210709 Machine Learning **Spring 2020**

Lecture 7. Support Vector Machine II

Lecturer: Jie Wang Date: April 9, 2020

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

1 The Problem Settings

Recall from the last lecture that, we consider the problem as follows.

$$\begin{aligned} \min_{\mathbf{x}} \, & f(\mathbf{x}) \\ \text{s.t.}\, & \mathbf{g}(\mathbf{x}) \leq 0, \\ & \mathbf{h}(\mathbf{x}) = 0, \\ & \mathbf{x} \in X, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ are all continuously differentiable. The problem in (1) is the so-called primal problem, and its feasible set is denoted by

$$D_0 = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X\}.$$

2 Lagrange Duality

2.1 The dual problem

We introduce the dual function, which is defined for $(\lambda, \mu) \in \mathbb{R}^{m+p}$ by

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

The dual problem is

$$\begin{aligned} & \max q(\lambda, \mu), \\ & \text{s.t. } \lambda \geq 0. \end{aligned}$$

The dual optimal value is defined by

$$q^* = \sup_{\{(\lambda,\mu):\lambda \geq 0\}} q(\lambda, \mu). \tag{2}$$

***Remark* 1.** Recall that, for each (λ, μ), the value of $\inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu)$ is indeed the *interception* of the hyperplane —that has normal $(\lambda, \mu, 1)$ and supports the set

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x}) : \mathbf{x} \in X \subseteq \mathbb{R}^n)\}$$

from below—with the vertical axis. Thus, solving the dual problem corresponds to find the supremum of this interception.

Notice that, $L(\mathbf{x}, \lambda, \mu)$ can be unbounded from below for some (λ, μ). If this is the case, we have $q(\lambda, \mu) = -\infty$. Thus, we introduce the domain of q, which is the set for which $q(\lambda, \mu)$ is finite:

$$\mathbf{dom}\ q = \{(\lambda, \mu) : q(\lambda, \mu) > -\infty\}.$$

We can similarly define the *dual feasible set* by

$$D_1 = \{(\lambda, \mu) : \lambda \geq 0\} \cap \mathbf{dom}\ q = \{(\lambda, \mu) : \lambda \geq 0, q(\lambda, \mu) > -\infty\}.$$

Remark 2. We do not require that $\lambda \geq 0$ for the points in $\mathbf{dom}\,(q)$.

Proposition 1. *The domain of q is convex and q is concave over $\mathbf{dom}\,(q)$.*

Proof. We first show that $\mathbf{dom}\,(q)$ is convex.

Suppose that $q(\lambda_1, \mu_1)$ and $q(\lambda_2, \mu_2)$ are finite and $(\lambda_1, \mu_1) \neq (\lambda_2, \mu_2)$. Let $\theta \in [0, 1]$.

$$\begin{aligned} q(\theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\mu_2) =& \inf_{\mathbf{x}\in X} L(\mathbf{x}, \theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\mu_2) \\ =& \inf_{\mathbf{x}\in X} \theta L(\mathbf{x}, \lambda_1, \mu_1) + (1-\theta)L(\mathbf{x}, \lambda_2, \mu_2) \\ \geq& \inf_{\mathbf{x}\in X} \theta L(\mathbf{x}, \lambda_1, \mu_1) + \inf_{\mathbf{x}\in X} (1-\theta)L(\mathbf{x}, \lambda_2, \mu_2) \\ =& \theta q(\lambda_1, \mu_1) + (1-\theta)q(\lambda_2, \mu_2) \\ >& -\infty. \end{aligned}$$

Thus, we have $\mathbf{dom}\,(q)$ is convex.

The concavity of q can easily be seen by noting that q is the infimum of a set of linear functions of (λ, μ). □

2.2 Weak duality

Theorem 1. Weak Duality Theorem *We have*

$$q^* \leq f^*. \tag{3}$$

Proof. For any (λ, μ), we have

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

Thus,

$$q^* = \sup_{\{(\lambda,\mu):\lambda\geq 0\}} q(\lambda, \mu) \leq f^*.$$

□

Definition 1. Duality gap is defined by

$$f^* - q^*.$$

Remark 3. Duality gap is a commonly used termination condition for a set of optimization algorithms.

Proposition 2.

1. *If there is no duality gap, the set of geometric multipliers is equal to the set of optimal dual solutions.*

2. *If there is duality gap, the set of geometric multipliers is empty.*

Proof. Recall the definition of geometric multipliers that, if (λ^*, μ^*) belongs to a set of geometric multipliers, we have $\lambda^* \geq 0$ and

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

On the other hand, notice that, the left hand side of the above equation is indeed $q(\lambda^*, \mu^*)$.

1. Suppose that $f^* = q^*$.

   We first show that the set of geometric multipliers belongs to the set of optimal dual solutions. Let (λ^*, μ^*) be one of the geometric multipliers. Thus

   $$f^* = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*) = q(\lambda^*, \mu^*) \leq \sup_{\{(\lambda,\mu):\lambda \geq 0\}} q(\lambda, \mu) = q^*.$$

   As $f^* = q^*$, we can see that

   $$q(\lambda^*, \mu^*) = \sup_{\{(\lambda,\mu):\lambda \geq 0\}} q(\lambda, \mu).$$

   This implies that the geometric multiplier (λ^*, μ^*) is one of the optimal dual solutions.

