CSC 665: Support Vector Machines

Chicheng Zhang

November 13, 2019

1 Support vector machines - the maximum margin hyperplane problem

We consider linear classification, where examples $(x_i, y_i)_{i=1}^m$ are such that $x_i \in \mathbb{R}^d$ are features, and $y_i \in \{\pm 1\}$ are binary labels.

Suppose that the training set $S = (x_i, y_i)_{i=1}^m$ is linearly separable, i.e. there exists a linear classifier $(w, b) \in \mathbb{R}^{d+1}$, such that for all i,

$$\begin{cases} \langle w, x_i \rangle + b > 0 & y_i = +1, \\ \langle w, x_i \rangle + b < 0 & y_i = -1. \end{cases} \tag{1}$$

One way to train a linear classifier would be to use the consistency algorithm, i.e. solving a linear program, that finds a (w, b) such that Equation (1) holds. However, note that not all consistent linear classifiers are created equal: some of them are closer to training examples than others. Formally, the distance of a point x in $\mathbb{R}^d$ to a hyperplane $H_{w,b} = \{x_0 : w \cdot x_0 + b = 0\}$ is defined as the shortest distance of x to any of the points in $H_{w,b}$:

$$d(x, H_{w,b}) = \min \left\{ \|x - x_0\| : w \cdot x + b = 0 \right\}. \tag{2}$$

Can we calculate this distance analytically? First, let us assume without loss of generality that $\|w\| = 1$, as any hyperplane $H_{w',b'}$ can be written as $H_{w,b}$ for $\|w\| = 1$ by letting $w = \frac{w'}{\|w'\|}$ and $b = \frac{b'}{\|w'\|}$. Now, consider a point $x_0 \in H_{w,b}$ such that $x_0 = x + \alpha w$ for some α. What is the value of α? Note that

$$\langle w, x + \alpha w \rangle + b = 0,$$

which implies that $\alpha = -(\langle w, x \rangle + b)$.

Claim 1. *For all x_1 in $H_{w,b}$,*

$$\|x_1 - x\| \geq \|x_0 - x\|. \tag{3}$$

Consequenty, $d(x, H_{w,b}) = |\langle w, x \rangle + b|$.

Proof. Note that x_0 and x_1 are both in $H_{w,b}$, $\langle w, x_0 \rangle + b = \langle w, x_1 \rangle + b = 0$. Therefore, $\langle x_1 - x_0, w \rangle = 0$. In other words,

$$\langle x_1 - x_0, x_0 - x \rangle = 0.$$

Now, by Pythagorean theorem,

$$\|x - x_1\|^2 = \|x - x_0\|^2 + \|x_0 - x_1\|^2 \geq \|x - x_0\|^2,$$

whch proves Equation (3). This implies that

$$d(x, H_{w,b}) = \|x - x_0\| = |\langle w, x \rangle + b|.$$

□

Here is a proposal:

> Find the linear classifier (w, b) that not only separates the examples but also maximizes the minimum distances to all examples.

Why is the proposal sensible? One observation is that this classifier is the most "robust". For example, if test examples happen to be just a little distance away from training examples (with the same labels), then this classifier would still classify such examples correctly.

Formally, we can describe the proposal as an optimization problem:

$$\begin{aligned} \underset{w,b,A}{\text{maximize}} \quad & A \\ \text{s.t.} \quad & A > 0, \quad \|w\| = 1, \\ & y_i(\langle w, x_i \rangle + b) > 0, && \forall i \in \{1, \dots, n\}, \\ & |\langle w, x_i \rangle + b| \geq A, && \forall i \in \{1, \dots, n\}. \end{aligned} \tag{4}$$

The above program is not a convex program, and is difficult to optimize directly. Let's make a few transformations to make it a convex program - i.e. finding a convex optimization problem whose solution is related to that of the above optimization problem.

Let's consider the following optimization problem:

$$\begin{aligned} \underset{w,b,A}{\text{maximize}} \quad & A \\ \text{s.t.} \quad & A > 0, \quad \|w\| = 1, \\ & y_i(\langle w, x_i \rangle + b) \geq A, && \forall i \in \{1, \dots, n\}, \end{aligned} \tag{5}$$

Our claim is that the above two optimization problems have the same solutions. Why? Because under $A > 0$, constraints $y_i(\langle w, x_i \rangle + b) > 0$ and $|\langle w, x_i \rangle + b| \geq A$, together, are equivalent to $y_i(\langle w, x_i \rangle + b) \geq A$, as $y_i \in \{\pm 1\}$. For every i, the quantity $y_i(\langle w, x_i \rangle + b)$ is the *margin* of halfspace $H_{w,b}$ on example (x_i, y_i). Therefore the above is also called the "maximum margin hyperplane" problem.

