Lecture 2 [Introduction to Learning]

In the following, we present a brief introduction to the theory of learning of a particular variety, i.e. learning with kernels.

Given a data sequence $\{(x_i, y_i)\}_{i=1}^{m}$ consisting of input/output pairs (or stimulus/response pairs, in the language of biology), where $x_i \in X \overset{\text{closed}}{\subset} \mathbb{R}^n$ and $y_i \in Y = \mathbb{R}^1$ for the time being, the idea is to build/learn a model

$$f : X \to Y$$

where f accounts for the data by approximating it well in a certain sense, AND has predictive/generalization power. It is the latter requirement that leads us to a key consideration: how complicated should f be? It turns out that high complexity, measured suitably, leads to poor predictive ability.

Thus the process of building a model from data is a process of trading off predictive ability against approximation error. This is encoded as an optimization problem on a suitable space of candidate models, known as the hypothesis space. The very definition of hypothesis space will be based on the concept of a kernel and an associated Hilbert space known as reproducing kernel Hilbert space, or RKHS for short.

A Hilbert space V is a vector space with an inner product $\langle \cdot, \cdot \rangle$ (for us usually of the strict positive definite kind) and satisfying the completeness property:

Every sequence $\{v_n\}_{n=1}^{\infty} \subset V$ satisfying the Cauchy condition

$$\lim_{m,n \to \infty} \|v_n - v_m\| = 0$$

is also convergent to a $v = \lim_{n \to \infty} v_n \in V$

The norm is defined by

$$\|v\| = \sqrt{\langle v, v \rangle}$$

($\|v\| \geq 0$ and $\|v\| = 0 \leftrightarrow v = 0$, by strict positive definiteness)

RKHS Hilbert spaces are function spaces.

We begin with an elementary setting.

(i) Suppose $X = \{1, 2, \ldots, L\}$

for each $f : X \to \mathbb{R}$, we have a row vector $(f(1), f(2), \ldots, f(L)) \in \mathbb{R}^L$ and conversely. Thus the function space on X is identified with $\mathbb{R}^L$.

A kernel K on X is a function

$$K : X \times X \to \mathbb{R}$$
$$(i, j) \mapsto K(i, j)$$

satisfying

(i) $K(i, j) = K(j, i)$ symmetry

(ii) $\mathbb{K} = [K(i, j)]$ is a positive definite* matrix with L rows and L columns.

---

* would be same as positive semidefiniteness in linear algebra

Define the set of functions

$$K_i : X \to \mathbb{R}$$

$$j \mapsto K_i(j) = K(i,j)$$

Thus K_i is the i^{th} row of the matrix $\mathbb{K}$. Such functions, through all possible linear combinations, generate the hypothesis space

$$H_{\mathbb{K}} = \left\{ f : X \to \mathbb{R} \;\middle|\; f = \sum_{i=1}^{L} a_i K_i \right\}$$
$$\forall\, a_i \in \mathbb{R}$$

It is clearly a vector space, of dimension L since the rows of $\mathbb{K}$ are linearly independent by the assumption of positive definiteness.

We give $H_{\mathbb{K}}$ an inner product $\langle \cdot, \cdot \rangle_K$ by first defining it on the functions K_i to be

$$\langle K_i, K_j \rangle_K = K(i,j)$$

and extending to all of H_K by linearity.

kernel

H_K is a Hilbert space (of dimension L);
here completeness follows from the completeness of the real line $\mathbb{R}$.

H_K is a reproducing kernel Hilbert space in the sense that

$$\langle f, K_i \rangle_K = f(i)$$

proof $\quad f \in H_K \iff f = \sum_{j=1}^{L} a_j K_j$

for some vector $(a_1, a_2, \ldots, a_L)$.

$$f(i) = \left(\sum_{j=1}^{L} a_j K_j \right)(i)$$

$$= \sum_{j=1}^{L} a_j K_j(i)$$

$$= \sum_{j=1}^{L} a_j K(j, i) .$$

On the other hand

$$\langle f, K_i \rangle_K = \left\langle \sum_{j=1}^{L} a_j K_j , K_i \right\rangle_K$$

$$= \sum_{j=1}^{L} a_j \langle K_j, K_i \rangle_K$$

$$= \sum_{j=1}^{L} a_j \, K(j, i)$$

Thus $f(i) = \langle f, K_i \rangle_K$ ▨

(ii) More generally, let $X \overset{\text{closed}}{\subset} \mathbb{R}^n$

and let $K : X \times X \to \mathbb{R}$

be a symmetric, positive definite function

i.e.

$$K(x, x') = K(x', x) \qquad x, x' \in X$$

and, for any choice of $c_1, \ldots, c_m \in \mathbb{R}$

and, for any choice of $\tilde{x}_1, \ldots, \tilde{x}_m \in X$

and for any $m \in \{1, 2, 3, \ldots\}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \, K(\tilde{x}_i, \tilde{x}_j) \geq 0 \, .$$

For any $\tilde{x}_i \in X$ define

$$K_{\tilde{x}_i} : X \longrightarrow \mathbb{R}$$

$$x \longmapsto K_{\tilde{x}_i}(x) = K(\tilde{x}_i, x)$$

Define

$$H_K^{pre} = \left\{ f : X \to \mathbb{R} \;\middle|\; f = \sum_{i=1}^{M} \alpha_i K_{\tilde{x}_i} \right.$$

$$\left. \forall\, \tilde{x}_i \in X, \quad \forall\, \alpha_i \in \mathbb{R}, \quad \forall\, M \in \{1, 2, \dots\} \right\}$$

.

