2017.09.13

ENEE 765 Lecture 4 (kernel method)

In discussing identification of systems, we spoke of fitting linear models (e.g. for indirect adaptive control). More generally one might speak of learning a (nonlinear) map $f : X \to Y$ from a set X of stimuli/input to response/output set Y, using an empirical data sequence $\{(x_i, y_i) : i = 1, 2, \cdots, m\}$. The index set $\{i = 1, 2, \cdots, m\}$ need not represent time instants. We refer to this setting as supervised learning or "learning with a teacher". One widely used approach to this task is outlined here and is known as kernel-based learning.

Some desiderata:

(i) Learned model $\hat{f}$ should account for the data by approximating it well in some sense.

(ii) Learned model $\hat{f}$ should prove effective in making generalizations/predictions i.e. error $(y - \hat{f}(x))$ between prediction $\hat{f}(x)$ and observed response y to a future stimulus x should be small.

(iii) For the most part, $X \subset \mathbb{R}^n$ is a closed set and $Y = \mathbb{R}^1$.

Various formulations of learning show that these are competing requirements. Models of high complexity that fit the data well pay a price in generalization performance. This trade-off is encoded as an optimization problem on a suitable space of candidate models, known as hypothesis space. The very definition of this space is based on the concept of a kernel and associated Hilbert space known as reproducing kernel Hilbert space (RKHS). Here we go with mathematical details:

A Hilbert space V over reals $\mathbb{R}$ is a vector space with a positive definite inner product $\langle \cdot, \cdot \rangle$, and associated norm satisfying completeness property. Thus

$$\langle \cdot , \cdot \rangle : V \times V \to \mathbb{R}$$
$$(v, w) \mapsto \langle v, w \rangle$$

Satisfies

(i) $\langle v, w \rangle = \langle w, v \rangle$

(ii) $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle \qquad \alpha \in \mathbb{R}$

(iii) $\langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle$

(iv) $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0 \Rightarrow v = 0$

Norm $\|v\| = (\langle v, v \rangle)^{1/2}$ defines a metric

$$d(v, w) = \|v - w\|$$

We say a sequence $\{v_n : n = 1, 2, 3, \cdots\} \subset V$ __converges__ to $v \in V$ if

$$\lim_{n \to \infty} \| v_n - v \| = 0 .$$

We say a sequence $\{v_n : n = 1, 2, 3, \cdots\} \subset V$ is a __Cauchy__ sequence if

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

It is easy to see that __every__ convergent sequence is a Cauchy sequence. In general the converse is not true. We say that $(V, \| \cdot \|)$ is a __complete__ normed linear space if __every__ Cauchy sequence is also a convergent sequence.

EXAMPLE Suppose $X = \{1, 2, \ldots, L\} \subset \mathbb{R}^1$, a discrete set of stimuli/inputs. Each $f : X \to Y = \mathbb{R}^1$ defines a row vector $(f(1), f(2), \cdots, f(L)) \in \mathbb{R}^L$. A kernel K is simply 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) For any real c_i, $i = 1, 2, \ldots, L$

$$\sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j K(i, j) \geq 0$$

Condition (ii) above is often referred to as <u>positive definiteness</u> of the kernel function, but in linear algebra this would correspond to <u>positive semidefiniteness</u> of the matrix

$$\mathbb{K} = [K(i,j)]$$

with L rows and L columns.

Define the set of functions

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

Again each such K_i defines a row vector $(K_i(1), K_i(2), \cdots, K_i(L)) \in \mathbb{R}^L$. Thus it makes sense to look for a model $f: X \to \mathbb{R}$ in the hypothesis space

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

Clearly this is at most L dimensional. In fact for any $f \in H_{\mathbb{K}}$, associated row vector

$$\big(f(1)\ f(2)\ \cdots\ f(L)\big) = (a_1\ a_2\ \cdots\ a_L) \begin{pmatrix} K_1(1) & K_1(2) & \cdots & K_1(L) \\ K_2(1) & K_2(2) & \cdots & K_2(L) \\ K_L(1) & K_L(2) & \cdots & K_L(L) \end{pmatrix}$$

or in short hand, the row vector

$$f = a\,\mathbb{K}$$

Kernel matrix

where $a = (a_1, a_2, \cdots, a_L)$ and $\text{dimension}(H_{\mathbb{K}}) = \text{dimension}(\text{range}(\mathbb{K}))$.