Now let $w' = \frac{w}{A}$, $b' = \frac{b}{A}$. Note that the above optimization problem is equivalent to

$$\begin{aligned} \underset{w',b',A}{\text{maximize}} \quad & A \\ \text{s.t.} \quad & A > 0, \quad \|w'\| = \frac{1}{A}, \\ & y_i(\langle w', x_i \rangle + b') \geq 1, && \forall i \in \{1, \dots, n\}, \end{aligned}$$

Furthermore, this is equivalent to

$$\begin{aligned} \underset{w',b'}{\text{minimize}} \quad & \|w'\| \\ \text{s.t.} \quad & y_i(\langle w', x_i \rangle + b') \geq 1, && \forall i \in \{1, \dots, n\}, \end{aligned}$$

As the function $x \mapsto \frac{1}{2}x^2$ is monotonically increasing for $x > 0$, we get that the above is equivalent to

$$\begin{aligned} \underset{w',b'}{\text{minimize}} \quad & \frac{1}{2}\|w'\|^2 \\ \text{s.t.} \quad & y_i(\langle w', x_i \rangle + b') \geq 1, && \forall i \in \{1, \dots, n\}, \end{aligned} \tag{6}$$

Optimization problem (4) is called the *support vector machine* (SVM). Note that its constraints are all linear inequalities, which defines a convex constraint set. In addition, its optimization objective is a quadratic function over optimization variables, which is a convex function. This implies that it is a *convex optimization* problem.

Recovering the optimal solution of (4)**.** Suppose we have a solution of (6), written as $(w'^\star, b'^\star)$. Note that the optimal A in (5) (thus, in (4)) is $1/\|w'^\star\|$, which is the value of the minimum margin. This implies that in (5) (thus, in (4)), $w^\star = A^\star w'^\star = \frac{w'^\star}{\|w'^\star\|}$, $b^\star = A^\star b'^\star = \frac{b'^\star}{\|w'^\star\|}$. The optimal hyperplane is simply $H_{w^\star,b^\star} = H_{w'^\star,b'^\star}$.

1.1 Optimality condition

To avoid notation clutter, let us drop the apostrophes in optimization problem (6):

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1, \qquad \forall i \in \{1,\dots,n\}, \end{aligned} \tag{7}$$

What property does the optimal solution $(w^\star, b^\star)$ have? We will take a detour and first discuss Lagrangian duality, a fundamental concept in constrained optimization. Let us first write (7) as an unconstrained optimization problem over a slightly more complicated objective:

$$\min_{w,b} \max_{\alpha \geq 0} L(w, b, \xi; \alpha), \tag{8}$$

where $L(w, b; \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i(\langle w, x_i\rangle + b))$.

Define $P(w, b) := \max_{\alpha \geq 0} L(w, b; \alpha)$. Observe that:

$$P(w,b) = \begin{cases} +\infty, & \exists i, 1 - y_i(\langle w, x_i\rangle + b) > 0 \\ \frac{1}{2}\|w\|^2, & \forall i, 1 - y_i(\langle w, x_i\rangle + b) \leq 0 \end{cases}$$

Therefore, optimization problem (8) is equivalent to (7). Now consider switching the orders of min and max in (8):

$$\max_{\alpha \geq 0} \min_{w,b} L(w, b; \alpha).$$

This is called the dual problem of (8) ((8) is called the primal problem). Let's call the optimal primal value $p^\star$ and the optimal dual value $d^\star$. What's the relationship between the primal and dual problems, and their respective optimal solutions?

We state the following result from numerical optimization. Consider a constrained convex optimization problem that has both equality and inequality constraints:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \qquad \forall i \in \{1,\dots,n\}, \\ & h_i(x) = 0, \qquad \forall i \in \{1,\dots,m\}. \end{aligned}$$

Similar as before, we can define Lagrange function $L(x, \alpha, \beta) = f(x) + \sum_{i=1}^n \alpha_i g_i(x) + \sum_{i=1}^m \beta_i h_i(x)$. Define

$$P(x) \triangleq \max_{\alpha \geq 0, \beta} L(x, \alpha, \beta),$$

$$D(\alpha, \beta) = \min_x L(x, \alpha, \beta),$$

$$p^\star = \min_x P(x) = \min_x \max_{\alpha \geq 0, \beta} L(x, \alpha, \beta),$$

$$d^\star = \max_{\alpha \geq 0, \beta} D(\alpha, \beta) = \max_{\alpha \geq 0, \beta} \min_x L(x, \alpha, \beta),$$

we have the following result.

Theorem 1. *Under mild assumptions*[1]*, we have that there exists* $x^\star$*,* $\alpha^\star$*, and* $\beta^\star$*, such that*

1. $x^\star$ *is optimal solution of the primal problem and* $\alpha^\star, \beta^\star$ *is the optimal solution of the dual problem.*
2. *Strong duality holds:*
$$p^\star = L(x^\star, \alpha^\star, \beta^\star) = d^\star.$$
3. *Karush-Kuhn-Tucker (KKT) condition holds:*

$$\begin{aligned}
& \nabla_x L(x^\star, \alpha^\star, \beta^\star) = 0, && \textit{Stationarity} \\
\forall i, \quad & g_i(x^\star) \le 0, h_i(x^\star) \le 0, && \textit{Primal feasible} \\
\forall i, \quad & \alpha_i \ge 0, && \textit{Dual feasible} \\
\forall i, \quad & \alpha_i g_i(x^\star) = 0. && \textit{Complementary slackness}
\end{aligned}$$

Applying the theorem to SVM optimization, we can also recover the primal optimal solution $(w^\star, b^\star)$ from dual solution $\alpha^\star$ by invoking the KKT condition. To see why, recall that in SVM, $L(w, b; \alpha) = \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i(\langle w, x_i\rangle + b))$, hence by stationarity condition,

$$\nabla_w L(w, b; \alpha) = \lambda w - \sum_{i=1}^n \alpha_i y_i x_i = 0,$$

which implies that

$$w^\star = \sum_{i=1}^n \alpha_i^\star y_i x_i.$$

that is, the optimal solution is a linear combination of the feature vectors of training exmaples.

