Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1-2 — Week 5

Instructor: S. Minsker *Scribe: S. Minsker*

1.1 Support Vector Machines (SVM)

The general version of the algorithm was proposed by V.Vapnik and C.Cortes around 1995.

Let $X \in \mathbb{H}$, where $\mathbb{H}$ is a separable Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle$ and the corresponding norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$, and $Y \in \{\pm 1\}$ is a binary label, as before. If the distribution Π of X is supported on a finite set and the Hilbert space is infinite-dimensional, then one can always find a separating hyperplane - the hyperplane which separates instances labeled $+1$ from the ones with label -1. There are many hyperplanes that might solve the problem; which one should we choose? One reasonable approach is to choose the hyperplane that represents the largest degree of separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized (figure 1.1).

Let $u \in \mathbb{H}$ such that $||u|| = 1$. Then for some $c \in \mathbb{R}$, the separating hyperplane is given by the affine subspace

$$L_{u,c} := \{x \in \mathbb{H} : \langle u, x \rangle + c = 0\}.$$

Let $y \in \mathbb{H}$. Then the distance between y and $L_{u,c}$ is

$$d(y, L_{u,c}) = |\langle u, y \rangle + c|.$$

Indeed, $y = \langle y, u \rangle u + y^{\perp}$, where $\langle y^{\perp}, u \rangle = 0$. $x \in L_{u,c}$ implies that $x = x_L + v$, where $v = -cu$ and $x_L \perp u$ since $\langle u, x \rangle + c = \langle u, x_L \rangle + \langle u, v \rangle + c = 0$. Finally,

$$d(y, L_{u,c}) = \inf_{x \in L} ||y - x|| = ||y - (y^{\perp} + v)|| = ||\langle y, u \rangle u - v|| = |\langle y, u \rangle + c|.$$

SVM aims to solve the following problem:

Problem 1. Find $u \in \mathbb{H},\ \|u\| = 1,\ c \in \mathbb{R}$ that maximize d (the margin) subject to

$$\begin{aligned} \langle u, X_j \rangle + c &\geq d, \qquad Y_j = 1 \\ \langle u, X_j \rangle + c &\leq -d, \qquad Y_j = -1 \end{aligned}$$

for $j = 1, \dots, n$.

This problem is equivalent to the following:

Problem 2. Minimize $\frac{1}{d}$ subject to

$$Y_j(\langle u, X_j \rangle + c) \geq d$$

for $j = 1, \dots, n$.

