Lecture 6. Support Vector Machine I

Lecturer: Jie Wang Date: April 2, 2020

The major references of this lecture are [2, 1].

1 Introduction

Suppose that we are given a set of data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$, where $y_i \in \mathcal{C} = \{-1, 1\}$. Support vector machine tries to find a linear function $f : \mathbb{R}^d \to \mathbb{R}$ in the form of

$$f(X; \mathbf{w}, b) = b + \sum_{j=1}^{d} w_j X_j,$$

such that

$$y_i = \mathbf{sign}\,(f(\mathbf{x}_i; \mathbf{w}, b)).$$

To fit the data, we need to put all the positive training instances in the positive half space and the negative training instances in the negative half space.

2 SVM for Linearly Separable Cases

To illustrate the idea of SVM, we consider a simple case where the training samples are linearly separable.

Definition 1. A training sample is linearly separable if there exists $(\hat{\mathbf{w}}, \hat{b})$ such that

$$y_i = \mathbf{sign}\,(f(\mathbf{x}_i; \hat{\mathbf{w}}, \hat{b})),\ \forall\, i \in [n], \tag{1}$$

which is equivalent to

$$y_i f(\mathbf{x}_i; \hat{\mathbf{w}}, \hat{b}) > 0,\ \forall\, i \in [n], \tag{2}$$

where $[n] = \{1, \ldots, n\}$.

In this section, we assume that the training sample is linearly separable.

Assumption 1. *The training sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$ is linearly separable.*

However, we can find infinitely many hyperplanes such that the inequality in (2) holds. Which one shall we choose? The SVM classifier makes the decision based on the notion of *geometric margin*.

Definition 2. The geometric margin $\gamma_f(\mathbf{z})$ of a linear classifier $f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ at a point $\mathbf{z}$ is its signed Euclidean distance to the hyperplane $\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$:

$$\gamma_f(\mathbf{z}) = \frac{y_i(\langle \mathbf{w}, \mathbf{z} \rangle + b)}{\|\mathbf{w}\|}.$$

The geometric margin γ_f of a linear classifier f for a sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$ is the minimum geometric margin over the points in the sample, that is

$$\gamma_f = \min_{i \in [n]} \gamma_f(\mathbf{x}_i).$$

***Remark* 1.** The geometric margin of a data instance to a hyperplane can be *negative*, which implies that it falls into the wrong side of the hyperplane. Given a training sample, a negative geometric margin implies that some of the data instances are misclassified.

SVM looks for the hyperplane which maximizes the geometric margin, and thus it is known as the *maximum margin classifier*. Specifically, we can model SVM by the following optimization problem:

$$\max_{\mathbf{w},b} \gamma_f = \max_{\mathbf{w},b} \min_{i\in[n]} \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)}{\|\mathbf{w}\|} = \max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|}\left(\min_{i\in[n]} y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\right). \tag{3}$$

Notice that, the geometric margin is unchanged if we multiply $(\mathbf{w}, b)$ by a *postive* scalar (why positive?). Thus, from the set $\{\lambda(\mathbf{w}, b) : \lambda > 0\}$, we can only consider the pair of parameter values that satisfy the constraint as follows.

$$\min_i y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) = 1. \tag{4}$$

This transforms the problem in (3) to

$$\begin{aligned} &\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|}, \\ &\text{s.t.} \min_i y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) = 1, \end{aligned}$$

which is equivalent to

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \\ &\text{s.t.} \min_i y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) = 1. \end{aligned} \tag{5}$$

By relaxing the constraint (4) to

$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1, \forall\, i \in [n],$$

the problem in (5) changes to

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \\ &\text{s.t.}\, y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1. \end{aligned} \tag{6}$$

Indeed, the problems (5) and (6) are equivalent, that is, one of the constraints in (6) must hold as an equality at the optimal solution.

Question 1.

1. Show there is at least one of the constraints holds as an equality at the optimum.
2. Show there exist at least one positive **and** negative samples such that the equality holds at the optimum.
3. Can we remove the inequalities that hold strictly at the optimum without affecting the solution?

