CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 8**

Representer theorem and kernel examples

Lecturer: Peter Bartlett *Scribe: Howard Lei*

1 Representer Theorem

Recall that the SVM optimization problem can be expressed as follows:

$$J(f^*) = \min_{f \in H} J(f)$$

where

$$J(f) = \frac{C}{n} \sum_{i=1}^{n} \text{hingeloss}\left(f(x_i), y_i\right) + ||f||_H^2$$

and H is a Reproducing Kernel Hilbert Space (RKHS).

Theorem 1.1. Fix a kernel k, and let H be the corresponding RKHS. Then, for a function L: $\mathbb{R}^n \to \mathbb{R}$ and non-decreasing Ω: $\mathbb{R} \to \mathbb{R}$, if the SVM optimization problem can be expressed as:

$$J(f^*) = \min_{f \in H} J(f) = \min_{f \in H} \left(L(f(x_1) \ldots f(x_n)) + \Omega\left(||f||_H^2\right)\right)$$

then the solution can be expressed as:

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

Furthermore, if Ω is strictly increasing, then all solutions have this form.

This shows that to solve the SVM optimization problem, we only need to solve for the α_i, which agrees with the solution obtained via the Lagrangian formulation of the problem. Furthermore, our solution lies in the span of the kernels.

Proof.

Suppose we project f onto the subspace:

$$\text{span}\{k(x_i, \cdot) \colon 1 \le i \le n\}$$

obtaining f_s (the component along the subspace) and $f_\perp$ (the component perpendicular to the subspace). We have:

$$f = f_s + f_\perp \Rightarrow ||f||^2 = ||f_s||^2 + ||f_\perp||^2 \ge ||f_s||^2$$

Since Ω is non-decreasing,

$$\Omega(||f||_H^2) \ge \Omega(||f_s||_H^2)$$

implying that $\Omega(\cdots)$ is minimized if f lies in the subspace. Furthermore, since the kernel k has the reproducing property, we have:

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle = \langle f_s, k(x_i, \cdot)\rangle + \langle f_\perp, k(x_i, \cdot)\rangle = \langle f_s, k(x_i, \cdot)\rangle = f_s(x_i)$$

Implying that:

$$L(f(x_1), \ldots, f(x_n)) = L(f_s(x_1), \ldots, f_s(x_n))$$

Hence, $L(\cdots)$ depends only on the component of f lying in the subspace: $\text{span}\{k(x_i, \cdot)\colon 1 \leq i \leq n\}$, and $\Omega(\cdots)$ is minimized if f lies in that subspace. Hence, $J(f)$ is minimized if f lies in that subspace, and we can express the minimizer as:

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

Note that if $\Omega(\cdot)$ is strictly non-decreasing, then $||f_\perp||$ must necessarily be zero for f to be the minimizer of $J(f)$, implying that f^* must necessarily lie in the subspace: $\text{span}\{k(x_i, \cdot)\colon 1 \leq i \leq n\}$.

□

2 Constructing Kernels

In this section, we discuss ways to construct new kernels from previously defined kernels. Suppose k_1 and k_2 are valid (symmetric, positive definite) kernels on $\mathcal{X}$. Then, the following are valid kernels:

1. $k(u, v) = \alpha k_1(u, v) + \beta k_2(u, v)$, for $\alpha, \beta \geq 0$

   Proof.

   Since $\alpha k_1(u, v) = \langle \sqrt{\alpha}\Phi_1(u), \sqrt{\alpha}\Phi_1(v)\rangle$ and $\beta k_2(u, v) = \langle \sqrt{\beta}\Phi_2(u), \sqrt{\beta}\Phi_2(v)\rangle$, then:

$$k(u, v) = \alpha k_1(u, v) + \beta k_2(u, v) \quad (1)$$
$$= \langle \sqrt{\alpha}\Phi_1(u), \sqrt{\alpha}\Phi_1(v)\rangle + \langle \sqrt{\beta}\Phi_2(u), \sqrt{\beta}\Phi_2(v)\rangle \quad (2)$$
$$= \langle [\sqrt{\alpha}\Phi_1(u)\ \sqrt{\beta}\Phi_2(u)], [\sqrt{\alpha}\Phi_1(v)\ \sqrt{\beta}\Phi_2(v)]\rangle \quad (3)$$

   and we see that $k(u, v)$ can be expressed as an inner product

   □

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

   Proof.

   Note that the gram matrix K for k is the Hadamard product (or element-by-element product) of K_1 and K_2 ($K = K_1 \bigodot K_2$). Suppose that K_1 and K_2 are covariance matrices of $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively. Then K is simply the covariance matrix of $(X_1Y_1, \ldots, X_nY_n)$, implying that it is symmetric and positive definite. □

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

   Proof.

   Since f is a transformation in the same domain, k is simply a different kernel in that domain:

$$k(u, v) = k_1(f(u), f(v)) = \langle \Phi(f(u)), \Phi(f(v))\rangle = \langle \Phi_f(u), \Phi_f(v)\rangle$$

   □

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

   PROOF.