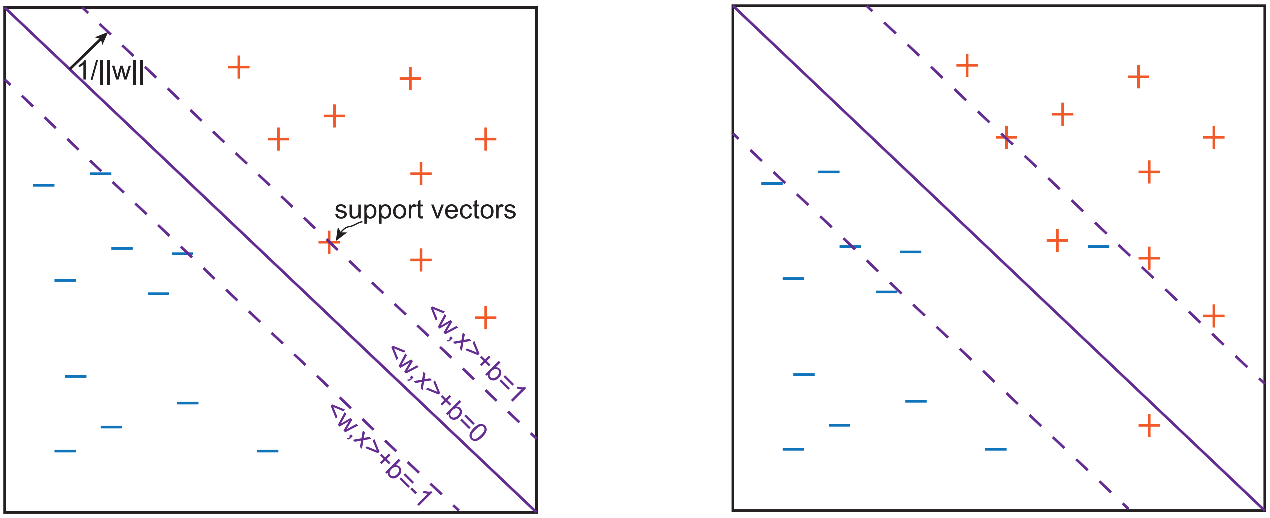

Figure 1.1. Hard-margin SVM (left) and soft-margin SYM (right)

Let $f_{u,c}(\cdot) = \langle u, \cdot \rangle + c$, then the constraint becomes $\min_j Y_j f_{u,c}(X_j) \geq d$, which is in turn equivalent to

$$\min_j Y_j(\langle u/d, X_j \rangle + c/d) \geq 1.$$

Define $w = u/d \in \mathbb{H}$ and $b = c/d \in \mathbb{R}$, so that $||w|| = 1/d$ and we now seek to minimize $\|w\| = 1/d$ subject to

$$\min_j Y_j f_{w,b}(X_j) \geq 1,$$

which we summarize as the following problem

Problem 3. Minimize $\|w\|$ subject to

$$\min_{j=1,\ldots,n} Y_j f_{w,b}(X_j) \geq 1.$$

This is a quadratic programming problem and is termed the "hard-margin SVM". The solution of Problem 3 has several key properties that we summarize below.

Theorem 1. (*Representer theorem*). The solution w^* of Problem 3 is in the linear span of $\{X_1, \ldots, X_n\}$.

Proof: Assume that $w^* = \tilde{w} + \tilde{w}^\perp$, where $\tilde{w} \in$ l.s.$\{X_1, \ldots, X_n\}$ and $\tilde{w}^\perp \perp$ l.s.$\{X_1, \ldots, X_n\}$. If w^* is feasible (i.e. satisfies the constraints), then $\tilde{w}$ is also feasible, because

$$\langle w^*, X_j \rangle = \langle \tilde{w}, X_j \rangle, \qquad \forall j.$$

Note $||w^*|| > ||\tilde{w}||$ if $\tilde{w}^\perp \neq 0$, and if $\tilde{w}^\perp \neq 0$ then the solution can be improved. □

Definition 1. The **support vectors** are a subset $\{X_{i_1}, ..., X_{i_k}\}$ of $\{X_1, \ldots, X_n\}$ such that

$$Y_{i_j} f_{w^*,b^*}(X_{i_j}) = 1, \qquad j = 1, 2, ..., k.$$

Proposition 1. The solution w^* of problem (c) is in the linear span of $\{X_{i_1}, ..., X_{i_k}\}$.

Proof: Use the KKT (Karush-Kuhn-Tucker) conditions. In our case, they become the Fritz-John optimality conditions. Allowing inequality constraints, the KKT conditions generalize the method of Lagrange multipliers. The KKT conditions make applicable in a numerical setting the idea that continuous functions on closed sets are optimized on their boundaries. Here, the optimization problem is

$$\begin{aligned} w^* &= \operatorname*{argmin}_{w\in\mathbb{H}} h(w)(=\|w\|^2) \text{ s.t.} \\ g_j(w) &:= -(Y_j f_{w,b}(X_j) - 1) \leq 0 \qquad \forall j. \end{aligned}$$

Then the KKT conditions state that

$$\nabla h(w^*) + \sum_{i\in I} \alpha_i \nabla g_i(w^*) = 0,$$

where $I = \{i : g_i(w^*) = 0\}$

Exercise 1. Compute the gradients and complete the proof.

The above "hard-margin SVM" enforces a strong restriction (strict separability of classes) on w and b. There is also "soft-margin SVM" which allows misclassification and can be applied cases in which the data are not linearly separable (figure. 1.1). Soft-margin SVMs solve the following problem:

Problem 4. Minimize $\lambda\|w\|^2 + \frac{1}{n}\sum_{j=1}^n \xi_j$ subject to

$$\min_j Y_j f_{w,b}(X_j) \geq 1 - \xi_j, \quad \xi_j \geq 0.$$

for $j = 1, \ldots, n$.