On $H_{\mathbb{K}}$ define inner product candidate

$$\langle f, g \rangle_{\mathbb{K}} = \left\langle \sum_{i=1}^{L} a_i k_i \, , \, \sum_{j=1}^{L} b_j k_j \right\rangle$$

$$= \sum_{i=1}^{L} \sum_{j=1}^{L} a_i b_j \langle k_i, k_j \rangle_{\mathbb{K}}$$

where $\langle k_i, k_j \rangle_{\mathbb{K}} \triangleq K(i, j)$.

By positive definiteness of the Kernel K, (hence positive semidefiniteness of the matrix $\mathbb{K}$),

$$\langle f, f \rangle_{\mathbb{K}} = \sum_{i=1}^{L} \sum_{j=1}^{L} a_i a_j K(i, j)$$

$$\geq 0$$

The r.h.s above can be written as $a \mathbb{K} a^T$ where a is a row vector and a^T is associated column vector. Since $\mathbb{K}$ is positive semidefinite

it can be factorized as $\mathbb{K} = N N^T$.
Then

$$\langle f, f \rangle_{\mathbb{K}} = 0 \iff a \mathbb{K} a^T = 0$$

$$\iff a N N^T a^T = 0$$

$$\iff a N = 0$$

$$\Rightarrow a N N^T = 0$$

$$\Rightarrow a \mathbb{K} = 0$$

$$\Rightarrow f = 0 .$$

Thus $\langle \cdot, \cdot \rangle_{\mathbb{K}}$ is a genuine positive definite inner product.

It follows that $(H_{\mathbb{K}}, \langle \cdot, \cdot \rangle_{\mathbb{K}})$ is Hilbert space of dimension = rank $(\mathbb{K})$ $\leq L$. (Here completeness follows from completeness of the real line and hence of $\mathbb{R}^k$ for any positive integer k.)

$H_{\mathbb{K}}$ is a reproducing kernel Hilbert space in the sense that for any $f \in H_{\mathbb{K}}$,

$$f(i) = \langle f, K_i \rangle_{\mathbb{K}}$$

<u>proof</u> $f = \sum_{j=1}^{L} a_j K_j$ for some row vector

$a = (a_1, a_2, \cdots, a_L)$

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

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

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

$$\langle f, K_i \cdot \rangle_{\mathbb{K}} = \left\langle \sum a_j K_j \cdot \, , K_i \cdot \right\rangle_{\mathbb{K}}$$

$$= \sum_{j=1}^{L} a_j \langle K_j, K_i \cdot \rangle_{\mathbb{K}}$$

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

Hence $f(i) = \langle f, K_i \cdot \rangle_{\mathbb{K}}$

for $i = 1, 2, \ldots, L$. $\boxtimes$

We can proceed from the $\boxed{\text{example}}$ above of a discrete, finite set of stimuli/inputs to the more general setting

$$X \overset{\text{closed}}{\subset} \mathbb{R}^n$$

possibly a continuum

and kernel

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

satisfying

(symmetry) $$K(x, x') = K(x', x) \quad , x, x' \in X$$

and, for any choice of $c_1, c_2, \cdots, c_m \in \mathbb{R}$,
and, for any choice of $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_m \in X$
and, for any positive integer m,

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

→ For now we are SILENT on continuity of K etc

Associated to such a positive definite kernel K, we define kernel functions, $K_{\tilde{x}}$ for any $\tilde{x} \in X$ by letting

$$K_{\tilde{x}} : X \to \mathbb{R}$$

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

Notice that when X is a continuum, the family of kernel functions is uncountably infinite.

