CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Reproducing Kernel Hilbert Spaces II

Lecturer: Michael I. Jordan *Scribe: Nemanja Isailovic*

1 The "kernel trick" from the RKHS point of view

Given a kernel $k(x, x')$, define an RKHS:

$$\mathcal{H} = \left\{ f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x') \right\}$$

with the following dot product:

$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j k(x_i, x_j),$$

which implies the reproducing property:

$$\begin{aligned} \langle k(\cdot, x), f \rangle &= f(x) \\ \langle k(\cdot, x), k(\cdot, x') \rangle &= k(x, x'). \end{aligned}$$

We can now interpret the "kernel trick" using the RKHS formalism. Recall the reproducing kernel map:

$$\Phi : x \longrightarrow k(\cdot, x).$$

which assigns to each x a kernel function $k(\cdot, x)$. From the reproducing property, we have:

$$\langle \Phi(x), \Phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$$

which is nothing but the "kernel trick"!

2 Example: a RKHS over $(0, 2\pi)$

Start with a symmetric and continuous kernel $k(x)$:

$$k(x) = \sum_{n=0}^{\infty} \lambda_n \cos nx,$$

and define a translation-invariant kernel:

$$k(x, x') = k(x - x') = 1 + \sum_{n=1}^{\infty} \lambda_n \sin nx \sin nx' + \sum_{n=1}^{\infty} \lambda_n \cos nx \cos nx'$$

Define: $\{\psi_n(x)\} = (1, \sin x, \cos x, \sin 2x, \cos 2x, ...)$.

Define $\mathcal{H}$ to be the set of linear combinations of $\{\psi_n(x)\}$.

Given $f(x)$ and $g(x)$ in $\mathcal{H}$, we can calculate the Fourier coefficients:

$$\begin{aligned} f_n^c &= \langle f, \cos nx \rangle \\ f_n^s &= \langle f, \sin nx \rangle \\ g_n^c &= \langle g, \cos nx \rangle \\ g_n^s &= \langle g, \sin nx \rangle \end{aligned}$$

which implies:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{n=0}^{\infty} (f_n^c g_n^c + f_n^s g_n^s)/\lambda_n$$

and also:

$$||f||_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{n=0}^{\infty} ((f_n^c)^2 + (f_n^s)^2)/\lambda_n$$

Recall:

$$\sum_n |\lambda_n| < \infty$$

for any Mercer kernel. This implies $\lambda_n \to 0$ as $n \to \infty$. Thus, in order for the norm of f to be finite ($||f||_{\mathcal{H}} < \infty$), we need both f_n^c and f_n^s to approach 0 fast (as $n \to \infty$). That is, the numerator in the above expression must approach 0 faster than the denominator.

3 RKHS norm of a support vector expansion

Recall that in the Support Vector Machine (SVM):

$$f(\cdot) = \langle w, x \rangle = \sum_{i=1}^{m} \alpha_i y_i k(\cdot, x_i)$$

But since the kernels $k(\cdot, x_i)$ span our space and $f(x)$ is a linear combination of the kernels $k(\cdot, x_i)$, we can conclude that $f(x) \in \mathcal{H}$. Moreover:

$$\begin{aligned} ||f||_{\mathcal{H}}^2 &= \langle \sum_i \alpha_i y_i k(\cdot, x_i), \sum_j \alpha_j y_j k(\cdot, x_j) \rangle_{\mathcal{H}} \\ &= \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle k(\cdot, x_i), k(\cdot, x_j) \rangle_{\mathcal{H}} \\ &= \sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j) \end{aligned}$$

This is exactly the term we're minimizing in the dual of SVM. So, from our new perspective, we're minimizing the norm of f in an RKHS. Indeed, from $w = \sum_i \alpha_i y_i \Phi(x_i)$, we have:

$$\begin{aligned} ||w||^2 &= w^T w = \sum_i \sum_j \alpha_i \alpha_j y_i y_j \Phi(x_i)^T \Phi(x_i) \\ &= \sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ &= ||f||_{\mathcal{H}}^2 \end{aligned}$$

Thus our Primal Problem, which is to minimize $w^T w$ subject to constraints, is equivalent to minimizing $||f||^2_{\mathcal{H}}$ subject to constraints.

4 Pointwise loss functions and the Representer Theorem

At each point x_i, we wish to measure the difference between $f(x_i)$ and the observed value y_i.

Example: SVM Regression

$$c((x_1, y_1, f(x_1)), ..., (x_m, y_m, f(x_m))) = (1/m) \sum_{i=1}^{m} |y_i - f(x_i)|_\epsilon$$

No "loss" (cost) near the observed value, then a linearly increasing loss beyond some value ϵ.

Example: SVM Classification

$$c((x_1, y_1, f(x_1)), ..., (x_m, y_m, f(x_m))) = (1/m) \sum_{i=1}^{m} \max(0, 1 - y_i f(x_i)) = \sum_i \xi_i$$

In general, our primal problem is a minimization of the form:

$$P : \min[c((x_1, y_1, f(x_1)), ..., (x_m, y_m, f(x_m))) + \Omega(||f||_{\mathcal{H}})].$$

The first term is the loss function, while the second term is a regularization term that smooths the result and avoids overfitting. (Side note: The Ω term that we've always used so far is $w^T w$.)

Representer Theorem (Kimeldorf and Wahba, 1971):

Each minimizer of P admits a representation of the form:

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

i.e. f is a sum of the kernel functions evaluated at the data points x_i. We'll prove this result in the next class.