Here, $\lambda > 0$ is a regularization parameter: as $\lambda \to 0$, we recover hard-margin SVM. Note that for any j, either $\xi_j^* = 0$, or $\xi_j^* > 0$ and

$$Y_j f_{w^*,b^*}(X_j) = 1 - \xi_j^*,$$

since otherwise we can make ξ_j smaller, hence making the value of the objective function smaller. We can equivalently express this fact as

$$\xi_j^* = (1 - Y_j f_{w^*,b^*}(X_j))_+ := \max(1 - Y_j f_{w^*,b^*}(X_j), 0).$$

Thus, Problem 4 becomes

Problem 5. Minimize $\frac{1}{n}\sum_{j=1}^n (1 - Y_j f_{w,b}(X_j))_+ + \lambda\|w\|^2$ over w and b.

Recall that $Y_j f_{w,b}(X_j)$ is called the *margin* in binary classification. To build a connection to the Bayes classifier, we introduce the hinge loss function

$$\ell_{\text{hinge}}(y, g(x)) = (1 - yg(x))_+,$$

which is a convex function that also bounds the 0-1 loss function from above (similar with the exponential loss function, see figure 1.2). Now we define the function space (the "base class") as

$$\mathcal{F} = \{f_{w,b} = \langle w, \cdot \rangle + b : \ w \in \mathbb{H}, \ b \in \mathbb{R}\}.$$

T hen Problem 5 can be recast as

Problem 6. Find

$$f_{w^*,b^*} = \operatorname*{argmin}_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{j=1}^{n} \ell_{\text{hinge}}(Y_j f_{w,b}(X_j)) + \lambda \|w\|^2 \right].$$

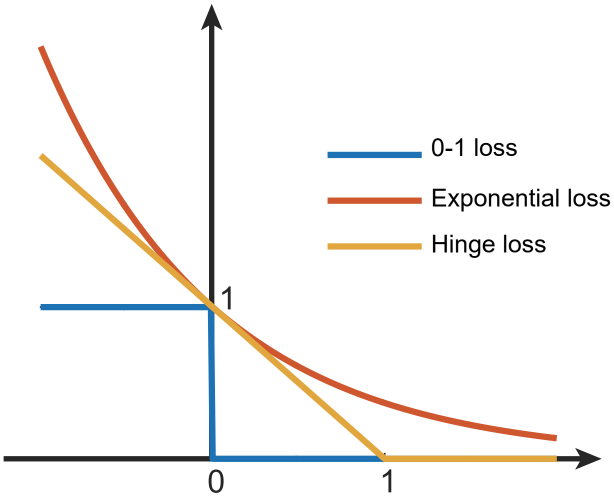

Figure 1.2. Loss fruntions: 0-1, exponential and hinge

Exercise 2. Let $f_* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(1 - Y f(X))_+$. Then $\operatorname{sign}(f_*) = \operatorname{sign}(\eta)$ where η is the Bayes classifier.

1.2 Reproducing kernel Hilbert spaces (RKHS)

Assume $(X, Y) \in S \times \{\pm 1\}$ and S is not a Hilbert space. To apply the ideas of "linear separation" that lead us to SVM, we will need the following *"kernel trick."*
Let $\mathbb{H}$ be separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$, and $\Phi : \mathbb{S} \to \mathbb{H}$ be a mapping, so that

$$(X_1, X_2, \cdots, X_n) \xrightarrow{\Phi} (\Phi(X_1), \Phi(X_2), \cdots, \Phi(X_n)).$$

The idea is that we can try to find a linear hyperplane that separates the images of $X_1, \ldots, X_n$ under Φ in $\mathbb{H}$. Let $f_{w,b}(\Phi(x)) = \langle w, \Phi(x) \rangle + b$, and recall that SVM solves

$$\min_{w \in \mathbb{H}, b \in \mathbb{R}} \frac{1}{n} \sum_{j=1}^{n} (1 - Y_j f_{w,b}(X_j))_+ + \lambda \|w\|^2.$$