Definition 3. Given a SVM classifier $\langle \mathbf{w}, \mathbf{x}\rangle + b = 0$, the marginal hyperplanes are determined by

$$|\langle \mathbf{w}, \mathbf{x}\rangle + b| = 1.$$

The support vectors are the data instances on the marginal hyperplanes, i.e.,

$$\{\mathbf{x} : |\langle \mathbf{w}, \mathbf{x}\rangle + b| = 1, \mathbf{x} \in \mathcal{S}\}.$$

3 SVM for Non-separable Cases

In most real applications, the training data instances are not linearly separable, that is, for any hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, there exists $\mathbf{x} \in \mathcal{S}$ such that

$$y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) < 0.$$

Thus, the constraints in (6) can not hold simultaneously. To address this problem, we introduce a set of nonnegative *slack variables* $\{\xi_i\}_{i=1}^n$ to relax the constraints as

$$y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1 - \xi_i,\ i \in [n].$$

We can see that the value of ξ_i measures the the vector $\mathbf{x}_i$'s violation of the corresponding inequality $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$. To limit the violations over all data instances, we add a penalty to the objective function in (6), which leads to

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i \in [n]. \end{aligned}$$

4 Elementary Lagrange Duality

Duality plays an important role in analyzing SVM. Besides interesting theoretical results, duality also motives many efficient algorithms for solving SVM.

4.1 Preliminary

Definition 4. [3] Consider a function $f : X \to Y$.

- The value $f(x) \in Y$ that it assumes at the element $x \in X$ is called the image of x.
- The image of a set $A \subset X$ under the mapping f is defined as the set

  $$f(A) := \{y \in Y : \exists x \in A,\ \text{s.t. } f(x) = y\},$$

  that is, $f(A)$ consists of the elements of Y that are images of elements of A.
- The pre-image of a set $B \subset Y$ is defined as

  $$f^{-1}(B) := \{x \in X : f(x) \in B\},$$

  consisting of the elements of X whose images belong to B.

Definition 5. [1] A hyperplane H in $\mathbb{R}^{d+1}$ is specified by a linear equation involving a nonzero vector (μ, μ_0) (called the normal vector of H), where $\mu \in \mathbb{R}^d$ and $\mu_0 \in \mathbb{R}$, and by a constant c as follows:

$$H = \{(\mathbf{w}, z) : \mathbf{w} \in \mathbb{R}^d, z \in \mathbb{R}, \mu_0 z + \langle \mu, \mathbf{w} \rangle = c\}.$$

Any vector $(\bar{\mathbf{w}}, \bar{z})$ that belongs to the hyperplane H specifies the constant c as

$$c = \mu_0 \bar{z} + \langle \mu, \bar{\mathbf{w}} \rangle.$$

Thus, the hyperplane with given normal (μ, μ_0) that pass through a given vector $(\bar{\mathbf{w}}, \bar{z})$ is the set of the points $(\mathbf{w}, z)$ that satisfy the equation:

$$\mu_0 z + \langle \mu, \mathbf{w} \rangle = \mu_0 \bar{z} + \langle \mu, \bar{\mathbf{w}} \rangle.$$

The hyperplane defines two halfspaces: the positive halfspace

$$H^+ = \{(\mathbf{w}, z) : \mu_0 z + \langle \mu, \mathbf{w} \rangle \geq \mu_0 \bar{z} + \langle \mu, \bar{\mathbf{w}} \rangle\}$$

and the negative halfspace

$$H^- = \{(\mathbf{w}, z) : \mu_0 z + \langle \mu, \mathbf{w} \rangle \leq \mu_0 \bar{z} + \langle \mu, \bar{\mathbf{w}} \rangle\}.$$

Hyperplane with normals (μ, μ_0) where $\mu_0 \neq 0$ are referred to as *nonvertical*. A nonvertical hyperplane can be normalized by dividing its normal vector by μ_0, and assuming this is done, we have $\mu_0 = 1$.

4.2 The problem setting

We consider the problem as follows.

