Lecture 3: SVM (Part I) and KKT Conditions

Week 3

Lecturer: Tianyu Wang

1 Maximizing the Margins

To start with, we consider a linearly separable dataset $\{(x_i, y_i)\}_{i=1}^n$. Linear separability is equivalent to existence of a hyperplane that perfectly classifies the dataset. Consider a linear model $f(x) = \begin{cases} +1, & \text{if } w^\top x + b \geq 0, \\ -1, & \text{otherwise.} \end{cases}$ This function defines a hyperplane $h(x) = w^\top x + b$. We want all observations to be far away from the decision boundary. In machine learning, margins are unsigned distances from the data points to the decision boundary.[1]

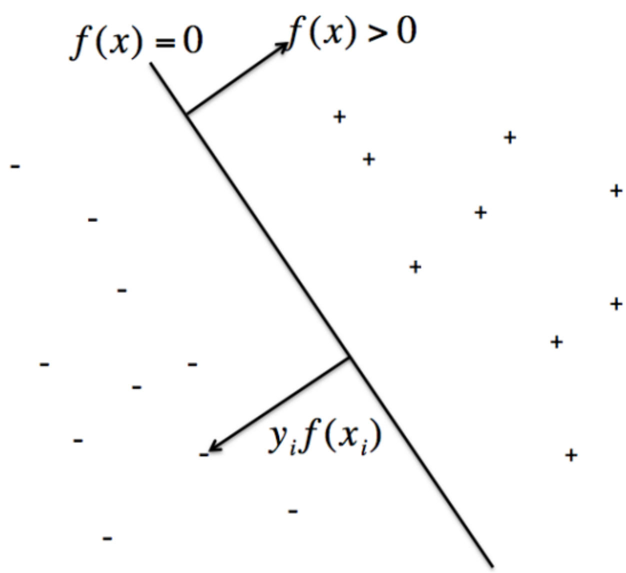

Figure 1: Margins. Picture Credit: C. Rudin.

Let γ be the signed distance from $x \in \mathbb{R}^d$ to the hyperplane $w^\top x + b = 0$. Then we know $x - \gamma \frac{w}{\|w\|_2}$ is on the plane $w^\top x + b = 0$. Thus we have

$$w^\top \left(x - \gamma \frac{w}{\|w\|_2} \right) + b = 0,$$

[1] Some authors uses different definitions for "margin", but it's always the case that the margin summarizes distances from the data points to the decision boundary.

which gives

$$\gamma = \frac{w^\top x + b}{\|w\|_2}.$$

On a (linearly separable) dataset $\{(x_i, y_i)\}_{i=1}^n$. The objective for margin maximization is

$$\max_{w,b,\gamma} \gamma, \quad \text{such that} \quad y_i \frac{w^\top x_i + b}{\|w\|_2} \geq \gamma \quad , \forall i = 1, 2, \cdots, n.$$

Note that the constraint satisfaction is invariant to scaling. Thus we can set $\|w\|_2 = \frac{1}{\gamma}$ and the above optimization problem becomes

$$\max_{w,b} \frac{1}{\|w\|_2}, \quad \text{such that} \quad y_i \left(w^\top x_i + b\right) \geq 1 \quad , \forall i = 1, 2, \cdots, n.$$

This problem can be further relaxed to

$$\min_{w,b} \frac{1}{2}\|w\|_2^2, \quad \text{such that} \quad y_i \left(w^\top x_i + b\right) \geq 1 \quad , \forall i = 1, 2, \cdots, n. \tag{1}$$

To solve this problem, we need some knowledge on constrained convex optimization.

2 Basics of Constrained Convex Optimization and KKT Conditions

Consider a convex optimization problem

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ i = 1, 2, \cdots, m \\ & h_i(x) = 0, i = 1, \cdots, p, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ are differentiable convex functions, and $h_i : \mathbb{R}^n \to \mathbb{R}$ are linear functions.

Alternatively, one can replace the equality constraints, since $h_i(x) = 0$ *iff* $h_i(x) \geq 0$ and $h_i(x) \leq 0$. For simplicity, we omit the equality constraints.

