i = 1,...,l
\end{align}$$

This is a quadratic function of $2l$ dual variables $(\boldsymbol{\gamma},\boldsymbol{\mu})$ with linear constrains, hence, one can use QP algorithm to solve it. As we showed before that the optimal solution has the form

$$\mathbf{w}^* = \sum_{i}(\gamma_i - \mu_i)\mathbf{x}_i \tag{9.17}$$

where $\gamma_i - \mu_i \ge 0$. Furthermore, it turns out that $\boldsymbol{\gamma} - \boldsymbol{\mu}$ vector is sparse, because we don't care about the errors which are smaller than ϵ. The $\mathbf{x}_i$ for which $\gamma_i - \mu_i \ge 0$ are called the **support vectors**; these are points for which the errors lie on or outside the ϵ tube. Once again, by the complementary slackness conditions, one can determine b^* as

$$\begin{align}
\alpha_i\xi_i^+ = (C - \gamma_i)\xi_i^+ = 0 \quad i = 1,...,l \tag{9.18}\\
\beta_i\xi_i^- = (C - \mu_i)\xi_i^- = 0 \quad i = 1,...,l \tag{9.19}\\
\gamma_i[y_i - (\mathbf{w}^T\mathbf{x}_i + b) - \epsilon - \xi_i^+] = 0 \quad i = 1,...,l \tag{9.20}\\
\mu_i[y_i - (\mathbf{w}^T\mathbf{x}_i + b) + \epsilon + \xi_i^-] = 0 \quad i = 1,...,l \tag{9.21}
\end{align}$$

hence, $b = y_i - \epsilon - \mathbf{w}^T\mathbf{x}_i$ for i such that $0 < \gamma_i < C$ and $b = y_i + \epsilon - \mathbf{w}^T\mathbf{x}_i$ for i such that $0 < \mu_i < C$.

$$b^* = mean\left(y_i - \epsilon - \mathbf{w}^T\mathbf{x}_i\right) + mean\left(y_i + \epsilon - \mathbf{w}^T\mathbf{x}_i\right) \tag{9.22}$$

Once the model is trained, one can predict new point $\mathbf{x}$ using

$$sign\left(\mathbf{w}^{*T}\mathbf{x} + b^*\right) \tag{9.23}$$

plugging in the definition of $\mathbf{w}^*$, we get

$$sign\left(\sum_i(\gamma_i-\mu_i)\mathbf{x}_i^T\mathbf{x}+b^*\right) \tag{9.24}$$

Finally, if we replace $\mathbf{x}_i^T\mathbf{x}$ with $k(x_i,x)$, we get a kernelized solution of kernel ϵ-SVR.

$$sign\left(\sum_i(\gamma_i-\mu_i)k(\mathbf{x}_i,\mathbf{x})+b^*\right) \tag{9.25}$$

Remark: ν-SVR will be added later.