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t.} \ & g_i(\mathbf{x}) \leq 0, \ i = 1, \ldots, m, \\ & h_i(\mathbf{x}) = 0, \ i = 1, \ldots, p, \\ & \mathbf{x} \in X, \end{aligned} \tag{7}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [m]$, $h_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [p]$, and $X \subseteq \mathbb{R}^n$. To simplify notations, let $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector function whose i^{th} component is g_i, and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ be a vector function whose i^{th} component is h_i. Then, the problem in (7) becomes

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t.} \ & \mathbf{g}(\mathbf{x}) \leq 0, \\ & \mathbf{h}(\mathbf{x}) = 0, \\ & \mathbf{x} \in X. \end{aligned} \tag{8}$$

We assume that f, $\mathbf{g}$, and $\mathbf{h}$ are continuously differentiable. We call the problem in (8) *the primal problem.*

The so-called feasible set is defined by:

$$D = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X\}. \tag{9}$$

Each element in D is called *feasible solution.* The optimal function value is defined by

$$f^* = \inf_{\mathbf{x} \in D} f(\mathbf{x}). \tag{10}$$

Assumption 2. Feasibility and Boundedness *The feasible set is nonempty and the objective function is bounded from below, that is,*

$$-\infty < f^* = \inf_{\mathbf{x} \in D} f(\mathbf{x}) < \infty.$$

4.3 The visualization lemma

We used to analyze and/or solve optimization problems by focusing on the problem domain. However, taking the perspective of the problems' *codomain* can provide us new insights. Specifically, we consider the set as follows.

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) : \mathbf{x} \in X\}. \tag{11}$$

Notice that, different from $\mathbb{R}^n$ where $\mathbf{x}$ lies in, the set S is a subset of $\mathbb{R}^{m+p+1}$.

Definition 6. Associated with the primal problem, we define the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}).$$

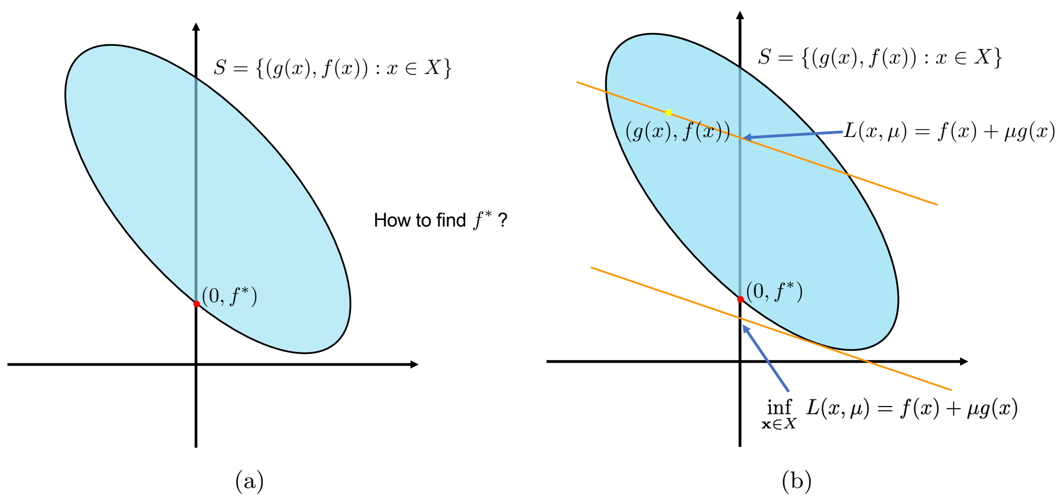

Figure 1: Illustration of the geometric multipliers.

Clearly, for fixed $\mathbf{x}$, the Lagrangian $L(\mathbf{x}, \lambda, \mu)$ is a linear function of (λ, μ).

Definition 7. A vector $(\lambda^*, \mu^*) = (\lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_p^*)$ is said to be a geometric multiplier vector (or simply geometric multiplier) for the primal problem if

$$\lambda_i^* \geq 0,\ i = 1, \ldots, m,$$

and

$$f^* = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*). \tag{12}$$

***Remark* 2.** Notice that, Eq. (12) is a requirement of the geometric multiplier instead of a definition of f^*, which is given in Eq. (10).