According to the *representer theorem*, solution $w^* \in \mathbb{H}$ of the optimization problem that is solved by SVM algorithm can be represented as

$$w^* = \sum_{j=1}^{n} \alpha_j^* \Phi(X_j).$$

Moreover, $\|w^*\|^2 = \sum_{i,j}^{N} \alpha_i^* \alpha_j^* \langle \Phi(x_i), \Phi(x_j) \rangle = \sum_{i,j}^{n} \alpha_i^* \alpha_j^* K(x_i, x_j)$.
Note that for any $x \in S$,

$$\begin{aligned} g^*(x) &= \operatorname{sign}(f_{w^*,b^*}(\Phi(x))) = \operatorname{sign}\left(\langle w^*, \Phi(x) \rangle + b^*\right) \\ &= \operatorname{sign}\left(\sum_{j=1}^{n} \alpha_j^* \langle \Phi(x_j), \Phi(x) \rangle + b^* \right). \end{aligned}$$

<u>Observation:</u> g^* depends on Φ only through the inner products of the form $\langle \Phi(x), \Phi(y) \rangle$. Let $K(x,y) := \langle \Phi(x), \Phi(y) \rangle$ be the "kernel". Our observation implies that to solve the SVM optimization problem, we don't need to know the mapping Φ itself, but only the kernel $K(\cdot,\cdot)$.

It follows from our definition that the kernel K must satisfy the following properties:

(1) $K(x_1, x_2) = K(x_2, x_1), \forall x_1, x_2 \in \mathbb{S}$;

(2) K is nonnegative definite, meaning that $\forall k \geq 1$, $x_1, \cdots, x_k \in \mathbb{S}$, $\alpha_1, \cdots, \alpha_k \in \mathbb{R}$,

$$\sum_{i,j}^{k} \alpha_i \alpha_j K(x_i, x_j) \geq 0.$$

Indeed, $\sum_{i,j=1}^{n} \alpha_i \alpha_j K(\Phi(x_i), \Phi(x_j)) = \|\sum_{i=1}^{n} \alpha_i \Phi(x_i)\|^2 \geq 0$.

1.2.1 Reproducing kernel Hilbert Spaces continued

Assume $K : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ be a kernel that satisfies properties (1) and (2). Let

$$\mathbb{H}_0 = \left\{ \sum_{j=1}^{k} \alpha_j K(x_j, \cdot), \forall k > 1, \alpha_1, \cdots, \alpha_k \in \mathbb{R}, x_1, \cdots, x_k \in \mathbb{S} \right\}$$

be the linear span of the functions of the form $\{K(x,\cdot),\ x \in S\}$. Define the inner product

$$\left\langle \sum_{j=1}^{k} \alpha_j K(x_j, \cdot), \sum_{j=1}^{l} \beta_j K(y_j, \cdot) \right\rangle = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j)$$

It's easy to check that our construction indeed defines an inner product by properties 1 and 2. Define the inner product on $\mathbb{H}_0$ via $\langle K(x_1,\cdot), K(x_2,\cdot) \rangle_{\mathbb{H}_0} := K(x_1, x_2)$ which makes $(\mathbb{H}_0, \langle \cdot,\cdot \rangle)$ an inner product space. Let $(\mathcal{H}, \langle \cdot,\cdot \rangle_{\mathcal{H}})$ be the completion of $(\mathbb{H}_0, \langle \cdot,\cdot \rangle_{\mathbb{H}_0})$. The resulting Hilbert space $\mathcal{H}_K$ (its elements are functions mapping S to $\mathbb{R}$) is called the RKHS with kernel K.