Furthermore, denote by $\mathcal{I} = \left\{i : y_i(\langle w^\star, x_i\rangle + b^\star = 1\right\}$ the set of examples that has margin exactly equal to 1. Complementary slackness says that for all i,

$$\alpha_i^\star(1 - y_i(\langle w^\star, x_i\rangle + b^\star)) = 0.$$

This implies that for an $i \notin \mathcal{I}$, as $y_i(\langle w^\star, x_i\rangle + b^\star > 1$, $\alpha_i = 0$. We call $\mathcal{I}$ the set of *support vectors*, which are the vectors that "contribute" to the optimal solution $w^\star$.

It can also be verified that there exists at least one i, $y_i(\langle w^\star, x_i\rangle + b^\star) = 1$. Pick one such i; $b^\star$ can be recovered by the formula $b^\star = y_i - \langle w^\star, x_i\rangle$.

1.2 Coping with linear non-spearability

Can we still train SVM if the data is not linearly separable? Note that optimization problem (6) will not find a solution, as now the constraint set become infeasible. Generally there are two ways to sidestep this problem: first, introduce nonlinear feature maps; second, relax the SVM formulation to allow for training examples to be classified incorrectly.

For the first approach, we can consider having a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$, so that every example (x_i, y_i) is transformed to $(\phi(x_i), y_i)$. Suppose $(\phi(x_i), y_i)$ is linearly separable, then we compute a SVM over these examples to get $w^\star \in \mathbb{R}^{m+1}$ and $b \in \mathbb{R}$. Our output linear classifier is $\text{sign}(\langle w^\star, \phi(x_i)\rangle + b^\star)$. For example, suppose we have a distribution D over $\mathbb{R}^2 \times \{\pm 1\}$ such that for all examples (x, y)'s on the support of D, $x_1^2 + x_2^2 \le 1 \Leftrightarrow y = +1$. In this case, we can introduce feature map $\phi(x) = (x_1^2, x_2^2)$ to make the dataset linearly separable.

[1]specifically, f, g_i's are convex, h_i's are linear, and there exists w, b, ξ such that all inequality constraints in are stictly satisfied, namely the Slater condition.

For the second approach, we introduce slack variables $\xi_i \geq 0$ for every example i, to allow some example to be misclassified. In addition, we introduce a regularization parameter $\lambda > 0$ that trades off misclassification and margin on correct examples:

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i & (9) \\ \text{s. t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, & \forall i \in \{1, \ldots, n\}, \\ & \xi_i \geq 0, & \forall i \in \{1, \ldots, n\}, \end{aligned}$$

Intuitively, when λ is larger, it focuses more on enforcing large margin on correct examples; when λ is smaller, it forces more on reducing misclassification. Notice that the last two lines can be summarized by: $\forall i \in \{1, \ldots, n\}, \xi_i \geq \max(0, 1 - y_i(\langle w, x_i \rangle + b))$. Therefore, the optimal choice of ξ_i equals $\max(0, 1 - y_i(\langle w, x_i \rangle + b))$. Let $\phi(z) = \max(0, 1 - z)$ and $R(w) = \frac{\lambda}{2}\|w\|^2$. We thus can rewrite optimization problem (9) as:

$$\underset{w,b}{\text{minimize}} \quad \lambda R(w) + \sum_{i=1}^{n} \phi(y_i(\langle w, x_i \rangle + b)). \tag{10}$$

As a convention, we call $\phi(y(\langle w, x \rangle + b))$ the *hinge loss* of linear classifier (w, b) on example (x, y), written as $\ell_{\text{hinge}}((w, b), (x, y))$. When the margin $y(\langle w, x \rangle + b)$ is larger, the hinge loss is smaller. The above form is also called a *regularized loss minimzation* formulation, which captures a wide range of optimization problems in machine learning (by changing loss function ϕ and regularizer R), such as logistic regression, ridge regression, lasso, etc.

Both approaches has its own advantages and drawbacks. For the feature transformation approach, it is unclear if a ϕ will guarantee that the transformed dataset satisfies linear separability. For the soft margin approach, if the dataset is highly linearly nonpseparable (e.g. the unit circle example discussed above), then as it is still learning a linear classifier, it will not perform well. It may be a good idea to combine nonlinear feature map with soft margin in practice.

2 The dual of SVM

Sometimes looking at the dual problem will yield unexpected insights about the original (primal) problem. Indeed, SVM is a canonical example for this statement - we have already seen that the KKT condition implies that we can write the optimal solution $w^\star$ in terms of dual optimal solution $\alpha^\star$. We have discussed the dual problem in an abstract way so far. But what exactly is the dual problem for SVM?

