RKHS

Mu Niu

December 17, 2014

1 Approximation in Reproducing Kernel Hilbert Spaces

we will derive the expression of the best predictor. A brief introduction can be found in Durrande et al. (2013)

Let $\mathcal{H}$ be a Hilbert space of real valued functions defined over D $\subset \mathbb{R}.\mathcal{H}$ is said to be a RKHS if and only if there exist a function $k(.,.) : D \times D \to \mathbb{R}$ such that for all $x \in D$

- $k(x,.) \in \mathcal{H}$
- $\exists f \in \mathcal{H}\ f(x) =< f(.), k(x,.) >_{\mathcal{H}}$

The function k satisfying these properties is unique and it is the reproducing kernel of $\mathcal{H}$. RKHS is completion of

$$\{\sum_{i=1}^{n} a_i k(x_i,.); n \in \mathbb{N}, a_i \in \mathbb{R}, x_i \in D\} \tag{1}$$

In other words, the element in RKHS can be represented as linear combination of $k(x_i,.)$, but n need to be ∞.

we will show how to approximate a function f that is observed in a finite number of points. Let $X = x_1, ..., x_n$ be a set of points where the value $y_i = f(x_i)$ is known and y be the vector of y_i. For a given RKHS $\mathcal{H}$, the best interpolator m is defined as the interpolator with minimal norm:

$$m = \arg\min_{h \in \mathcal{H}} (\|h\|_{\mathcal{H}} \mid h(x_i) = y_i, i \in 1, ..., n\ ) \tag{2}$$

It can be shown that m corresponds to the orthogonal projection of f onto $\mathcal{H}_x$ which is the subspace of $\mathcal{H}$ and spanned by $k(x_i,.)$.

$$\mathcal{H}_x = span(k(x_i,.), x_i \in X) \tag{3}$$

$k(x_i,.)$ corresponds to a basis of $\mathcal{H}_x$

Example: suppose $\mathcal{H}_x$ have two basis function $k(x_1,.)$ and $k(x_2,.)$, and there is another substance of $\mathcal{H}$, $\mathcal{H}_o$

$$\{v \in \mathcal{H}_o\ st\ v(x_i) = 0\} \tag{4}$$

g is orthogonal to the element $k(x_i,.) \in \mathcal{H}$. So $\mathcal{H}_o$ is orthogonal to all the basis function $k(x_i,.)$ of $\mathcal{H}_x$

$$< k(x_i,.), v(.) >= 0 \tag{5}$$

we also have

$$f - m = v \tag{6}$$

combined above two equation we have

$$< f(.) - a_1 k(x_1,.) - a_2 k(x_2,.), k(x_i,.) >= 0 \tag{7}$$

$$\begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) \\ k(x_2,x_1) & k(x_2,x_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \end{bmatrix} \tag{8}$$

and the best predictor m become

$$\begin{aligned} m &= [k(x_1,.)\ k(x_2,.)] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \\ &= k(\bar{x},.)' K^{-1} y(\bar{x}) \end{aligned} \tag{9}$$

$\bar{x}$ is vector of x_i

In the probabilistic framework, this expression corresponds to the conditional expectation of a centred Gaussian process Z with covariance k knowing the observations. Furthermore, GP provide naturally some prediction variance for the model:

$$\begin{aligned} m(x) = E[Z(x)|Z(x_i) = y_i] = k(\bar{x},x)'K^{-1}y(\bar{x}) \\ v(x) = Var[Z(x)|Z(x_i) = y_i] = k(x,x) - k(\bar{x},x)'K^{-1}k(\bar{x},x) \end{aligned} \tag{10}$$

The squared norm $\|m\|^2_{\mathcal{H}_x}$ is the inner produce $< m, m > = < \sum a_i k(x_i,.), \sum a_j k(.,x_j) > = \bar{a}'k(x,x)\bar{a}$

b_j in equation 6 of Heinonen and d'Alché Buc (2014) is actually the product of the inverse of the gram matrix and measurements $K^{-1}y$

1.1 Representer theorem

Theorem. *Let $\mathcal{X}$ be a nonempty set and k a positive-definite real-valued kernel on $\mathcal{X} \times \mathcal{X}$ with corresponding reproducing kernel Hilbert space H_k. Given a training sample $(z_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$, a strictly monotonically increasing real-valued function $g : [0, +\infty) \to \mathbb{R}$, and an arbitrary empirical risk function $E : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \infty$, then for any $f_m \in H_k$ satisfying*

$$f_m = \arg\min_{f \in \mathcal{H}_k} \{\ E(\ (x_1, y_1, f(x_1)\ ), ..., (x_n, y_n, f(x_n)\ )\ ) + g(\|f\|)\}$$

f_m admits a representation of the form:

$$f_m = \sum_{i=1}^{n} a_k(., x_i), a_i \in \mathbb{R} \tag{11}$$

In other words, if we have a function satisfy equation above (f_m) this function must lie in the subspabce with basis $k(x_i, .)$ in H_k

If we define $\varphi(x) = k(., x)$ Given any $x_1, ..., x_n$. we can use orthogonal projection to decompose any $f \in H_k$ into a sum of two function, one lying in $span\{\varphi(x_1), ..., \varphi(x_n)\}$ and the other lying in the orhtogonal complement:

$$f = \sum_{i=1}^{n} a_i \varphi(x_i) + v, \quad < v, \varphi(x_i) >= 0 \tag{12}$$

for all i Using orthogonal decomposition and reproducing property together,

$$f(x_j) =< \sum a_i \varphi(x_i) + v, \varphi(x_j) >= \sum_{i=1}^{n} a_i < \varphi(x_i), \varphi(x_j) > \tag{13}$$

so $f(x_j)$ is indpendent of v. Consequently, the value of E is indepdent of v. The second term the regularization term,

$$\begin{aligned} g(\|f\|) = g(\| \sum a_i \varphi(x_i) + v\|) = g(\sqrt{\| \sum a_i \varphi(x_i)\|^2 + \|v\|^2}) \\ \geq g(\| \sum a_i \varphi(x_i)\|) \end{aligned} \tag{14}$$

Therefore setting $v = 0$ does not affect the first term of (*), while it strictly decreasing the second term. Consequently, any minimizer f_m in (*) must have $v = 0$, i.e., it must be of the form

$$f_m(.) = \sum a_i \varphi(x_i) = \sum a_i k(., x_i) \tag{15}$$

[not relevent to proof   Riesz representation theorem: If we define H a Hilbert space and H^* is its dual space. if f is a element in H , function $\phi_f(.)$ is defind as a map $\phi_f : H \to \mathbb{R}$, for g in H we have $\phi_f(g) =< g, f >$ where $< ., . >$ is inner product. And $\phi_f(.)$ is one element in H^*

Theorem. *The mapping $\psi : H \to H^*$ defined by $\psi(f) = \phi_f(.)$ is isometric isomorphism: ψ is bijective. and $\|f\| = \|\psi(f)\| = \|\phi_f(.)\| = \| < ., f > \|$ and ......*

]

References

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. D. (2013). Gaussian process models for periodicity detection. *arXiv preprint arXiv:1303.7090.*

Heinonen, M. and d'Alché Buc, F. (2014). Learning nonparametric differential equations with operator-valued kernels and gradient matching. *arXiv preprint arXiv:1411.5172.*