Define

$$H_k^{pre} = \left\{ f : X \to \mathbb{R} \;\middle|\; f = \sum_{i=1}^{m} a_i K_{\tilde{x}_i}, \quad \begin{array}{l} m \text{ any positive integer} \\ \forall\, a_i \in \mathbb{R},\ \forall\, \tilde{x}_i \in X \end{array} \right\}$$

H_K^{pre} is a (in general infinite dimensional) vector space.

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

and extend this by linearity to ~~an~~ a positive definite inner product on all of H_K^{pre}.

Next, define $\|f\|_K = \langle f, f \rangle_K^{1/2}$ as norm on H_K^{pre}. Then, our hypothesis space

$$H_K = \text{Completion of } H_K^{pre}$$

(hypotheses about K such as continuity matter here)

in the above norm, and one can verify that the inner product on H_K^{pre} extends to one on H_K, satisfying the reproducing property,

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

We are now ready to discuss the modeling problem:

Given a finite set of input-output data $\{(x_i, y_i) : i = 1, 2, \ldots, m\} \subset X \times Y$, find $f \in H_K$

~~$f : H_K \to \mathbb{R}$~~ ~~$f \in H_K$ such that~~

minimizing

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

Here Φ is a loss function measuring fit error, e.g. $\Phi(a,b) = (a-b)^2$. $\|f\|_K^2$ is a measure of complexity of f.

The constant $\gamma > 0$ is chosen to reflect the relative importance given to the two terms in the cost function C.

Hypotheses about Φ will dictate learning context

$$\boxed{\begin{array}{l} \text{If } f \text{ minimizes } C \text{ then} \\ \left.\dfrac{d}{d\varepsilon} C(f + \varepsilon \bar{f})\right|_{\varepsilon=0} = 0 \quad \forall\ \bar{f} \in H_K \end{array}}$$

(1st order necessary cond.)

Compute:

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

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

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

here $D_2 \Phi$ = partial derivative of Φ w.r.t second argument

Set this $= 0$.

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

Thus we have the 1^{st} order necessary condition,

$$0 = \frac{1}{m} \sum_{i=1}^{m} D_2 \Phi(y_i, f(x_i)) K_{x_i}(x) + 2\gamma \langle f, K_x \rangle_K$$

$$= \frac{1}{m} \sum_{i=1}^{m} D_2 \Phi(y_i, f(x_i)) K_{x_i}(x) + 2\gamma f(x)$$

(by reproducing property)

Hence

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

This is known as the <u>representer theorem</u>, since it says that optimal f is necessarily a linear combination of kernel functions with coefficients

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

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

$i = 1, 2, \ldots, m.$

This is a system of equations for the unknown coefficients.

Suppose $\Phi(a, b) = (a-b)^2$. Then $D_2 \Phi(a, b) = -2(a-b)$. Hence

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

$$\Longleftrightarrow \quad \gamma m c_i + \sum_{j=1}^{m} c_j K_{x_j}(x_i) = y_i \quad i = 1, 2, \ldots, m$$

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

where we have defined $\mathbb{K} = [K_{x_j}(x_i)]$

$$= [K(x_i, x_j)],$$

$c = (c_1, \ldots, c_m)^T$, $y = (y_1, \ldots, y_m)^T$ (column vectors)

The matrix

$$\gamma m \mathbb{1} + \mathbb{K}$$

is invertible since $\gamma m > 0$ and $\mathbb{K}$ is positive semi-

-definite, with $\mathbb{1}$ denoting the identity matrix. Hence we can solve for unique c to determine the minimizer of $\mathcal{C}$. For general loss functions Φ, the system of equations for c_i, $i = 1, 2, \cdots, m$

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

would be nonlinear and possibly admit multiple solutions.