   We next show that the set of optimal dual solutions belongs to the set of geometric multipliers. Let (λ^*, μ^*) be one of the optimal dual solutions. Then

   $$q(\lambda^*, \mu^*) = \sup_{\{(\lambda,\mu):\lambda \geq 0\}} q(\lambda, \mu) = q^*.$$

   In view of $f^* = q^*$, we can see that

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

   This implies that (λ^*, μ^*) is one of the geometric multipliers. The proof is complete.

2. Suppose that $f^* - q^* > 0$. This implies that, for any (λ, μ),

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

   Therefore, there is no (λ^*, μ^*), such that $q(\lambda, \mu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu) = f^*$. Thus, the set of geometric multipliers is empty.

□

Remark 4. If we can find a geometric multiplier, then there is no duality gap. However, the converse is not true.

2.3 Primal and dual optimal solutions

Proposition 3. Optimality Conditions *A pair* $\mathbf{x}^*$ *and* (λ^*, μ^*) *is an optimal solution and geometric multiplier pair if and only if*

$$\mathbf{x}^* \in X, \, \mathbf{g}(\mathbf{x}^*) \leq 0, \mathbf{h}(\mathbf{x}^*) = 0, \qquad \text{(Primal Feasibility)}, \tag{4}$$

$$\lambda^* \geq 0, \qquad \text{(Dual Feasibility)}, \tag{5}$$

$$\mathbf{x}^* \in \mathop{\mathbf{argmin}}_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*), \qquad \text{(Lagrangian Optimality)}, \tag{6}$$

$$\lambda_i^* g_i(\mathbf{x}^*) = 0, \, i = 1, \ldots, m, \qquad \text{(Complementary Slackness)}. \tag{7}$$

Proof.

1. $\Rightarrow$ Suppose that $\mathbf{x}^*$ and (λ^*, μ^*) is an optimal solution and geometric multiplier pair. Then, the primal feasibility and dual feasibility hold.

   Moreover,

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

   which implies the Lagrangian optimality and the complementary slackness.

2. $\Leftarrow$ Suppose that the conditions in (4) to (7) hold. Then

$$f(\mathbf{x}^*) = L(\mathbf{x}^*, \lambda^*, \mu^*) = \min_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*) \leq \inf_{\mathbf{x} \in D_0} L(\mathbf{x}, \lambda^*, \mu^*) \leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) \leq f(\mathbf{x}^*),$$

   which implies that $\mathbf{x}^*$ is the optimal solution and (λ^*, μ^*) is the geometric multiplier.

□

Proposition 4. Saddle Point Theorem *A pair $\mathbf{x}^*$ and (λ^*, μ^*) is an optimal solution-geometric multiplier pair if and only if $\mathbf{x}^* \in X$, $\lambda^* \geq 0$, and $(\mathbf{x}^*, \lambda^*, \mu^*)$ is a saddle point of the Lagrangian, in the sense that*

$$L(\mathbf{x}^*, \lambda, \mu) \leq L(\mathbf{x}^*, \lambda^*, \mu^*) \leq L(\mathbf{x}, \lambda^*, \mu^*), \forall \, \mathbf{x} \in X, \, \lambda \geq 0, \, \mu \in \mathbb{R}^p. \tag{8}$$

Proof.

1. $\Rightarrow$ As the pair $\mathbf{x}^*$ and (λ^*, μ^*) is an optimal solution-geometric multiplier pair, we have (4) to (7) hold. Clearly, we can see that $\mathbf{x}^* \in X$, $\lambda^* \geq 0$, and the Lagrangian optimality in (6) implies that

$$L(\mathbf{x}^*, \lambda^*, \mu^*) \leq L(\mathbf{x}, \lambda^*, \mu^*), \forall \, \mathbf{x} \in X.$$

   Moreover, in view of the definition of geometric multiplier, we have

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

   Thus, combining the feasibility of $\mathbf{x}^*$ and $\lambda \geq 0$ leads to

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

   which completes the proof.

2. $\Leftarrow$ In view of Proposition 3, it suffices to show that (4) and (7) hold. The left half of the saddle point property of the Lagrangian in (8) implies that

$$\begin{aligned} & L(\mathbf{x}^*, \lambda, \mu) \leq L(\mathbf{x}^*, \lambda^*, \mu^*), \, \forall \, \lambda \geq 0, \\ \Rightarrow & f(\mathbf{x}^*) + \langle \lambda, \mathbf{g}(\mathbf{x}^*) \rangle + \langle \mu, \mathbf{h}(\mathbf{x}^*) \rangle \leq L(\mathbf{x}^*, \lambda^*, \mu^*), \, \forall \, \lambda \geq 0. \end{aligned}$$

   In other words, $L(\mathbf{x}^*, \lambda, \mu)$ is upper bounded for any $\lambda \geq 0$. Consequently, we have

$$\mathbf{g}(\mathbf{x}^*) \leq 0, \, \mathbf{h}(\mathbf{x}^*) = 0,$$

   i.e., the primal feasibility (4) holds (otherwise $L(\mathbf{x}^*, \lambda, \mu)$ can not be upper bounded).