   We can express the gram matrix K as the outer product of the vector $\gamma = [g(x_1), \ldots, g(x_n)]'$. Hence, K is symmetric and positive semi-definite with rank 1. (It is positive semi-definite because the non-zero eigenvalue of $\gamma\gamma'$ is the trace of $\gamma\gamma'$ which is the trace of $\gamma'\gamma$ which is simply $\gamma'\gamma$ which is greater than or equal to 0).

   □

5. $k(u, v) = f(k_1(u, v))$, where f is a polynomial with positive coefficients.

   PROOF.

   Since each polynomial term is a product of kernels with a positive coefficient, the proof follows by applying 1 and 2.

   □

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

   PROOF.

   Since:

   $$\exp(x) = \lim_{i\to\infty}\left(1 + x + \cdots + \frac{x_i}{i!}\right)$$

   The proof follows from 5 and the fact that:

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

   □

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

   PROOF.

   $$k(u, v) = \exp\left(\frac{-||u-v||^2}{\sigma^2}\right) = \exp\left(\frac{-||u||^2 - ||v||^2 + 2u'v}{\sigma^2}\right) \tag{4}$$

   $$= \left(\exp\left(\frac{-||u||^2}{\sigma^2}\right)\exp\left(\frac{-||v||^2}{\sigma^2}\right)\right)\exp\left(\frac{2u'v}{\sigma^2}\right) \tag{5}$$

   $$= (g(u)g(v))\exp(k_1(u, v)) \tag{6}$$

   $g(u)g(v)$ is a kernel according to 4, and $\exp(k_1(u, v))$ is a kernel according to 6. According to 2, the product of two kernels is a valid kernel.

   □

Note that the Gaussian kernel is translation-invariant, where $k(u, v)$ can be expressed as $f(u - v) = f(x)$.

Example: Translation-invariant kernels

Consider the function f: $[-\pi, \pi] \to \mathbb{R}$, and suppose that f is continuous and even (i.e. $f(x) = f(-x)$). Then, we can express f via the Fourier expansion as:

$$f(x) = \sum_{n=0}^{\infty} a_n \cos(nx)$$

where $a_n \geq 0$.

If we let x be the difference of u and v, then we have:

$$f(x) = f(u - v) = a_0 + \sum_{n=1}^{\infty} a_n(\sin(nu)\sin(nv) + \cos(nu)\cos(nv)) \tag{7}$$

$$= \sum_{i=0}^{\infty} \lambda_i \Psi_i(u)\Psi_i(v), \tag{8}$$

where $\{\Psi_i\} = \{\sin(nu) : n \geq 1\} \cup \{\cos(nu) : n \geq 0\}$.

We see that $f(u - v)$ is a valid kernel that's translation invariant. This example shows that we can choose the kernel by choosing the a_i coefficients, which is equivalent to choosing a filter.

Example: Bag-of-words kernel

Suppose that $\Phi_w(d)$ is the number of times word w appears in document d. If we want to classify documents by their word counts, we can use the kernel $k(d_1, d_2) = \langle \Phi(d_1), \Phi(d_2) \rangle$. (In practice, these counts are weighted to take into account the relative frequency of different words.)

Example: Marginalized kernel

Given the probability distribution $p(x, h)$ (and hence $p(h|x)$) and a kernel defined for (x,h) pairs ($k((x, h), (x', h'))$), we can obtain a kernel on only the x's as follows:

$$k_m(x, x') = \sum_{h,h'} k((x, h), (x', h'))p(h|x)p(h'|x')$$

Exercise: Prove that this is a valid kernel!

Example: Convolution kernel (or "string" kernel)

Define a_i to be a letter of the alphabet, $s = (s_i, \ldots, s_\ell)$ to be a string of letters, and Σ^* to be the space of all possible letter sequences.

Suppose that s has $a = (a_1, \ldots, a_n)$ as a subsequence if there exists a sequence of indices $I = (i_1, \ldots, i_n)$, where $i_1 < i_2 < \cdots < i_n$ with $s_{i_j} = a_j$, where $j = 1, \ldots, n$. Define the length of the set of indices $(i_1, \ldots, i_n)$ forming the subsequence as $\ell(I) = i_n - i_1 + 1$. For simplicity, we use the notation $s[I] = a$.

Define, for fixed n, the feature map for a particular sequence a and string s:

$$\Phi_a(s) = \sum_{I:\ s[I]=a} \lambda^{\ell(I)}$$

where $\lambda \in (0, 1)$. To compare two strings s and s', we can use the following kernel:

$$k(s, s') = \sum_{a \in \Sigma^n} \Phi_a(s)\Phi_a(s')$$

We can also derive the above kernel via convolution. Define the following kernel:

$$k_0((s,i),(s',i')) = 1[s(i) = s'(i')]$$

Set

$$k_n((s,i),(s',i')) = k_0((s,i),(s',i'))(h * k_{n-1})((s,i),(s',i'))$$

where $h(i-j) = 1[i-j>0]\lambda^{-(i-j)}$, and $*$ is the convolution operator. Then:

$$(h * k_{n-1})((s,i),(s',i')) = \sum_{j,j'} h(i-j)h(i'-j')k_{n-1}((s,i),(s',i'))$$

and

$$k(s,s') = \sum_{i,i'} k_n((s,i),(s',i'))$$