RKHS $\mathcal{H}_K$ has the following **reproducing property:** for any $f \in \mathcal{H}_K$, $\langle f, K(x,\cdot)\rangle_{\mathcal{H}} = f(x)$. Indeed, if $f \in \mathbb{H}_0$, then $f = \sum_{j=1}^k \alpha_j K(x_j, \cdot)$ for some $\alpha_1, \cdots, \alpha_k, x_1, \cdots, x_k$. In this case, $\langle f, K(x,\cdot)\rangle_{\mathbb{H}_0} = \sum_{j=1}^k \alpha_j K(x_j, x) = f(x)$. Since $\mathcal{H}$ is the completion of $\mathbb{H}_0$, the reproducing property holds for any $f \in \mathcal{H}_K$ as well. From this point of view, the soft margin SVM can be recast as penalized risk minimization problem in RKHS $\mathcal{H}_K$, namely

$$\hat{f} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{j=1}^n (1 - Y_j f(X_j))_+ + \lambda \|f\|^2_{\mathcal{H}_K} \right].$$

1.2.2 More on RKHS (optional)

In this section, we will look at RKHS from a different, "functional-analytic" point of view. Let $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space of functions $f : \mathbb{S} \to \mathbb{R}$ where $\mathbb{S}$ is a metric space. Let $\mathbb{H}'$ be the dual space – the space of continuous linear functionals $\ell : \mathbb{H} \to \mathbb{R}$. Given $x \in \mathbb{S}$, let $\delta_x : \mathbb{H} \to \mathbb{R}$ be the "delta-functional" $\delta_x(f) = f(x)$; clearly, δ_x is a linear functional.

Definition 2. The RKHS is a Hilbert space $\mathbb{H}$ such that all point evaluation functionals $\{\delta_x(\cdot)\}_{x \in \mathbb{S}}$ are continuous.

Remark 1. Point evaluation functionals in the space $L_2([0,1])$ of square-integrable functions are not continuous.

By Riesz Representation Theorem, any continuous linear functional can be represented as $\ell(f) = \langle g, f\rangle$ for some $g \in \mathbb{H}$. Hence, for any $x \in \mathbb{S}$, $\exists h_x \in \mathbb{H}$ such that $\delta_x(f) = \langle h_x, f\rangle$ for all $f \in \mathbb{H}$. Define

$$K(x,y) := \langle h_x, h_y\rangle$$

and note that $K(x,y) = h_x(y) = h_y(x)$. Clearly, for any $f \in \mathbb{H}$, $\langle f, K(x,\cdot)\rangle = \langle f, h_x(\cdot)\rangle = f(x)$, so that K has the reproducing property.

1.2.3 RKHS and compact integral operators

Let S be a compact metric space and let v be finite measure with support S. Let $K(\cdot,\cdot) : S \times S \mapsto \mathbb{R}$ be a symmetric, non-negative definite function that is continuous in each variable.

Given $f \in L_2(\nu)$, define the integral operator $T_K : \mathbb{L}_2(v) \to \mathbb{L}_2(v)$ via

$$(T_K f)(x) = \int_{\mathbb{S}} K(x,y) f(y) dv(y)$$

Under our assumptions, T_K is a compact operator, meaning that it maps every bounded set into an relatively compact set. Theory of such compact operators tells us that T_K has eigenvalues $\{\lambda_j(T_K)\}_{j=1}^\infty$ and corresponding eigenfunctions $\{\varphi_j(x)\}_{j=1}^\infty$ such that $(T_K\varphi_j)(x) = \lambda_j \varphi_j(x)$. Moreover, $\lambda_j \to 0$ as $j \to \infty$ and $\{\varphi_j\}_{j=1}^\infty$ form an orthogonal basis of $L_2(\nu)$, meaning that $\langle \varphi_j, \varphi_k\rangle_{\mathbb{L}_2(\nu)} = 0$, $\|\varphi_j\|_{\mathbb{L}_2(\nu)} = 1$.