To show that the complementary slackness in (7) holds, we combine the primal feasibility of $\mathbf{x}^*$ and left half of (8)

$$f(\mathbf{x}^*) + \langle \lambda, \mathbf{g}(\mathbf{x}^*) \rangle \leq f(\mathbf{x}^*) + \langle \lambda^*, \mathbf{g}(\mathbf{x}^*) \rangle, \forall \lambda \geq 0,$$
$$\xrightarrow{\lambda \to 0} \langle \lambda^*, \mathbf{g}(\mathbf{x}^*) \rangle = \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) \geq 0.$$

On the other hand, in view of the facts that $\lambda^* \geq 0$ and $\mathbf{g}(\mathbf{x}^*) \leq 0$, we have

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

All together, we have

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

Thus, the complementary slackness holds and the proof is complete.

□

2.4 Strong duality

We discuss conditions that ensure the duality gap is zero.

Proposition 5. Strong Duality Theorem - Linear Constraints *Consider the problem in (1). Suppose that f is convex, X is a polyhedron (that is, $X = \{\mathbf{x} : \langle \mathbf{a}_i, \mathbf{x} \rangle \leq b_i, i = 1, \ldots, r\}$), and the optimal value f^* is finite. Then, there is no duality gap and there exists at least one geometric multiplier.*

Proposition 6. Linear and Quadratic Programming Duality *Consider the problem in (1). Suppose that f is convex quadratic, X is a polyhedron, and the optimal value f^* is finite. Then, the primal and dual problems have optimal solutions, and the duality gap is* 0.

3 The Dual Problem of SVM

Recall that the soft margin SVM takes the form of

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \; & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i, \\ \text{s.t.} \; & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, i \in [n], \\ & \xi_i \geq 0, \; i \in [n]. \end{aligned} \tag{9}$$

By Proposition (6), the strong duality holds.

To find the dual problem of (9), we first construct the Lagrangian:

$$L(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) - \sum_{i=1}^{n} \mu_i \xi_i,$$

where $\alpha_i, \mu_i \geq 0$, $i = 1, \ldots, n$.

We next find the dual function:

$$
\begin{aligned}
q(\alpha, \mu) &= \inf_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, \mu) && (10)\\
&= \inf_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle && (11)\\
&+ \inf_{b} -b \sum_{i=1}^{n} \alpha_i y_i && (12)\\
&+ \inf_{\xi} \sum_{i=1}^{n} (C - \alpha_i - \mu_i)\xi_i. && (13)
\end{aligned}
$$

For fixed (α, μ), let $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi})$ be the optimal solution to the above problem. The first order optimal condition implies that

$$
\begin{aligned}
\nabla_{\mathbf{w}} L(\mathbf{w}, b, \xi, \alpha, \mu)|_{\mathbf{w}=\hat{\mathbf{w}}} = 0 &\Rightarrow \hat{\mathbf{w}} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0,\\
\nabla_{b} L(\mathbf{w}, b, \xi, \alpha, \mu)|_{b=\hat{b}} = 0 &\Rightarrow -\sum_{i=1}^{n} \alpha_i y_i = 0,\\
\nabla_{\xi_i} L(\mathbf{w}, b, \xi, \alpha, \mu)|_{\xi_i=\hat{\xi}_i} = 0 &\Rightarrow C - \alpha_i - \mu_i = 0.
\end{aligned}
$$

Plugging the above equations into Eq. (10) leads to

$$
q(\alpha, \mu) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \alpha_i. \tag{14}
$$

Thus, the dual problem of the soft margin SVM in (9) is

$$
\begin{aligned}
\max_{\alpha} \; & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \alpha_i\\
\text{s.t.} \; & \sum_{i=1}^{n} \alpha_i y_i = 0,\\
& C - \alpha_i - \mu_i = 0,\\
& \alpha_i \geq 0,\\
& \mu_i \geq 0, \, i = 1, \ldots, n.
\end{aligned}
$$

By simple algebraic manipulation, we can remove μ from the problem, that is

$$
\begin{aligned}
\min_{\alpha} \; & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^{n} \alpha_i && (15)\\
\text{s.t.} \; & \sum_{i=1}^{n} \alpha_i y_i = 0,\\
& \alpha_i \in [0, C], i = 1, \ldots, n.
\end{aligned}
$$

Proposition 7. *Let* α^* *be one of the optimal solutions to* (15). *Suppose that* α_k^* *is one of the entries of* α^* *and* $\alpha_k^* \in (0, C)$, *then we can find a primal optimal solution by*

$$\begin{aligned}\mathbf{w}^* =& \sum_{i=1}^{n} \alpha^* y_i \mathbf{x}_i, \\ b^* =& y_k - \langle \mathbf{w}^*, \mathbf{x}_k \rangle.\end{aligned}$$

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.