2.1 Intuition behind the Lagrangian

If we receive an infinitely large penalty whenever a constraint is violated, the objective for the above optimization problem can be written as

$$f(x) + \infty \cdot \sum_{i=1}^m \mathbb{I}_{[g_i(x) > 0]}. \tag{2}$$

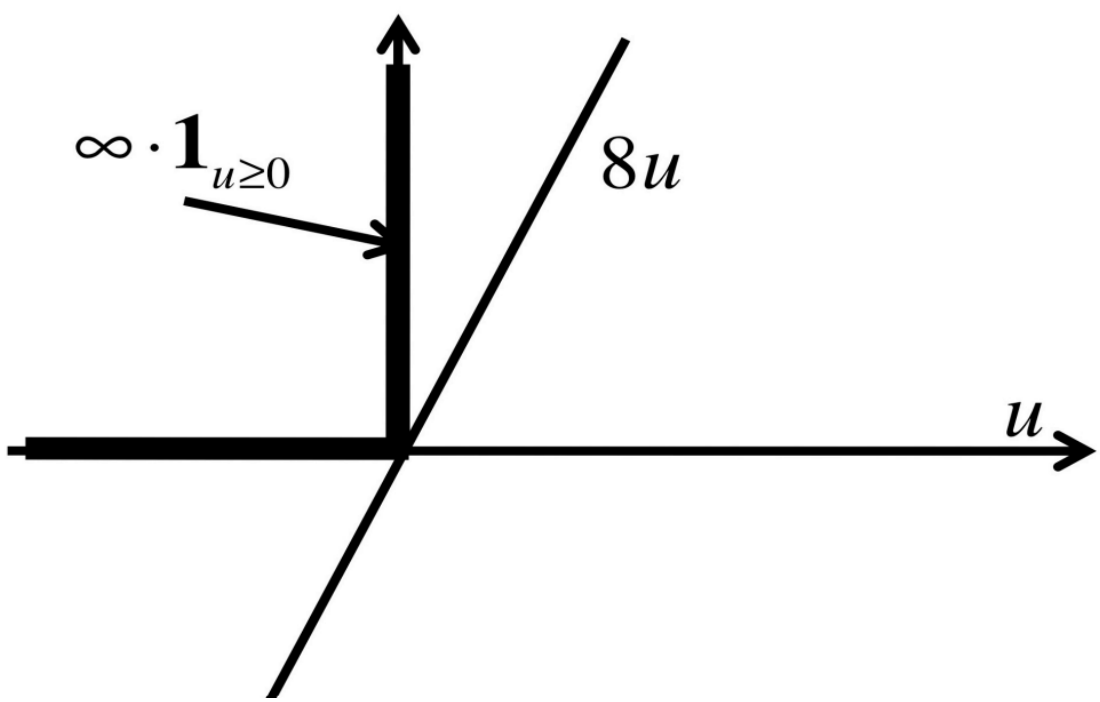

Figure 2: Approximation of the $\infty \cdot \mathbb{I}_{[x>0]}$ function. Picture credit: C. Rudin.

The $\infty \cdot \mathbb{I}_{[x>0]}$ function can be approximated by a linear function, as shown in Figure 2.

We can approximate $\infty \cdot \mathbb{I}_{[x>0]}$ by αx ($\alpha \geq 0$). By using this approximation, we get the Lagrangian for the optimization problem:

$$\boldsymbol{L}(x, \alpha) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x),$$

where α are called dual variables.

The primal objective can be written as

$$\Theta_p(x) := \max_{\alpha_i \geq 0, i=1,2,\cdots,m} \boldsymbol{L}(x, \alpha),$$

and the dual objective can be written as

$$\Theta_d(\alpha) := \min_x \boldsymbol{L}(x, \alpha)$$

2.2 Duality Properties

Proposition 2.1 (max-min inequality)**.** *Consider a real-valued function ϕ and two sets X, Y on which ϕ is defined. It holds that*

$$\max_{y \in Y} \min_{x \in X} \phi(x, y) \leq \min_{x \in X} \max_{y \in Y} \phi(x, y).$$

Proof. For any fixed x, y, it holds that

$$\min_{x' \in X} \phi(x', y) \leq \max_{y' \in Y} \phi(x, y').$$

Since the above inequality holds for any (x, y), we finish the proof by taking the minimum over $x \in X$on the right-hand side, then the maximum over $y \in Y$ on the left-hand side. $\square$

By the max-min inequality, we know that

$$\max_{\alpha \geq 0, i=1,2,\cdots,m} \Theta_d(\alpha) = \max_{\alpha \geq 0, i=1,2,\cdots,m} \min_x \boldsymbol{L}(x, \alpha) \leq \min_x \max_{\alpha \geq 0, i=1,2,\cdots,m} \boldsymbol{L}(x, \alpha) = \min_x \Theta_p(x),$$

which is known as weak duality.