Let us first calculate the dual objective function $D(\alpha) = \min_{w,b} L(w, b; \alpha)$, where

$$L(w, b; \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(\langle w, x_i \rangle + b)).$$

We can write $D(\alpha)$ as follows:

$$D(\alpha) = \sum_{i=1}^{n} \alpha_i + \min_{w} \left(\frac{1}{2}\|w\|^2 - \left\langle w, \sum_{i=1}^{n} \alpha_i y_i x_i \right\rangle \right) + \min_{b} \left(\sum_{i=1}^{n} \alpha_i y_i b \right).$$

Define $g(z) = \begin{cases} -\infty & z = 0 \\ 0 & z \neq 0 \end{cases}$, then

$$D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i y_i x_i \|^2 + g(\sum_{i=1}^{n} \alpha_i y_i).$$

Therefore, $\max_{\alpha \geq 0} D(\alpha)$ is equivalent to

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \| \sum_{i=1}^{n} \alpha_i y_i x_i \|^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \qquad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{11}$$

writing the objective more explicitly, it is

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \left\langle x_i, x_j \right\rangle \alpha_i \alpha_j \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \qquad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{12}$$

Compared to (6), this is also a quadratic program, however, its objective becomes a complicated quadratic function, and its constraints says that α lies in the positive orthant of $\mathbb{R}^n$, which is simpler than the linear inequality constraints in (6).

3 The kernel trick

The dual of SVM (12) uncovers an interesting fact: if we would like to compute the optimal solution of (6), it suffices to solve the dual optimization problem, whose objective function only depends on the pairwise inner product between training examples (as opposed to the original feature vectors of training examples).

This opens up a new opportunity: suppose we have a feature map that is extremely high dimensional (say has dimensionality M) but has succinct representation on pairwise inner product $\left\langle \phi(x), \phi(x') \right\rangle$ (say can be evaluated with time m), then we may avoid paying a time complexity of M in learning the SVM classifier on the transformed examples. Here is the full proposal:

1. Define $k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle$ be the *kernel function* associated with feature mapping ϕ.
2. Solve the dual optimization problem (12), get $(\alpha_i)_{i=1}^m$.
3. By KKT condition, we can recover
$$w^\star = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i),$$
but we only store $w^\star$ *implicitly*, i.e. storing the value of all α_i's.
4. To recover $b^\star$, find an j such that $\alpha_j > 0$, and let
$$b^\star = y_j - \left\langle w^\star, x_j \right\rangle = y_j - \sum_{i=1}^{n} \alpha_i^\star y_i k(x_i, x_j),$$
where we directly evaluate $k(x_i, x_j)$ as opposed to calculating $\phi(x_i)$, $\phi(x_j)$ and take their inner product.
5. To make prediction on future example x, we compute
$$\left\langle w^\star, \phi(x) \right\rangle + b^\star = \sum_{i=1}^{n} \alpha_i^\star k(x_i, x) + b^\star.$$
Same as before, we directly evaluate the kernel function.

As discussed before, each feature map corresponds to a kernel function. Some feature map gives succinct kernel functions, whereas others may not. For example, for input domain $\mathbb{R}^2$, define $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$. It can be checked that its associated kernel function has a succinct form:

$$\left\langle \phi(x), \phi(x') \right\rangle = (x_1x_1' + x_2x_2')^2 = (\left\langle x, x' \right\rangle)^2.$$

However, if we define $\phi(x) = (x_1^2, x_1x_2, x_2^2)$, then its corresponding $k(x, x')$ does not have a succinct form.

Basic properties of kernel functions:

1. if K is the kernel function of ϕ, then for positive c, cK is the kernel function of $\sqrt{c}\phi$.
2. if K_1 (resp. K_2) is the kernel function of ϕ_1 (resp. ϕ_2), then $K_1 + K_2$ is the kernel function of $\phi(x) = (\phi_1(x), \phi_2(x))$.
3. if K_1 (resp. K_2) is the kernel function of ϕ_1 (resp. ϕ_2), then $K_1 \cdot K_2$ is the kernel function of $\phi(x) = \phi_1(x) \otimes \phi_2(x)$, where the $\otimes$ notation denotes the Kronecker product. Suppose $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_m)$. Then,

$$a \otimes b = \begin{bmatrix} a_1 b \\ \ldots \\ a_n b \end{bmatrix} = \begin{bmatrix} a_1b_1 \\ \ldots \\ a_1b_m \\ \ldots \\ a_nb_1 \\ \ldots \\ a_nb_m \end{bmatrix}.$$

   The claim follow from a basic fact about Kronecker product:

$$\langle a \otimes b, c \otimes d \rangle = \langle a, c \rangle \cdot \langle b, d \rangle .$$

4. if K_1 is the kernel function of ϕ_1, and f is an arbitrary scalar function, then $K_2(x, x') = K_1(x, x')f(x)f(x')$ is the kernel function of $\phi_2(x) = f(x)\phi_1(x)$.

Examples of kernel functions:

1. Linear kernel $K_1(x, x') = (1 + \sum_{i=1}^d x_i x_i') = 1 + \left\langle x, x' \right\rangle$. Define feature map $\phi_1(x) \triangleq (1, x_1, \ldots, x_d)$. It can be checked that $k_1(x, x') = \left\langle \phi_1(x), \phi(x') \right\rangle$.
2. Polynomial kernel $K_2(x, x') = (1 + \left\langle x, x' \right\rangle)^s$ for $s \geq 1$. Then define $\phi_2(x) \triangleq \phi_1(x)^{\otimes s} = \phi_1(x) \otimes \ldots \otimes \phi_1(x)$. It can be checked by property of Kronecker product that $\left\langle \phi_2(x), \phi_2(x') \right\rangle = (\left\langle \phi_1(x), \phi_1(x') \right\rangle)^s = k_2(x, x')$.
3. Radial basis function (RBF) kernel $K_3(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$ for $\sigma > 0$. First, note that $K_3(x, x') = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{\|x'\|^2}{2\sigma^2}\right) \cdot K_4(x, x')$, where

$$K_4(x, x') = \exp\left(\frac{\left\langle x, x' \right\rangle}{\sigma^2}\right) = \sum_{i=0}^{\infty} \frac{(\left\langle x, x' \right\rangle)^i}{\sigma^{2i} i!}.$$

