CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 9**

Representer Theorem and Constructing Kernels

Lecturer: Peter L. Bartlett *Scribe: R.J. Honicky*

1 Recap

We saw in previous lectures that Support Vector Machines can be viewed as calculating:

$$J^* = \min_{f\in\mathcal{H}} \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i))$$

Recall that the optimal f^* achieving J^* above could be represented as

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\ .$$

2 Representer Theorem

We can generalize this and prove a simple but powerful theorem called the "Representer Theorem," which shows that even in an *infinite dimensional* reproducing kernel Hilbert space, the minimizer f^* can be represented as an (at most) n-dimensional linear combination of kernel functions.

Theorem 2.1. For a kernel k, its RKHS $\mathcal{H}$ a function $L : \mathbb{R}^n \to \mathbb{R}$ and a non-decreasing $\Omega : \mathbb{R} \to \mathbb{R}$, if

$$\begin{aligned} J^* &= \min_{f\in\mathcal{H}} J(f) \\ &= \min_{f\in\mathcal{H}}\ L(f(x_1), \dots, f(x_n)) + \Omega(\|f\|^2_{\mathcal{H}}) \end{aligned}$$

then for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$,

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \tag{1}$$

has

$$J(f^*) = J^*.$$

Furthermore, if Ω is strictly increasing, any minimizer f of J can be expressed in the form (1).

Proof.

Consider the subspace spanned by the evaluation functionals:

$$\mathrm{span}\{k(x_i, \cdot) : 1 \le i \le n\}.$$

Consider also the projection $f_{\|}$ of f on this subspace:

$$f = f_{\|} + f_{\perp}.$$

Since

$$\begin{aligned} f(x_i) &= \langle f, k(x_i, \cdot) \rangle \\ &= \langle f_{\|}, k(x_i, \cdot) \rangle + \langle f_{\perp}, k(x_i, \cdot) \rangle \\ &= \langle f_{\|}, k(x_i, \cdot) \rangle \\ &= f_{\|}(x_i), \end{aligned}$$

the value $L(f(x_i), \ldots, f(x_n))$ depends only on $f_{\|}$.

But

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \|f_{\|}\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2 \\ &\geq \|f_{\|}\|_{\mathcal{H}}^2, \end{aligned}$$

and thus

$$\begin{aligned} J(f) &= L(f) + \Omega(\|f\|_{\mathcal{H}}^2) \\ &= L(f_{\|}) + \Omega(\|f_{\|}\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2). \end{aligned}$$

But this since Ω is non-decreasing, it must be the case that for any function f that minimizes J, $\Omega(\|f\|_{\mathcal{H}}^2) = \Omega(\|f_{\|}\|_{\mathcal{H}}^2)$. This implies that $J(f_{\|}) = J^*$. But $f_{\|}$ is in $\mathrm{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}$, and thus can be represented as (1). Furthermore, if Ω is strictly increasing, then $\|f_{\perp}\|_{\mathcal{H}}^2$ must be zero in order to minimize J, and thus any minimizer f must be in the span, and have a representation in the form (1).

□

3 Constructing Kernels

In this section, we discuss how to combine kernels to create more complicated kernels.

If $k_1, k_2, \ldots$ are kernels, then so are the following:

1. $a_1 k_1 + a_2 k_2$ for $a_1, a_2 \in \mathbb{R}^+ = (0, \infty)$.

   Proof. Letting $(K_1)_{ij} = k_1(x_i, x_j)$, and $(K_2)_{ij} = k_2(x_i, x_j)$ as usual, we have

   $$\begin{aligned} \alpha'(a_1 K_1 + a_2 K_2)\alpha &= a_1 \alpha' K_1 \alpha + a_2 \alpha' K_2 \alpha \\ &\geq 0 \end{aligned}$$

   Where the inequality follows since $\alpha' K_1 \alpha \geq 0, \alpha' K_2 \alpha \geq 0, a_1 \geq 0, a_2 \geq 0$. □

2. $k(u, v) = k_1(u, v) k_2(u, v)$

   K can be written as $\alpha'(K_1 \odot K_2)\alpha$ (where $\odot$ denotes the Hadamard or element-by-element product) It turns out that the Hadamard product of two positive semi-definite matrices is also positive semi-definite, and so k is a kernel.