Let p^* be the optimal value of the primal problem and let d^* be the optimal value of the dual problem. By weak duality, the number $p^* - d^*$ is called *duality gap*.

When the equality holds, we know strong duality. That is, the optimal value of the dual objective equals the optimal value of the primal objective. Next we discuss strong duality. A sufficient condition for strong duality is called Slater's condition.

Theorem 2.2 (Strong duality)**.** *When the constraint sets has one strict interior point, then strong duality holds. More formally, if there exists $\widetilde{x}$ such that $g_i(\widetilde{x}) < 0$ for all inequality constraints $i = 1, 2, \cdots, m$ and $h_j(\widetilde{x}) = 0$ for all equality constraints $j = 1, 2, \cdots, p$ (Slater's condition), then*

$$\max_{\alpha \geq 0, i=1,2,\cdots,m} \Theta_d(\alpha) = \min_x \Theta_p(x).$$

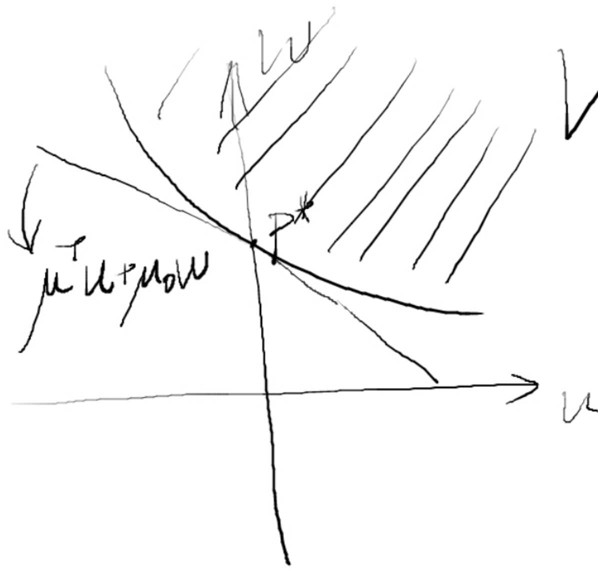

Figure 3: Strong duality proof illustration.

The proof idea is illustrated in Figure 3. We will provide a proof for the the case where there is no equality constraints. The proof for cases with equality constraints uses a similar argument.

Proof. Consider the set

$$V = \{(u, w) \in \mathbb{R}^m \times \mathbb{R} : \exists x, \text{ such that } f(x) \leq w \text{ and } g_i(x) \leq u_i \ \forall i = 1, 2, \cdots, m\}.$$

Step 1. By convexity of f and g_i, the set V is convex. This can be verified by definition.

Step 2. Let p^* be the solution to the primal problem. The point $(0, p^*)$ is on the boundary of V. Otherwise there is a contradiction to the optimality of p^*.

By Claim 1 and Claim 2, there exist $(\mu, \mu_0) \in \mathbb{R}^m \times \mathbb{R}$ and $(\mu, \mu_0) \neq 0$ such that

$$\mu^\top u + \mu_0 w \geq 0^\top u + \mu_0 p^* = \mu_0 p^* \tag{3}$$

for all $(u, w) \in V$.

Step 3. It holds that $\mu_i \geq 0$ for all $i = 0, 1, 2, \cdots, m$. Note that (3) implies

$$\mu^\top u + \mu_0 (w - p^*) \geq 0, \quad \forall (u, w) \in V.$$

Since u_i, w can be arbitrarily large, $\mu_i < 0$ would contradict to the above inequality.

Step 4. It holds that $\mu_0 \neq 0$.

Suppose, in order to get a contradiction, that $\mu_0 = 0$. In this case

$$\inf_{(u,w) \in V} \mu^\top u \geq 0.$$

At the same time,

$$\inf_{(u,w)\in V, u_i\leq 0} \mu^\top u = \inf_x \sum_{i=1}^m \mu_i g_i(x) \leq \sum_{i=1}^m \mu_i g_i(\widetilde{x}) < 0,$$

where the last inequality uses Claim 3 and the Slater's condition.

Step 5. Finish up the proof.

Let $\alpha = \frac{\mu}{\mu_0}$. Note that $\alpha_i \geq 0$ for all $i = 1, 2, \cdots, m$. Plugging this back to (3) implies that

$$\alpha^\top u + w \geq p^*,$$

for all $(u, w) \in V$. Thus we have

$$\Theta_d(\alpha) = \inf_x \left(f(x) + \sum_{i=1}^m \alpha_i g_i(x) \right) = \inf_{u,w} \left(\alpha^\top u + w \right) \geq p^*, \tag{4}$$

which implies

$$\max_{\alpha_i \geq 0, i=1,2,\cdots,m} \Theta_d(\alpha) \geq p^* = \min_x \Theta_p(x). \tag{5}$$

Together with the weak duality theorem, the above result finishes the proof.