   Let us define $\phi_4(x) = \left(\frac{1}{\sigma^i \sqrt{i!}} x^{\otimes i}\right)_{i=0}^{\infty}$; it can be easily seen that it is the feature map of K_4. Therefore, $\phi_3(x) = \left(\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \cdot \frac{1}{\sigma^i \sqrt{i!}} x^{\otimes i}\right)_{i=0}^{\infty}$ is the feature map of K_3. Note that this is an example of a kernel with infinite dimensional feature map, and the primal SVM problem 6 cannot even be explicitly written down.

4. String kernel. Suppose the strings are over a finite alphabet Σ (e.g. $\Sigma = \{A, T, C, G\}$ for DNA sequences). For two strings s and s', define $K_5(s, s') = \left|\left\{t \in \Sigma^\star : t \text{ is a common substring of } s \text{ and } s'\right\}\right|$. Define feature map $\phi_5(s) = (\mathbf{1}(t \text{ is a substring of } s))_{t \in \Sigma^\star}$. It can be checked that $K_5(s, s') = \left\langle \phi_5(x), \phi_5(x') \right\rangle$. This is also an example of a kernel with infinite dimensional feature map, and moreover the input domain is the set of strings as opposed to the familiar Euclidean space.

4 Margin bounds for linear classification - why does SVM work well?

Recall that in PAC learning, we have seen that, given a distribution D over $\mathbb{R}^d \times \{\pm 1\}$ realizable by the set of linear classifiers, any consistent classifier will have an error rate of $O(\frac{d \ln \frac{m}{d}}{m})$ with high probability. Note that the generalization bound depends crucially on the dimensionality of the data. However, it has been observed that SVM works quite well in practice, even if it uses a function kernel whose feature map is extremely-high (or even, infinite) dimensional. What is going on in SVM that makes it effective?

In this section, we give evidence sheding lights on the effecitiveness of SVM in practice. For simplicity, we only consider SVM for homogeneous linear classifiers, that is

$$\begin{aligned} \underset{w'}{\text{minimize}} \quad & \frac{1}{2}\|w'\|^2 \\ \text{s. t.} \quad & y_i \left\langle w', x_i \right\rangle \geq 1, \qquad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{13}$$

We show the following theorem.

Theorem 2. *Fix $B, R > 0$. Suppose S is a set of examples $(x_i, y_i)_{i=1}^n$ drawn iid from distribution D on $\left\{x \in \mathbb{R}^d : \|x\|_2 \leq R\right\} \times \{\pm 1\}$. Then with probability $1 - \delta$, for all classifiers $w \in \left\{w : \|w\|_2 \leq B\right\}$, and all margin parameters $\gamma \in (0, BR]$, we have*

$$\mathop{\mathbb{P}}_{D}(y \left\langle w, x \right\rangle \leq 0) \leq \mathop{\mathbb{P}}_{S}(y \left\langle w, x \right\rangle < \gamma) + \frac{BR}{\gamma}\sqrt{\frac{8 + 4\ln\left(\frac{2}{\delta}\right) + 2\ln\left(1 + \log_2(\frac{BR}{\gamma})\right)}{m}}.$$

The theorem is called a "margin bound", in the sense that the generalization error bound of the classifier depends on two quantities: first, the empirical "margin error" of the classifier, where an example is counted as error when it has a margin smaller than γ; second, a concentration term that decreases with margin γ. Importantly, the theorem is *dimension-free* - it holds as long as examples and linear predictors have bounded norm.

Another feature in the above statement is that it is invariant under positive scaling of B and γ: consider a positive number α. Given a w such that $\|w\|_2 \leq B$ and a margin $\gamma \in (0, BR]$, consider their scaling $w' = \alpha w$ (with norm at most $B' = \alpha B$) and $\gamma' \in (0, \alpha BR]$. Then we have the following identities on events

$$\left\{y \left\langle w, x \right\rangle \leq 0\right\} = \left\{y \left\langle w', x \right\rangle \leq x\right\}, \quad \left\{y \left\langle w, x \right\rangle \leq \gamma\right\} = \left\{y \left\langle w', x \right\rangle \leq \gamma'\right\}.$$

In addition, the generalization bounds depends only on B/γ, which is equal to B'/γ'. This implies that it is impossible to "game" the theorem by initially obtaining a statement with some alternative value of B and use the above scaling reasoning to get a sharper margin bound.

By this theorem, we immediately have the following important consequence regarding SVM.

