5.6 Non-Linear SVM

In general, hyperplanes in the input space may not be sufficiently flexible to attain the smallest error rate for a given problem. As noted, the linear SVM solution and prediction of a new case x depends on the x_i's only through the inner product $x_i'x_j$ and $x_i'x$. This fact leads to a straightforward generalization of the linear SVM to the nonlinear case by taking a basis expansion. To enlarge the feature space from the original input space, transformations of x, say, $\phi_m(x)$, $m = 1, \ldots, M$, are considered. Let $\Phi(x) := (\phi_1(x), \ldots, \phi_M(x))'$ be the so-called feature mapping from $\mathbb{R}^d$ to a higher dimensional feature space, which can be even infinite dimensional. Then by replacing the dot product $x_i'x_j$ with $\Phi(x_i)'\Phi(x_j)$, the formulation of the linear SVM is easily extended. The main idea of the nonlinear SVM is to map the data in the original input space to the feature space and find the hyperplane with a large margin in the feature space. For instance, suppose the input space is $\mathbb{R}^2$ and $x = (x_1, x_2)'$. Define $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$ such that $\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)'$. Then the feature mapping gives a new dot product in the feature space,

$$\Phi(x)'\Phi(t) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)(t_1^2, t_2^2, \sqrt{2}t_1t_2)' = (x_1t_1 + x_2t_2)^2 = (x't)^2.$$

In fact, for this nonlinear generalization to work, there is no need to specify the feature mapping Φ explicitly. Specification of the bivariate function $K(x, t) := \Phi(x)'\Phi(t)$ suffices. Then the nonlinear decision function is given by $\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i y_i K(x_i, x) + \hat{\beta}_0$, and essentially it is in the span of $K(x_i, x)$ $i = 1, \ldots, n$. So, the classification boundary is $\{x \in \mathbb{R}^d : \hat{f}(x) = 0\}$ and its shape is determined by K.

From the property of the dot product, such a bivariate function is non-negative definite. Replacing the Euclidean inner product in a linear method with a non-negative definite bivariate function $K(x, t)$ known as a kernel function to obtain its nonlinear generalization is referred to as the "kernel trick" in the machine learning literature. The only condition for a kernel to be valid is that it is a symmetric non-negative (semi-positive) definite function: for every $N \in \mathbb{N}$, $a_i \in \mathbb{R}$, and $z_i \in \mathbb{R}^d$ $(i = 1, \ldots, N)$, $\sum_{i,j}^{N} a_i a_j K(z_i, z_j) \geq 0$. In other words, $K_N := [K(z_i, z_j)]$ is a non-negative definite matrix. Some popularly used kernels are polynomial kernels with dth degree, $K(x, t) = (1 + x't)^d$ or $(x't)^d$ for some positive integer d and the radial basis (or Gaussian) kernel $K(x, t) = \exp(-\|x - t\|^2/2\sigma^2)$ for $\sigma > 0$.

It turns out that this generalization of the linear SVM is closely linked to the nonparametric function estimation procedure, known as the reproducing kernel Hilbert space method (e.g. smoothing splines) and mainly developed in statistics. A rich theory is available for the function estimation technique for nonparametric regression and classification. And the kernelized SVM is a special case of the reproducing kernel Hilbert space methods.

5.7 Kernel Methods

Kernel methods can be viewed as a method of regularization in a function space characterized by a kernel. For general treatment of the function estimation technique in such a space, consider a Hilbert space (complete inner product space) of real-valued functions defined on a domain $\mathcal{X}$, $\mathcal{H} = \{f : f : \mathcal{X} \to \mathbb{R}\}$ with an inner product $\langle f, g \rangle_{\mathcal{H}}$ for $f, g \in \mathcal{H}$. A Hilbert space is a reproducing kernel Hilbert space (RKHS) if there is a kernel function (called reproducing kernel) $K(\cdot, \cdot) : \mathcal{X}^2 \to \mathbb{R}$ such that

i) $K(x, \cdot) \in \mathcal{H}$ for every $x \in \mathcal{X}$, and

ii) $\langle K(x,\cdot), f(\cdot)\rangle_{\mathcal{H}} = f(x)$ for every $f \in \mathcal{H}$ and $x \in \mathcal{X}$.

The second condition is called the *reproducing* property for the obvious reason that K reproduces every f in $\mathcal{H}$. Let $K_x(t) := K(x,t)$ for a fixed x. Then the reproducing property gives the following useful identities

$$K(x,t) = \langle K_t(\cdot), K_x(\cdot)\rangle_{\mathcal{H}} = \langle K_x(\cdot), K_t(\cdot)\rangle_{\mathcal{H}} = K_x(t) = K_t(x).$$

For a comprehensive treatment of the RKHS, see Aronszajn (1950), *Theory of reproducing kernels.*

Lemma 5. *Reproducing kernels are non-negative definite.*

Proof. For every n, and every $x_1, \ldots, x_n \in \mathcal{X}$, and every $a_1, \ldots, a_n \in \mathbb{R}$,

$$\begin{aligned}
\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) &= \sum_{i,j=1}^{n} a_i a_j \langle K(x_i,\cdot), K(x_j,\cdot)\rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{n} a_i K(x_i,\cdot), \sum_{j=1}^{n} a_j K(x_j,\cdot)\rangle_{\mathcal{H}} \\
&= \Big\| \sum_{i=1}^{n} a_i K(x_i,\cdot) \Big\|_{\mathcal{H}}^2 \geq 0.
\end{aligned}$$

□

Conversely, by the Moore-Aronszajn Theorem, for every non-negative definite function $K(x,t)$ on $\mathcal{X}$, there corresponds a unique RKHS $\mathcal{H}_K$ that has $K(x,t)$ as its reproducing kernel. The corresponding RKHS is the completion of the linear space spanned by the functions of the form $\sum_{i=1}^{n} a_i K_{x_i}$ for all choices of n, $a_1, \ldots, a_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in \mathcal{X}$ with the inner product $\langle K_x(\cdot), K_t(\cdot)\rangle_{\mathcal{H}_K} = K(x,t)$.