□

For cases with equality constraints, one can create another set of variables for the equality constraints, and the general idea is similar.

2.3 KKT Conditions

Theorem 2.3 (Karush-Kuhn-Tucker)**.** *Consider a constraint convex optimization problem where strong duality holds. The following conditions are satisfied at* (x^*, α^*)

- *Primal feasibility:* $g_i(x^*) \leq 0$ *for all* $i = 1, 2, \cdots, m$*;*
- *Dual feasibility:* $\alpha_i^* \geq 0$ *for all* $i = 1, 2, \cdots, m$*;*
- *Stationarity:* $\nabla_x \boldsymbol{L}(x^*, \alpha^*) = 0$*;*
- *Complementary Slackness:* $\alpha_i^* g_i(x^*) = 0$ *for all* $i = 1, 2, \cdots, m$*,*

if and only if x^* *optimally solves the primal problem and* α^* *optimally solves the dual problem.*

The four conditions (primal/dual feasibility, stationarity, complementary slackness) are called Karush-Kuhn-Tucker (KKT) conditions.

Proof. **Necessity.** Let x^* and α^* be primal and dual solutions with zero duality gap, then

$$f(x^*) = \Theta_p(x^*) = \boldsymbol{L}(x^*, \alpha^*) \overset{(i)}{=} \min_x \{f(x) + \sum_{i=1}^m \alpha_i^* g_i(x)\} \leq f(x^*) + \sum_{i=1}^m \alpha_i^* g_i(x^*) \overset{(ii)}{\leq} f(x^*).$$

Thus all of the above inequalities are actually equalities. Thus we have

- Primal feasibility and dual feasibility, which directly follows from that x^* and α^* are primal solution and dual solution.
- Stationarity, since x minimizes $\boldsymbol{L}(x, \alpha^*)$ (by Eq. (i)).
- Complementary Slackness, by Eq. (ii).

Sufficiency. Let the KKT conditions hold. We have

$$\min_x \boldsymbol{L}(x, \alpha^*) \overset{(iii)}{=} \boldsymbol{L}(x^*, \alpha^*) = f(x^*) + \sum_{i=1}^m \alpha_i^* g_i(x^*) \overset{(iv)}{=} f(x^*),$$

where (iii) uses stationarity and (iv) uses complementary slackness. □

3 Back to SVM

Now we have the tools we need to solve for the SVM objective. Recall the SVM objective is

$$\min_{w,b} \frac{1}{2}\|w\|_2^2, \quad \text{such that} \quad y_i\left(w^\top x_i + b\right) \geq 1 \quad, \forall i = 1, 2, \cdots, n.$$

The Lagrangian is

$$\boldsymbol{L}(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^n \alpha_i (1 - y_i\left(w^\top x_i + b\right)).$$

The KKT conditions are

- $1 - y_i(w^\top x_i + b) \leq 0$ for $i = 1, 2, \cdots, n$. (Primal feasibility)
- $\alpha_i \geq 0$ for $i = 1, 2, \cdots, n$. (Dual feasibility)
- $w - \sum_{i=1}^n \alpha_i y_i x_i = 0$ and $\sum_{i=1}^n \alpha_i y_i = 0$. (Stationarity)
- $\alpha_i (1 - y_i\left(w^\top x_i + b\right))$ for $i = 1, 2, \cdots, n$. (Complementary slackness)

The KKT conditions are satisfied at the optimal solution (w^*, b^*, α^*).

3.1 Support Vectors

Look at the complementary slackness condition (and feasibility conditions):

$$\alpha_i(1 - y_i\left(w^\top x_i + b\right)) = 0 \Rightarrow \begin{cases} 1 - y_i\left(w^\top x_i + b\right) < 0 \;\;\&\;\; \alpha_i = 0 & \text{(inactive constraint)} \\ 1 - y_i\left(w^\top x_i + b\right) = 0 \;\;\&\;\; \alpha_i > 0 & \text{(active constraint)} \end{cases} .$$

The support vectors are those data points whose constraint is active (See Figure 4).