Corollary 1. *Same setting as above. Suppose $w^\star$ is such that*

$$\mathop{\mathbb{P}}_{D}(y \left\langle w^\star, x \right\rangle \geq \gamma) = 1.$$

Then, with probability $1-\delta$, SVM returns a classifier $\hat{w}$, such that

$$\mathop{\mathbb{P}}_{D}(y\left\langle \hat{w}, x\right\rangle \leq 0) \leq \frac{\|w^\star\|R}{\gamma}\sqrt{\frac{8+4\ln\left(\frac{2}{\delta}\right)+2\ln\left(1+\log_2\left(\frac{\|w^\star\|R}{\gamma}\right)\right)}{m}}.$$

When $\frac{\|w^\star\|R}{\gamma} \ll d$, this theorem provides much stronger guarantees than $O(\frac{d}{m})$ generalization error bound as guaranteed by VC theory. In addition, using advance techniques, one can show that the generalization bound is in fact $O(\frac{\|w^\star\|^2R^2}{\gamma^2 m})$; therefore, so long as $\frac{\|w^\star\|^2R^2}{\gamma^2}$ is smaller than d, this provides more favorable guarantees.

Proof. Note that $w^\star/\gamma$ is a feasible solution of 13. By the optimality of $\hat{w}$, we know that $\|\hat{w}\| \leq \|w^\star/\gamma\|$. Now consider vector $w' = \gamma\hat{w}$. It can be seen that $\mathbb{P}_D(y\left\langle w, x\right\rangle \leq 0) = \mathbb{P}_D(y\left\langle w', x\right\rangle \leq 0)$, $\|w'\| \leq \|w^\star\|$ and for all i, $y_i\left\langle w', x_i\right\rangle \geq \gamma$.

Now, applying Theorem 2 with $w = w'$, along with $B = \|w^\star\|$ and γ, we have that

$$\mathop{\mathbb{P}}_{D}(y\left\langle w', x\right\rangle \leq 0) \leq \mathop{\mathbb{P}}_{S}(y\left\langle w', x\right\rangle < \gamma) + \frac{\|w^\star\|R}{\gamma}\sqrt{\frac{8+4\ln\left(\frac{2}{\delta}\right)+2\ln\left(1+\log_2\left(\frac{\|w^\star\|R}{\gamma}\right)\right)}{m}}.$$

As $\mathbb{P}_S(y\left\langle w', x\right\rangle < \gamma) = 0$, we immediately have the theorem statement. □

The proof of Theorem 2 is slightly involved. It will be based on the following two key steps: first, relating 0-1 error and margin error to a new loss function named ramp loss; second, conduct a Rademacher complexity-based analysis of the ramp loss class, through the way we will also develop general tools bounding the Rademacher complexity of function classes.

Step 1: relating 0-1 error to ramp loss. Define the ramp loss as follows: $\ell_\gamma(w,(x,y)) = \phi_\gamma(y\left\langle w, x\right\rangle)$, where

$$\phi_\gamma(z) = \begin{cases} 1, & z \leq 0, \\ 1-\frac{z}{\gamma}, & 0 < z < \gamma, \\ 0, & z \geq \gamma. \end{cases}$$

Observe that $\mathbf{1}(z \leq 0) \leq \phi_\gamma(z) \leq \mathbf{1}(z \leq \gamma)$. In addition, ϕ_γ is $\frac{1}{\gamma}$-Lipschitz. Therefore,

$$\mathop{\mathbb{P}}_{(x,y)\sim\Delta}(y\left\langle w, x\right\rangle \leq 0) \leq \mathop{\mathbb{E}}_{(x,y)\sim\Delta}\phi_\gamma(y\left\langle w, x\right\rangle) \leq \mathop{\mathbb{P}}_{(x,y)\sim\Delta}(y\left\langle w, x\right\rangle \leq \gamma).$$

For both $\Delta = \mathrm{U}(S)$ or D.

Therefore, it suffices to show the following theorem:

Theorem 3. *Suppose function ϕ is L-Lipschitz. Then with probability $1-\delta'$, for all w such that $\|w\| \leq B$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\phi(y\left\langle w, x\right\rangle) \leq \mathop{\mathbb{E}}_{(x,y)\sim S}\phi(y\left\langle w, x\right\rangle) + LBR\sqrt{\frac{8+4\ln\left(\frac{2}{\delta'}\right)}{m}}.$$

Why does this imply the original theorem statement? fix $\gamma \in (0, BR]$, consider $\phi = \phi_\gamma$ which is $\frac{1}{\gamma}$-Lipschitz. Then, with probability $1-\delta'$, for all w such that $\|w\| \leq B$,

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\phi_\gamma(y\left\langle w, x\right\rangle) \leq \mathop{\mathbb{E}}_{(x,y)\sim S}\phi_\gamma(y\left\langle w, x\right\rangle) + \frac{BR}{\gamma}\sqrt{\frac{8+4\ln\left(\frac{2}{\delta'}\right)}{m}}.$$

Consider the above statement with $\gamma_i = \frac{BR}{2^i}$ and $\delta_i = \frac{\delta}{2i^2}$ for $i = 1, 2, \ldots$. Then by a union bound, with probability $1 - \delta$, for all w such that $\|w\| \le B$, and all $i \in \mathbb{N}_+$,

$$\mathop{\mathbb{E}}_{(x,y)\sim D} \phi_{\gamma_i}(y \langle w, x\rangle) \le \mathop{\mathbb{E}}_{(x,y)\sim S} \phi_{\gamma_i}(y \langle w, x\rangle) + \frac{BR}{\gamma_i}\sqrt{\frac{8 + 4\ln\left(\frac{2}{\delta_i}\right)}{m}}.$$