Given $f \in L_2(\nu)$, we can write $f = \sum_{j=1}^\infty f_j \varphi_j$, where $f_j = \langle f, \varphi_j\rangle_{L_2(\nu)}$, hence $\|f\|^2_{\mathbb{L}_2(v)} = \sum_1^\infty f_j^2$ and

$$T_K(f) = \sum_{j=1}^\infty \lambda_j f_j \varphi_j.$$

Consider an "ellipsoid" $\{f : \langle Kf, f\rangle \le 1\}$. This ellipsoid is the unit ball for the norm $\|f\|_K := \langle T_K f, f\rangle$. Finally, let $\|\cdot\|_{\mathbb{H}_K}$ be the dual norm of $\|\cdot\|_K$ defined via

$$\|f\|_{\mathbb{H}_K} := \sup_{\|g\|_K \le 1} \langle f, g\rangle .$$

Let $\mathbb{H}_K := \{f \in L_2(\nu) : \ \|f\|_{\mathbb{H}_K}\} < \infty$. It is an easy exercise to show that $\|\cdot\|_{\mathbb{H}_K}$ admits the following equivalent characterization:

$$\|f\|^2_{\mathbb{H}_K} = \|T_K^{-1/2} f\|^2_{L_2(\nu)} = \sum_{j\ge 1} \frac{f_j^2}{\lambda_j}.$$

This norm is naturally associated with an inner product

$$\langle f_1, f_2\rangle_{\mathbb{H}_K} = \sum_{j=1}^{\infty} \frac{f_{1,j} f_{2,j}}{\lambda_j}$$

which makes $(\mathbb{H}_K, \langle\cdot,\cdot\rangle_{\mathbb{H}_K}$ a Hilbert space.

Claim: $(\mathbb{H}_K, \langle\cdot,\cdot\rangle_{\mathbb{H}_K}$ is a RKHS with kernel K. To prove this claim, we recall **Mercer's theorem** which states that under our assumptions on the K it admits the following series expansion:

$$K(x, y) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x)\varphi_j(y).$$

It is now easy to check the reproducing property: take $f \in \mathbb{H}_K$ so that $\sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j} < \infty$. Then

$$\langle f, K(\cdot, x)\rangle_{\mathbb{H}_K} = \sum_{j=1}^{\infty} \frac{f_j \lambda_j \varphi_j(x)}{\lambda_j} = \sum_{j=1}^{\infty} f_j \varphi_j(x) = f(x)$$

since $K(\cdot, x) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x)\varphi_j(\cdot)$.

Going back to our original motivation and feature maps, we see that the RKHS $\mathbb{H}_K$ can be obtained via the feature map $\Phi : \mathbb{S} \to \ell^2$ (where ℓ^2 is a space of square summable sequences)

$$\Phi(x) := (\sqrt{\lambda_1}\varphi_1(x), \cdots, \sqrt{\lambda_k}\varphi_k(x), \cdots).$$

We can easily check that

$$\langle \Phi(x), \Phi(y)\rangle_{\ell^2} = \sum_{j=1}^{\infty} \sqrt{\lambda_i}\sqrt{\lambda_i}\varphi_i(x)\varphi_j(y) = \sum_{j=1}^{\infty} \lambda_i \varphi_i(x)\varphi_j(y) = K(x, y),$$

so that this mapping indeed yields RKHS $\mathbb{H}_K$.

1.2.4 Examples of kernels

The function $f \in \mathbb{C}[0, \infty)] \bigcap \mathbb{C}^{\infty}(0, \infty)$ is called a **Completely Monotone Function** (CMF) if $(-1)^k f^{(k)}(r) \ge 0$ for all $r > 0, k \in \mathbb{N}$.

Lemma 1. If f is a CMF iff $f(\|\cdot\|_2^2)$ is a positive definite function.

In particular, if f is a CMF, then $K(x,y) = f(\|x-y\|_2^2)$ is a positive definite kernel. Example of kernels:

(a) Gaussian kernel $K_\sigma(x,y) = \exp\left(\dfrac{-\|x-y\|_2^2}{2\sigma^2}\right)$;

(b) $K(x,y) = \dfrac{1}{(c^2 + \|x-y\|_2^2)^\alpha}, \alpha > 0, c \neq 0$;

(c) Linear Kernel $K(x,y) = \langle x, y\rangle$;

(d) $K(x,y) = (a\langle x,y\rangle + 1)^d, a \in \mathbb{R}, d \in \mathbb{N}$;

(e) Laplace kernel $K(x,y) = \exp\left(\dfrac{\|x-y\|_2}{\sigma}\right), \sigma > 0.$