Lemma 1. Visualization Lemma

1. *The hyperplane with normal $(\lambda, \mu, 1)$ that passes through a vector $(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x}))$ intercepts the vertical axis $\{(\mathbf{0}, z) : z \in \mathbb{R}\}$ at the level $L(\mathbf{x}, \lambda, \mu)$.*

2. *Among all hyperplanes with normal* $(\lambda, \mu, 1)$ *that contains in their positive halfspace the set* S *defined in (11), the highest attained level of interception of the vertical axis is* $\inf_{\mathbf{x}\in X} L(\mathbf{x}, \lambda, \mu)$.
3. (λ^*, μ^*) *is a geometric multiplier if and only if* $\lambda^* \geq 0$ *and among all hyperplanes with normal* $(\lambda^*, \mu^*, 1)$ *that contain in their positive halfspace the set* S*, the highest attained level of of interception of the vertical axis is* f^*.

Proof.

1. The hyperplane with normal $(\lambda, \mu, 1)$ that passes through a vector $(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x}))$ can be written as

$$\langle \lambda, \mathbf{y}\rangle + \langle \mu, \mathbf{w}\rangle + z = \langle \lambda, \mathbf{g}(\mathbf{x})\rangle + \langle \mu, \mathbf{h}(\mathbf{x})\rangle + f(\mathbf{x}).$$

We note that the right hand side of the above equation is indeed $L(\mathbf{x}, \lambda, \mu)$. Thus, we can write the aforementioned hyperplane as

$$\langle \lambda, \mathbf{y}\rangle + \langle \mu, \mathbf{w}\rangle + z = L(\mathbf{x}, \lambda, \mu).$$

Clearly, we can see that this hyperplane intercepts the vertical axis at the level $L(\mathbf{x}, \lambda, \mu)$ by setting $\mathbf{y} = 0$ and $\mathbf{w} = 0$.

2. The hyperplane H with normal $(\lambda, \mu, 1)$ which intercepts the vertical axis at the level c takes the form of

$$\langle \lambda, \mathbf{y}\rangle + \langle \mu, \mathbf{w}\rangle + z = c.$$

Suppose that S is in the positive halfspace of H. This implies that

$$L(\mathbf{x}, \lambda, \mu) = \langle \lambda, \mathbf{g}(\mathbf{x})\rangle + \langle \mu, \mathbf{h}(\mathbf{x})\rangle + f(\mathbf{x}) \geq c, \forall\, (\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) \in S,$$

which is equivalent to

$$L(\mathbf{x}, \lambda, \mu) \geq c, \forall\, \mathbf{x} \in X.$$

Thus, we have

$$\inf_{\mathbf{x}\in X} L(\mathbf{x}, \lambda, \mu) \geq c.$$

We can see that the maximum value of c is $\inf_{\mathbf{x}\in X} L(\mathbf{x}, \lambda, \mu)$.

3. The claim follows immediately by noting the first two parts.

□

***Remark* 3.** Let $(\mathbf{y}, z) \in \mathbb{R}^{d+1}$. We can define a linear function $\ell : \mathbb{R}^{d+1} \to \mathbb{R}$ as

$$\ell(\mathbf{y}, z) = \langle \lambda, \mathbf{y}\rangle + z,$$

where $\lambda \in \mathbb{R}^d$. By letting

$$\ell(\mathbf{y}, z) = c,$$

where $c \in \mathbb{R}$, we have a hyperplane in $\mathbb{R}^{d+1}$. The linear function takes value c at the points all over the hyperplane. An interesting point we should note is that the level of interception of the vertical axis is c.

If the geometric multiplier (λ^*, μ^*) is known, then we can solve for the optimal solutions $\mathbf{x}^*$ by minimizing the Lagrangian $L(\mathbf{x}, \lambda^*, \mu^*)$ over $\mathbf{x} \in X$. However, for vectors

$$\hat{\mathbf{x}} = \mathop{\mathbf{argmin}}_{\mathbf{x}\in X} L(\mathbf{x}, \lambda^*, \mu^*),$$

it is possible that $\hat{\mathbf{x}}$ is infeasible, that is, some of the constraints may be violated.