This implies that for all i in $\mathbb{N}_+$,

$$\mathop{\mathbb{P}}_{(x,y)\sim D}(y \langle w, x\rangle \le 0) \le \mathop{\mathbb{P}}_{(x,y)\sim S}(y \langle w, x\rangle \le \gamma_i) + \frac{BR}{\gamma_i}\sqrt{\frac{8 + 4\ln\left(\frac{2}{\delta_i}\right)}{m}}.$$

Now consider a general $\gamma \in (0, BR]$. Note that we can always find a i in $\mathbb{N}_+$, such that $\gamma_i < \gamma \le 2\gamma_i$. For this choice of i, we have that $\frac{1}{\gamma_i} \le \frac{2}{\gamma}$, implying $i \le 1 + \log_2(\frac{BR}{\gamma})$; in addition,

$$\mathbb{P}(y \langle w, x\rangle \le \gamma_i) \le \mathbb{P}(y \langle w, x\rangle < \gamma).$$

Therefore, for all $\gamma \in (0, BR]$,

$$\mathop{\mathbb{P}}_{(x,y)\sim D}(y \langle w, x\rangle \le 0) \le \mathop{\mathbb{P}}_{(x,y)\sim S}(y \langle w, x\rangle < \gamma) + \frac{BR}{\gamma}\sqrt{\frac{8 + 4\ln\left(\frac{2}{\delta}\right) + 2\ln\left(1 + \log_2(\frac{BR}{\gamma})\right)}{m}}.$$

Step 2: The uniform convergence of Lipschitz losses via Rademacher complexity based analysis. Now our goal comes down to proving Theorem 3. Define loss function class $\mathcal{F} = \left\{\ell_{\phi,w} : \|w\| \le B\right\}$, where $\ell_{\phi,w}(x, y) = \phi(y \langle w, x\rangle)$ is the ϕ-loss induced by classifier w. It can be straightforwardly seen that Theorem 3 is a statement on the uniform convergence of $\mathbb{E}_S f(Z)$ to $\mathbb{E}_D f(Z)$.

By exactly the same reasoning as in the uniform convergence proof (see "Rademacher complexity" note), we can easily show that with probability $1 - \delta'$,

$$\mathop{\mathbb{E}}_{(x,y)\sim D} \phi(y \langle w, x\rangle) - \mathop{\mathbb{E}}_{(x,y)\sim S} \phi(y \langle w, x\rangle) \le 2\,\mathrm{Rad}_m(\mathcal{F}) + LBR\sqrt{\frac{2\ln\left(\frac{2}{\delta'}\right)}{m}}, \tag{14}$$

where $\mathrm{Rad}_m(\mathcal{F}) = \mathbb{E}\,\mathrm{Rad}_S(\mathcal{F})$, and

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m\sim \mathrm{R}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)$$

are the population and empirical Rademacher complexities respectively.

Remark. A careful reader may notice that the empirical Rademacher complexity $\mathrm{Rad}_S(\mathcal{F})$ is defined a bit differently from before; however, the original proof goes through with almost no changes: when we apply McDiarmid's inequality, we need to check the sensitivity of function

$$\sup_{w:\|w\|\le B} \mathop{\mathbb{E}}_{(x,y)\sim D} \phi(y \langle w, x\rangle) - \mathop{\mathbb{E}}_{(x,y)\sim S} \phi(y \langle w, x\rangle),$$

it can be shown that the above function is $\frac{2LBR}{m}$-sensitive. The symmetrization step is also almost identical to before, except that we don't have absolute value operation on the deviation between empirical loss and generalization loss.

Now let us look at $\mathrm{Rad}_m(\mathcal{F})$ more closely; first recall that $\mathrm{Rad}_m(\mathcal{F}) = \mathbb{E}_{S\sim D^m}\,\mathrm{Rad}_S(\mathcal{F})$, where

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m\sim \mathrm{R}} \sup_{w:\|w\|_2\le B} \sum_{i=1}^{m} \sigma_i \phi(y_i \langle w, x_i\rangle).$$

Also, note that $\mathcal{F}$ can be written as the following composite class: $\mathcal{F} = \{\phi \circ g : g \in \mathcal{G}\}$, where $\mathcal{G} = \{m_w : \|w\| \leq B\}$, where $m_w(x, y) = y \langle w, x \rangle$ is the margin function. We have the following lemma that relates the Rademacher complexity of $\mathcal{F}$ and that of $\mathcal{G}$.