Now, consider a regularization method in the RKHS, $\mathcal{H}_K$ with reproducing kernel K:

$$\min_{f \in \{1\} \oplus \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda \|h\|_{\mathcal{H}_K}^2, \tag{5.18}$$

where $f(x) = \beta_0 + h(x)$ with $h \in \mathcal{H}_K$ and the penalty $J(f)$ is given by $\|h\|_{\mathcal{H}_K}^2$. In general, the null space can be extended to a larger linear space than $\{1\}$. As an example, $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H}_K = \{h(x) = \beta^\top x \mid \beta \in \mathbb{R}^d\}$ with $K(x,t) = x't$. For $h_1(x) = \beta_1' x$ and $h_2(x) = \beta_2' x \in \mathcal{H}$, $\langle h_1, h_2\rangle_{\mathcal{H}_K} = \beta_1'\beta_2$. Then for $h(x) = \beta' x$, $\|h\|_{\mathcal{H}_K}^2 = \|\beta' x\|_{\mathcal{H}_K}^2 = \|\beta\|^2$. Taking $f(x) = \beta_0 + \beta' x$ and the hinge loss $L(f(x), y) = (1 - yf(x))_+$ gives the linear SVM as a regularization method in $\mathcal{H}_K$. So, encompassing the linear SVM as a special case, the SVM can be cast as a regularization method in an RKHS $\mathcal{H}_K$, which finds $f(x) = \beta_0 + h(x) \in \{1\} \oplus \mathcal{H}_K$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|h\|_{\mathcal{H}_K}^2. \tag{5.19}$$

See Wahba (1997), *Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV* for further discussion of the perspective.

Let $\hat{f}$ be the solution to the optimization problem (5.18). The following *representer* theorem reveals a useful representation of the minimizer, which is explained by a simple geometry of Hilbert spaces.

Theorem 15 (Kimeldorf and Wahba, 1971)**.** *The minimizer of* (5.18) *has a representation of the form*

$$\hat{f}(x) = b + \sum_{i=1}^{n} c_i K(x_i, x)$$

where b *and* $c_i \in \mathbb{R}$, $i = 1, \ldots, n$.

Proof. Consider the linear subspace of $\mathcal{H}_K$ spanned by $K(x_i, x)$, $i = 1, \ldots, n$ and the decomposition of $\mathcal{H}_K$ into the subspace and its orthogonal complement space. Then for each $h \in \mathcal{H}_K$, $h(x) = \sum_{i=1}^{n} c_i K(x_i, x) + \rho(x)$ for some ρ, an element in $\mathcal{H}_K$ such that $\rho \perp K(x_i, x)$. Let $f(x) = b + h(x) = b + \sum_{i=1}^{n} c_i K(x_i, x) + \rho(x)$. Then

$$\begin{aligned} f(x_i) &= b + h(x_i) = b + \langle h, K(x_i, \cdot)\rangle_{\mathcal{H}_K} \\ &= b + \sum_{j=1}^{n} c_j K(x_j, x_i) + \langle \rho, K(x_i, \cdot)\rangle_{\mathcal{H}_K} \\ &= b + \sum_{j=1}^{n} c_j K(x_j, x_i). \end{aligned}$$

So, the empirical risk does not depend on ρ. However,

$$\begin{aligned} J(f) &= \|h\|^2_{\mathcal{H}_K} = \langle h, h\rangle_{\mathcal{H}_K} \\ &= \langle \sum_{j=1}^{n} c_j K_{x_j} + \rho, \sum_{j=1}^{n} c_j K_{x_j} + \rho\rangle_{\mathcal{H}_K} \\ &= \langle \sum_{j=1}^{n} c_j K_{x_j}, \sum_{j=1}^{n} c_j K_{x_j}\rangle_{\mathcal{H}_K} + \langle \rho, \rho\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) + \|\rho\|^2_{\mathcal{H}_K}. \end{aligned}$$

Therefore, the minimizer $\hat{f}$ must have $\rho(x) = 0$, which gives the form stated above. □

Notice that the theorem holds for any loss L. For example, convexity is not necessary. As previously mentioned, the kernel trick leads to the SVM solution given as

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i y_i K(x_i, x) + \hat{\beta}_0.$$

It agrees with what the representer theorem generally implies for the SVM formulation. The theorem says that even if $\mathcal{H}_K$ is infinite dimensional, the minimizer resides in a finite dimensional space. The representation renders it feasible to characterize the solution although actual computation necessary for finding $\hat{f}$ depends largely on the loss L. In particular, the SVM solution for (5.19) can be found by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\{1 - y_i(b + \sum_{j=1}^{n} c_j K(x_j, x_i))\}_+ + \lambda \sum_{i,j} c_i c_j K(x_i, x_j)$$

over b and c_i's. Again, it is a quadratic programming problem. As with any other regularization methods, the choice of the tuning parameter λ is important for classification accuracy and one may choose it by cross validation in practice.