Proposition 1. *Let (λ^*, μ^*) be a geometric multiplier. Then, $\mathbf{x}^*$ is a global minimum of the primal problem (8) if and only if $\mathbf{x}^*$ is feasible and*

$$\mathbf{x}^* \in \mathop{\mathbf{argmin}}_{\mathbf{x}\in X} L(\mathbf{x}, \lambda^*, \mu^*),\ \lambda_i^* g_i(\mathbf{x}^*) = 0,\ i = 1, \ldots, m.$$

Proof.

1. $(\Rightarrow)$ Suppose that $\mathbf{x}^*$ is a global minimum of the problem (8). Then, $\mathbf{x}^*$ must be feasible, and thus

   $$f(\mathbf{x}^*) \geq L(\mathbf{x}^*, \lambda^*, \mu^*) \geq f^*.$$

   The definition of f^* leads to $f^* = f(\mathbf{x}^*)$, which implies that the above inequality is an equality. Thus,

   $$f(\mathbf{x}^*) = L(\mathbf{x}^*) = f^* = \inf_{\mathbf{x}\in X} L(\mathbf{x}, \lambda^*, \mu^*).$$

   This leads to

   $$\mathbf{x}^* \in \mathop{\mathbf{argmin}}_{\mathbf{x}\in X} L(\mathbf{x}, \lambda^*, \mu^*), \tag{13}$$

   and

   $$f(\mathbf{x}^*) = L(\mathbf{x}^*) = f(\mathbf{x}^*) + \langle \lambda^*, \mathbf{g}(\mathbf{x}^*)\rangle + \langle \mu^*, \mathbf{h}(\mathbf{x}^*)\rangle.$$

   As $\mathbf{x}^*$ is feasible, that is, $\mathbf{g}(\mathbf{x}^*) \leq 0$ and $\mathbf{h}(\mathbf{x}^*) = 0$, we have

   $$\lambda_i^* g_i(\mathbf{x}^*) = 0,\ i = 1, \ldots, m. \tag{14}$$

2. $(\Leftarrow)$ Suppose that $\mathbf{x}^*$ is feasible and (13) and (14) hold.

   In view of (13) and the fact that (λ^*, μ^*) is the geometric multiplier, we have

   $$L(\mathbf{x}^*, \lambda^*, \mu^*) = f^* = \inf_{\mathbf{x}\in D} f(\mathbf{x}).$$

   Moreover, the feasibility of $\mathbf{x}^*$ and (14) imply that

   $$L(\mathbf{x}^*, \lambda^*, \mu^*) = f(\mathbf{x}^*).$$

   Combining the above two equations leads to

   $$f(\mathbf{x}^*) = \inf_{\mathbf{x}\in D} f(\mathbf{x}),$$

   which implies that $\mathbf{x}^*$ is a global minimum of the primal problem in (8).

□

***Remark* 4.** Let (λ^*, μ^*) be the geometric multiplier. It is possible that none of the elements in the set $\mathbf{argmin}_{\mathbf{x}\in X}\ L(\mathbf{x}, \lambda^*, \mu^*)$ is feasible. A counterexample is

$$\min f(x) = \begin{cases} e^x,\ x \leq 0, \\ 1 - x,\ x \in [0, 1], \\ 0,\ x > 1. \end{cases}$$

$$\text{s.t.}\quad g(x) = x \leq 0.$$

***Remark* 5.** The major motivation for introducing the Lagrangian is to transforming a constrained optimization problem $(\mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X)$ to an unconstrained optimization problem $(\mathbf{x} \in X)$, while the optimal function value remains the same.

References

[1] D. P. Bertsekas. *Nonlinear Programming, 3ed.* Athena Scientific, 2016.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning, 2ed.* The MIT Press, 2018.

[3] V. A. Zorich. *Mathematical Analysis I.* Springer, 2016.