Lemma 1 (Contraction Lemma)**.** *Suppose $S = \{z_1, \ldots, z_m\}$ is a dataset of size m. In addition, suppse $\mathcal{G}$ is a function class, and ϕ is an L-Lipschitz function. Then, define $\mathcal{F} = \{\phi \circ g : g \in \mathcal{G}\}$, we have:*

$$\mathrm{Rad}_S(\mathcal{F}) \leq L \, \mathrm{Rad}_S(\mathcal{G}).$$

What do we know about $\mathrm{Rad}_S(\mathcal{G})$?

$$\begin{aligned}
\mathrm{Rad}_S(\mathcal{G}) &= \frac{1}{m} \mathbb{E} \sup_{w:\|w\|_2 \leq B} \sum_{i=1}^{m} \sigma_i y_i \langle w, x_i \rangle \\
&= \frac{1}{m} \mathbb{E} \sup_{w:\|w\|_2 \leq B} \sum_{i=1}^{m} \sigma_i \langle w, x_i \rangle \\
&= \frac{B}{m} \mathbb{E} \| \sum_{i=1}^{m} \sigma_i x_i \| \\
&\leq \frac{B}{m} \sqrt{\mathbb{E} \| \sum_{i=1}^{m} \sigma_i x_i \|^2} \\
&\leq \frac{B}{m} \sqrt{\mathbb{E} \sum_{i=1}^{m} \|x_i\|^2} \leq \frac{BR}{\sqrt{m}}.
\end{aligned}$$

Therefore, $\mathrm{Rad}_S(\mathcal{F}) \leq L \, \mathrm{Rad}_S(\mathcal{G}) \leq LBR \cdot \sqrt{\frac{1}{m}}$. Plugging into Equation (14), and using the elementary inequality that $\sqrt{C} + \sqrt{D} \leq \sqrt{2(C + D)}$, we have that with probability $1 - \delta'$,

$$\mathop{\mathbb{E}}_{(x,y)\sim D} \phi(y \langle w, x \rangle) \leq \mathop{\mathbb{E}}_{(x,y)\sim S} \phi(y \langle w, x \rangle) + LBR \cdot \sqrt{\frac{8 + 4 \ln\left(\frac{2}{\delta'}\right)}{m}}. \tag{15}$$

Proof of the contraction lemma. Recall that

$$m \, \mathrm{Rad}_S(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \phi(f(z_i)).$$

Our high-level strategy is that, consider removing one ϕ at one location at a time, that is, to show

$$\begin{aligned}
& \mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \sigma_i \phi(f(z_i)) \right) \\
\leq \; & \mathbb{E} \sup_{f \in \mathcal{F}} \left(L\sigma_1 f(z_1) + \sum_{i=2}^{m} \sigma_i \phi(f(z_i)) \right) \\
\leq \; & \mathbb{E} \sup_{f \in \mathcal{F}} \left(L\sigma_1 f(z_1) + L\sigma_2 f(z_2) + \sum_{i=3}^{m} \sigma_i \phi(f(z_i)) \right) \\
\leq \; & \ldots \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left(L\sigma_1 f(z_1) + L\sigma_2 f(z_2) + \ldots + L\sigma_m f(z_m) \right).
\end{aligned}$$

We only show the first inequality; the rest steps are fairly similar. We first expand the left hand side by explicitly averaging over random choices of σ_1:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i\phi(f(z_i))\right)=\mathbb{E}\,\frac{1}{2}\left[\sup_{f\in\mathcal{F}}\left(\phi(f(z_1))+\sum_{i=2}^{m}\sigma_i\phi(f(z_i))\right)+\sup_{f'\in\mathcal{F}}\left(-\phi(f'(z_1))+\sum_{i=2}^{m}\sigma_i\phi(f'(z_i))\right)\right]$$

Note that the right hand side can also be written as:

$$\mathbb{E}\,\frac{1}{2}\left[\sup_{f,f'\in\mathcal{F}}\left(\phi(f(z_1))-\phi(f'(z_1))+\sum_{i=2}^{m}\sigma_i\phi(f(z_i))+\sum_{i=2}^{m}\sigma_i\phi(f'(z_i))\right)\right] \tag{16}$$

which can be bounded as follows:

$$\begin{aligned}
(4) \quad &\leq \quad \mathbb{E}\,\frac{1}{2}\left[\sup_{f,f'\in\mathcal{F}}\left(L\big|f(z_1)-f'(z_1)\big|+\sum_{i=2}^{m}\sigma_i\phi(f(z_i))+\sum_{i=2}^{m}\sigma_i\phi(f'(z_i))\right)\right]\\
&\leq \quad \mathbb{E}\,\frac{1}{2}\left[\sup_{f,f'\in\mathcal{F}}\left(Lf(z_1)-Lf'(z_1)+\sum_{i=2}^{m}\sigma_i\phi(f(z_i))+\sum_{i=2}^{m}\sigma_i\phi(f'(z_i))\right)\right]\\
&= \quad \mathbb{E}\,\frac{1}{2}\left[\sup_{f\in\mathcal{F}}\left(Lf(z_1)+\sum_{i=2}^{m}\sigma_i\phi(f(z_i))\right)+\sup_{f'\in\mathcal{F}}\left(-Lf'(z_1)+\sum_{i=2}^{m}\sigma_i\phi(f'(z_i))\right)\right]\\
&= \quad \mathbb{E}\sup_{f\in\mathcal{F}}\left(L\sigma_1 f(z_1)+\sum_{i=2}^{m}\sigma_i\phi(f(z_i))\right).
\end{aligned}$$

where the first inequality uses the Lipschitzness of ϕ, the first equality is based on the observation that there is always a pair of f and f' such that $f(z_1)-f'(z_1)\geq 0$ that achieves the supremum[2] - if $f(z_1)-f'(z_1)<0$, then switch the settings of f and f' will give the same objective value inside the parenthesis. The second inequality is by unpacking the double supremum to the sum of two suprema (roughly speaking, "undoing" the operation in), and the third inequality is by introducing the Rademacher random variable σ_1 back.

[2]within arbitrary precision.