H_K^{pre} is a vector space.

It inherits an inner product $\langle \cdot, \cdot \rangle_K$

from the definition

$$\langle K_{\tilde{x}_i}, K_{\tilde{x}_j} \rangle_K = K(\tilde{x}_i, \tilde{x}_j)$$

and extension by linearity. The

inner product on H_K^{pre} gives rise to a ~~can be~~ completed

~~to a~~ norm

$$\| f \|_K = \sqrt{\langle f, f \rangle_K}$$

Then

H_K = Completion of H_K^{pre}

in the above norm

Again H_K^{pre} is a reproducing kernel space:

$$\langle f, K_x \rangle_K = f(x)$$

and the reproducing property extends to H_K.

We are now ready to discuss the modeling problem in H_K. Given a set of input-output data $\{(x_i, y_i) : i = 1, 2, \ldots, m\}$ we seek $f : H_K \to \mathbb{R}$ such that

$$C(f) = \frac{1}{m} \sum_{i=1}^{m} V(y_i, f(x_i)) + \gamma \|f\|_K^2$$

is minimized. Here V denotes a loss function, e.g., $V(a, b) = (a-b)^2$, that measures fit error, and with $\gamma > 0$, the term $\gamma \|f\|_K^2$ is a measure of complexity of f.

The minimization problem can be attacked by first looking for a first order necessary condition. Clearly, if f is a minimum, then

$$\frac{d}{d\varepsilon} C(f + \varepsilon \bar{f})\Big|_{\varepsilon=0} = 0 \qquad \forall\ \bar{f} \in H_K$$

Compute

$$\frac{d}{d\varepsilon} C(f + \varepsilon \bar{f})\Big|_{\varepsilon=0}$$

$$= \frac{d}{d\varepsilon}\left\{ \frac{1}{m} \sum_{i=1}^{m} V(y_i, f(x_i) + \varepsilon \bar{f}(x_i)) + \gamma \langle f + \varepsilon \bar{f}, f + \varepsilon \bar{f} \rangle_K \right\}\Big|_{\varepsilon=0}$$

$$= \frac{1}{m} \sum_{i=1}^{m} D_2 V(y_i, f(x_i))\, \bar{f}(x_i) + 2\gamma \langle f, \bar{f} \rangle_K$$

(here $D_2 V$ = partial derivative of V w.r.t. its second argument)

$$= 0 \qquad \forall\ \bar{f} \in H_K.$$

Let $\bar{f} = K_x$. Then $\bar{f}(x_i) = K_x(x_i)$

$= K_{x_i}(x)$. Thus we have the

necessary condition:

$$0 = \frac{1}{m} \sum_{i=1}^{m} D_2 V(y_i, f(x_i)) K_{x_i}(x)$$

$$+ \ 2\gamma \langle f, K_x \rangle_K$$

$$= \frac{1}{m} \sum_{i=1}^{m} D_2 V(y_i, f(x_i)) K_{x_i}(x)$$

$$+ \ 2\gamma f(x) \qquad \text{(reproducing property)}$$

Hence $$f = -\frac{1}{2\gamma m} \sum_{i=1}^{m} D_2 V(y_i, f(x_i)) K_{x_i}.$$

This is known as the <u>representer theorem</u>, since it tells us the form of an extremal f.

Let $c_i = -\frac{1}{2\gamma m} D_2 V(y_i, f(x_i))$.

$$= -\frac{1}{2\gamma m} D_2 V\left(y_i, \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i)\right)$$

(by representer theorem)

$$= -\frac{1}{2\gamma m} D_2 V\left(y_i, \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i)\right)$$

This is a (nonlinear) equation for the unknown c_i. If $V(a,b) = (a-b)^2$

Then $D_2 V(a,b) = -2(a-b)$. In that case

$$c_i = -\frac{1}{2\gamma m}(-2)\left(y_i - \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i)\right)$$

$$\leftrightarrow \quad \gamma m c_i + \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i) = y_i$$

$$\leftrightarrow \quad (\gamma m \mathbb{1} + \mathbb{K})\, \underline{c} = \underline{y}$$

where we have used $K_{x_j}(x_i) = K(x_j)_{x_i}$

and let $\underline{y} = (y_1, \ldots, y_m)^T$; $\underline{c} = (c_1, \ldots, c_m)^T$

and $\mathbb{1}$ denotes the $m \times m$ identity matrix. Since $\mathbb{K}$ is positive definite, it follows from $\gamma_m > 0$ that $(\gamma_m \mathbb{1} + K)$ is strictly positive definite and hence invertible leading to unique c (minimum).

For general loss functions V, the coupled nonlinear equations for c_i will typically have multiple solutions.

Reading Assignment

1. T. Poggio and S. Smale (2003). The mathematics of learning: dealing with data, Notices of the AMS, Vol 50 (5), pp 537-544.

2. S. Simic (2003). A learning theory approach to sensor networks. Pervasive Computing, October-December 2003, pp 44-49.