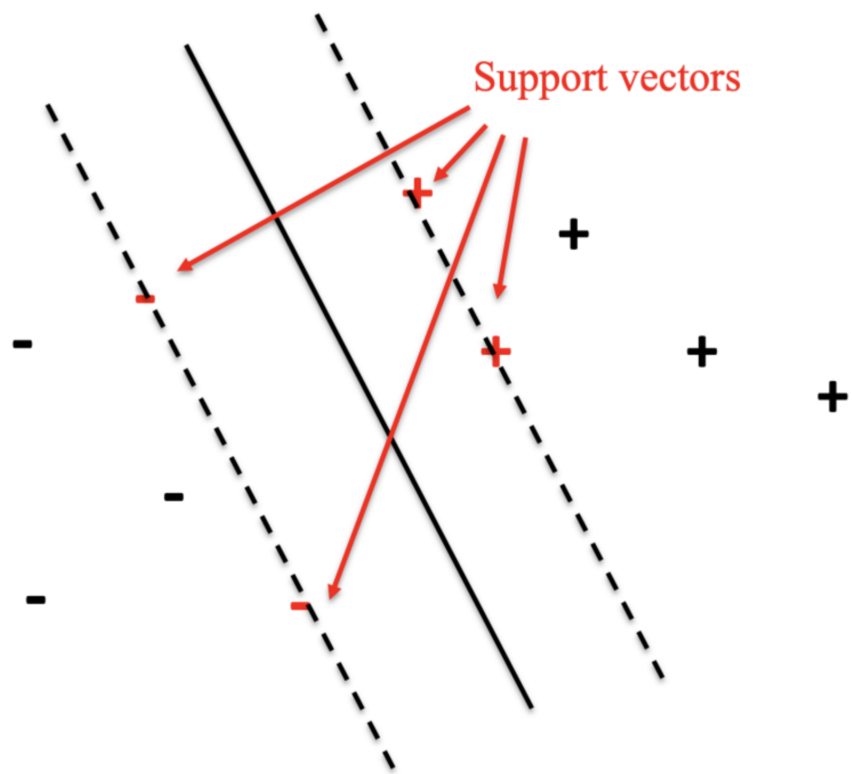

Figure 4: Support Vectors. Picture Credit: C. Rudin.

3.2 Solve for the SVM model

As shown before, the optimal SVM model is specified by (w^*, b^*) that solves the constrained optimization problem. The optimal solution (w^*, b^*) satisfies

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i, \quad b^* = y_{i_0} - x_{i_0}^\top w^* \text{ for some support vector } x_{i_0}.$$

Thus the dual solution α_i^* determines the model. The Lagrangian is

$$\begin{aligned}
\boldsymbol{L}(w^*, b^*, \alpha^*) &= \frac{1}{2}\|w^*\|_2^2 + \sum_{i=1}^n \alpha_i^* \left(1 - y_i \left(x_i^\top w^* + b\right)\right) \\
&= \frac{1}{2}\|w^*\|_2^2 + \sum_{i=1}^n \alpha_i^* - \sum_{i=1}^n \alpha_i^* y_i x_i^\top w^* + \sum_{i=1}^n \alpha_i^* y_i b^* \\
&= \frac{1}{2}\|w^*\|_2^2 + \sum_{i=1}^n \alpha_i^* - \sum_{i=1}^n \alpha_i^* y_i x_i^\top w^* && (\textstyle\sum_{i=1}^n \alpha_i^* y_i = 0) \\
&= \frac{1}{2}\|w^*\|_2^2 + \sum_{i=1}^n \alpha_i^* - \|w^*\|_2^2 && (w^* = \textstyle\sum_{i=1}^n \alpha_i^* y_i x_i) \\
&= \sum_{i=1}^n \alpha_i^* - \frac{1}{2}\|w^*\|_2^2 \\
&= \sum_{i=1}^n \alpha_i^* - \sum_{i=1}^n \sum_{j=1}^n \alpha_i^* \alpha_j^* y_i y_j x_i^\top x_j^\top .
\end{aligned}$$

Thus the dual objective is

$$\Theta_d(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^\top x_j \quad \text{subject to } \begin{cases} \alpha_i \geq 0, \ i = 1, 2, \cdots, n \\ \sum_{i=1}^n \alpha_i y_i = 0. \end{cases} \tag{6}$$

The dual problem is often easier to solve in practice, since the constraints are simpler than the primal problem.

Next time we will continue otn SVM.

Acknowledgement

The SVM was proposed by BE Boser, IM Guyon and VN Vapnik, and later many researchers have contributed to it. TW used lecture notes by Cynthia Rudin to compile this notes. The proof for strong duality is from *Convex Optimization* by Boyd and Vandenberghe.