CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Regularization and RKHS

Lecturer: Michael I. Jordan *Scribe: Barbara Rosario*

1 Introduction

The goal is to find a function f that minimizes the Risk $R_{reg}(f)$:

$$\min_f R_{reg}(f) = \min_f R_{emp}(f) + \lambda\Omega(f) \tag{1}$$

where $R_{emp}(f)$ is the empirical risk that depends on the data, $R_{emp}(f) = c((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n)))$ and $\Omega(f)$ is a measure of the smoothness of f.

There are several ways to achieve smoothness:

- $\Omega(f)$ penalizes big values of f and its derivatives
- $\Omega(f)$ penalizes high frequencies
- f is a sum of small numbers of simple (smooth) basis functions
- small coefficients in sum expansion.

We will analyze each of these methods in detail.

2 Penalizing derivatives

This is a common way to achieve smoothness:

$$\min_f R_{reg}(f) = \min_f R_{emp}(f) + \lambda\|Lf\|^2 \tag{2}$$

where the norm $\|\cdot\|$ is the L_2 norm and L is a differential operator (for example, $L = \frac{\partial}{\partial x}$ or $L = \frac{\partial^2}{\partial^2 x}$).

It turns out that there is a relationship between regularization and Reproducing Kernel Hilbert Space (RKHS): the reproducing Kernel k is obtained as the Green's function of the operator L^*L [1]. This is a very nice result, in that it ties together things that seem very different, regularization and RKHS. It shows that in reality *RKHS and regularization are equivalent*.

[1] L^* is the adjoint of L: $\langle Lx, y\rangle = \langle x, L^*y\rangle$

3 Penalizing high frequency

We can also express regularization in terms of a penalty in the frequency domain:

$$R_{reg}(f) = R_{emp}(f) + \lambda \frac{1}{2\pi} \int \frac{|\hat{f}(\omega)|^2}{V(\omega)} d\omega \tag{3}$$

where $\hat{f}(\omega)$ is the Fourier transform of f

$$\hat{f}(\omega) = \int f(x) e^{-i\omega x} dx \tag{4}$$

We'll return later to an explanation of how this relates to the earlier version of regularization, and to RKHS ideas.

4 Green's functions and Partial Differential Equations (PDEs)

Let u be an arbitrary function of any number of variables and L a partial differential operator. The problem if to find u given L and F and the equation $Lu = f$. One way of solving this is with Green's functions.[2]

We want to find the inverse of L, L^{-1}. Let us assume that L^{-1} is an integral operator based on a kernel (matrix), i.e. let's assume that for some k :

$$(L^{-1}v)(x) = \int k(x, y) v(y) dy$$

We redefine our problem as the problem of finding a k for which this is true. We want

$$LL^{-1}u = u$$

$$L \int k(x, y) u(y) dy = u(x)$$

$$\int (Lk(x, y)) u(y) dy = u(x)^3$$

We can therefore define the problem as follows: find $k(x, y)$ such that $Lk(x, y) = \delta(x - y)$. *Such a function $k(x, y)$ is called a Green's function for L.*[4]

If we can find the the Green's function (GF hereafter), we can solve the PDE $Lu = f$

$$u = L^{-1} f$$

$$u(x) = (L^{-1} f)(x) = \int k(x, y) f(y) dy$$

To verify this:

$$Lu(x) = \int Lk(x, y) f(y) dy = \int \delta(x - y) f(y) dy = f(x)$$

[2]For example, if $L = \frac{\partial}{\partial x}$ and $\frac{\partial u(x)}{\partial x} = f(x)$ we'll need to integrate to find $u(x)$.
[3]L is linear and we can move it inside the integral.
[4]This is the definition of a Green's function.

5 Relation to RKHS

Define a Kernel $k(\cdot, x)$ as a GF for L^*L:

$$\|Lf\|^2 = \langle Lf, Lf \rangle$$

For any f:

$$\langle Lf(t), Lk(t,x) \rangle = \langle f(t), L^*Lk(t,x) \rangle = \langle f(t), \delta(t-x) \rangle = f(x)$$

Note that the result obtained:

$$\langle f(t), L^*Lk(t,x) \rangle = f(x) \tag{5}$$

implies the reproducing property. In particular:

$$\langle Lk(\cdot, x), Lk(\cdot, x) \rangle = k(x, x') \tag{6}$$

which is the reproducing property.

Recall that the Hilbert space H is the space of all functions f such that

$$\|Lf\| < \infty$$

This implies that the solution f to (2) is in the Hilbert space H and

$$\lambda \|Lf\|^2 = \lambda \|f\|^2_{\mathcal{H}} \tag{7}$$

That is, eq. (2) is equivalent to minimizing

$$\min_f R_{reg}(f) = \min_f R_{emp}(f) + \lambda \|f\|^2_{\mathcal{H}} \tag{8}$$

$$\Longrightarrow f(x) = \sum_i \alpha_i k(x, x_i) \tag{9}$$

from the representer theorem.

We started from eq. (2) and we proved that f is not *any* function in L_2 but is given by Eq. (9): the regularization (given by the term $\lambda\|Lf\|^2$) constrains f to be of the form given by (9), i.e. f is in a *finite* dimensional space.

Note that in Eq. (2) we say nothing about kernel functions. The equivalence between regularization and RKHS implies that we're using a kernel function, however, and thus we obtain (9). The problem reduces to finding the α.

We can also go backwards, i.e. we can choose a kernel k and derive L. For example if we choose the Gaussian for the kernel, L is an infinite sum of derivatives: the Gaussian is a really smooth function that penalizes *all* derivatives.