Hint. To prove this, one utilizes the correspondence between positive semidefinite matrices and covariance matrices of multivariate distributions. Given $n \times n$ positive semidefinite matrices K_1 and K_2, consider independent $n \times 1$ random vectors X and Y with covariance matrices K_1 and K_2 respectively. Can you now create a random vector whose covariance matrix is $K_1 \odot K_2$?

3. $k(u, v) = k_1(g(u), g(v))$ where $g : \mathcal{X} \to \mathcal{X}$

   Since g is a transformation in the domain of k_1, k is just a different kernel on the same domain. We have
   $$k(u, v) = \langle \Phi_g(u), \Phi_g(v) \rangle = \langle \Phi(g(u)), \Phi(g(v)) \rangle.$$

4. $k(u, v) = p(k_1(u, v))$, where $p : \mathbb{R} \to \mathbb{R}$ is a polynomial with positive coefficients.

   This follows directly from 1 and 2, for polynomials with positive real coefficients (a sufficient but not necessary condition).

5. $k(u, v) = g(u)g(v)$ for $g : \mathcal{X} \to \mathbb{R}$ and rank one u, v.

   We have $K = zz'$ where $z = (g(x_1), \ldots, g(x_n))'$. Then we have $\alpha' K \alpha = (\alpha' z)^2 \geq 0$. Thus our kernel matrix with entries $K_{ij} = k(x_i, x_j) = g(x_i)g(x_j)$ is positive semidefinite.

6. $k(u, v) = \lim_{i \to \infty} k_i(u, v)$

   If the limit always exists and convergence is uniform for all u, v.

7. $k(u, v) = \exp(k_1(u, v))$

   Follows from 4 and 6 as
   $$\exp(x) = \lim_{i \to \infty} \left(1 + x + \ldots + \frac{x^i}{i!} \right) .$$

   The same reasoning will hold for $k(u, v) = f(k_1(u, v))$ if $f(x)$ has a Taylor expansion involving positive coefficients only.

8. $k(u, v) = \exp\left(-\frac{\|u-v\|^2}{\sigma^2}\right)$

   We can see this by expanding the quadratic and then rewriting the sums in the exponent to a product of exponentials:
   $$k(u, v) = \exp\left(\frac{-\|u\|^2}{\sigma^2}\right) \exp\left(\frac{-\|v\|^2}{\sigma^2}\right) \exp\left(\frac{2u'v}{\sigma^2}\right)$$

   The first two terms give a kernel by 5, and the third term is a kernel by 1 and 7. Finally, we can combine all of the terms into a kernel by 2.

The Gaussian kernel is an example of a *translation-invariant kernel*.

Definition. A kernel is *translation-invariant* if $k(u, v) = k(u - v)$ (i.e. if the kernel depends only on the value $(u - v)$, but not the values u and v themselves).

As an example, suppose $\mathcal{X} = [-\pi, \pi]$ and that k is continuous, even function (i.e. $k(-\alpha) = k(\alpha)$). In that case we can write k's Fourier expansion
$$k(x) = \sum_{n=0}^{\infty} a_n \cos(nx)$$

where $a_n \geq 0$.

So for the translation invariant kernel $k(u,v) = k(u-v)$, we have

$$\begin{aligned}
k(u,v) &= k(u-v) \\
&= a_0 + \sum_{n=1}^{\infty} a_n \cos(n(u-v)) \\
&= a_0 + \sum_{n=1}^{\infty} a_n (\sin(nu)\sin(nv) + \cos(nu)\cos(nv)) \\
&= \sum_{i=1}^{\infty} \lambda_i \psi_i(u) \psi_i(v)
\end{aligned}$$

Where ψ_i's are elements of the set

$$\{1, \sin(u), \cos(u), \sin(2u), \cos(2u), \ldots\}\ ,$$

which consists of all orthogonal trigonometric functions.