Reproducing Kernel Hilbert Space

Zhe Li

November 12, 2014

1 Reproducing Kernel Hilbert Space

Firstly, we would like to use one concrete example to show the concept input space, feature space, feature map and kernel function. Given the data $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^2$ is the representation of feature and y_i is the target value. Here, $\mathbf{x}_i$ is two dimension. Often it is very hard to find a linear classifier in this two dimensional space to classifier $\mathbf{x}$ properly, therefore one attempt to map the original input space $\mathcal{X}$ to a high dimension space $\mathcal{H}$, corresponding every $\mathbf{x} \in \mathcal{X}$ mapping to $\Phi(\mathbf{x}) \in \mathcal{H}$. For specifical $\mathbf{x}_1 \in \mathcal{X}$, we show the feature map from $\mathbf{x}_1$ to $\Phi(\mathbf{x}_1)$:

$$x_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \xrightarrow{\Phi(\cdot)} \begin{bmatrix} x_{11}^2 \\ x_{12}^2 \\ \sqrt{2}x_{11}x_{12} \\ \sqrt{2}x_{11} \\ \sqrt{2}x_{12} \\ 1 \end{bmatrix}$$

In this example, two dimensional space $\mathcal{X}$ is the input space; six dimensional space $\mathcal{H}$ is the feature space(very often X is also called feature space in many book, in this case, we can consider the feature map is linear map), the maping $\Phi(\cdot)$ from this two dimensional input space $\mathcal{X}$ to six dimensional feature space $\mathcal{H}$ is called feature map. In most linear classifier or regression model, we need to deal with the term $\mathbf{x}_i^T\mathbf{x}_j$, which can be seen as the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ in the input space. Similar, for the feature space $\mathcal{H}$, one needs to compute $\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$, which is the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ in the feature space. For the above example, Let's compute $\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$,

$$\begin{aligned} \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j) &= x_{i1}^2x_{j1}^2 + x_{i2}^2x_{j2}^2 + \sqrt{2}x_{i1}x_{i2}\sqrt{2}x_{j1}x_{j2} + \sqrt{2}x_{i1}\sqrt{2}x_{j1} + \sqrt{2}x_{i2}\sqrt{2}x_{j2} + 1 \\ &= (1 + x_{i1}x_{j1} + x_{i2}x_{j2})^2 \\ &= (1 + \mathbf{x}_i^T\mathbf{x}_j)^2 \end{aligned}$$

It turns out that $\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$ can be computed via the quadratic function in the input space with less compuational cost compared with computing it in the feature space. Thus, we define this quadratic function

$$k(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')^2 \tag{1}$$

as kernel function. There are two benefits following this, one is less computation cost to compute $\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$ for high deminsonal feature space and another is that one does not necessary explicitly construct feature space $\mathcal{H}$, just use the similarity between two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ as $k(\mathbf{x}_i, \mathbf{x}_j)$. Here, we explain some notations, used later on.

$$k(\mathbf{x}_i, \cdot) = (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}, 1)^T \tag{2}$$

Since $k(\cdot, \cdot)$ is symmetric, we have $k(\mathbf{x}_i, \cdot) = k(\cdot, \mathbf{x}_i) = \Phi(\mathbf{x}_i)$. Note that they are vectors in this six dimention feature space. There is slight difference between $k(\mathbf{x}_i, \cdot)$ and $\Phi(\mathbf{x}_i)$, we will explain that later.

In this high dimension feature space, we define a linear function(Here, linear is in terms of this high dimensional feature space)

$$f(x) =< f(\cdot), k(x, \cdot) > \tag{3}$$

The notation $f(\cdot), f$ are same and $k_x(\cdot), k(x, \cdot), k(\cdot, x)$ are same. $f(\cdot)$ is a vector $(f_1\ f_2\ \cdots\ f_6)^T$ representing linear function,

$$f_1 * \dim_1 + f_2 * \dim_2 + \cdots + f_6 * \dim_6 \tag{4}$$

So the $f(x)$ is

$$f(\mathbf{x}_i) = f_1x_{i1}^2 + f_2x_{i2}^2 + f_3\sqrt{2}x_{i1}x_{i2} + f_4\sqrt{2}x_{i1} + f_5\sqrt{2}x_{i2} + f_6 \tag{5}$$

The vector f is kind of similar to w in linear case. f is vector, but representing a linear function, so is $k(x, \cdot)$. Previously we say there is slight difference between $k(x, \cdot)$ and $\Phi(x)$, even though both of them are same vector, but the vector $k(x, \cdot)$ represents a linear function in that six dimension feature space. How can we find vector f or function $f(x)$ so that

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} l(y_i, f^T\Phi_i(\mathbf{x}_i)) + \frac{\lambda}{2}||f||^2 \tag{6}$$

The representer theorem tells us that the vector f is the combination of the vectors $k(\cdot, \mathbf{x}_1), k(\cdot, x_2), \cdots, k(\cdot, x_n)$. Using the linear algebra language, the six dimensional vector

f lies in the space $\mathcal{H}_s = span\{k(\cdot, \mathbf{x}_1), \cdots, k(\cdot, \mathbf{x}_i), \cdots, k(\cdot, \mathbf{x}_n)\}$. The entire six dimensial space $\mathcal{H}$ can be orthogonally decomposed into $\mathcal{H}_s$ and $\mathcal{H}_\perp$, we can write f as

$$f = \alpha_1 k(\cdot, \mathbf{x}_1) + \alpha_2 k(\cdot, \mathbf{x}_2) + \cdots + \alpha_n k(\cdot, \mathbf{x}_n) = \sum_{i=1}^{n} k(\cdot, x_i)\alpha_i$$

$$= \alpha_1 \begin{bmatrix} x_{11}^2 \\ x_{12}^2 \\ \sqrt{2}x_{11}x_{12} \\ \sqrt{2}x_{11} \\ \sqrt{2}x_{12} \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} x_{21}^2 \\ x_{22}^2 \\ \sqrt{2}x_{21}x_{22} \\ \sqrt{2}x_{21} \\ \sqrt{2}x_{22} \\ 1 \end{bmatrix} + \cdots + \alpha_n \begin{bmatrix} x_{n1}^2 \\ x_{n2}^2 \\ \sqrt{2}x_{n1}x_{n2} \\ \sqrt{2}x_{n1} \\ \sqrt{2}x_{n2} \\ 1 \end{bmatrix}$$

So we can write

$$\begin{aligned} f(\mathbf{x}_j) = f^T\Phi(\mathbf{x}_j) &= \alpha_1 k(\cdot, \mathbf{x}_1)^T k(\cdot, \mathbf{x}_j) + \alpha_2 k(\cdot, \mathbf{x}_2)^T k(\cdot, \mathbf{x}_j) + \cdots + \alpha_n k(\cdot, \mathbf{x}_n)^T k(\cdot, \mathbf{x}_j) \\ &= \sum_{i=1}^{n} k(\mathbf{x}_j, \mathbf{x}_i) \\ &= (K\alpha)_j \end{aligned}$$

Via representer theorem, the Eq. (6) can be rewriten as

$$\underset{\alpha \in \mathbb{R}^n}{min} \sum_{i=1} l((K\alpha)_i) + \frac{\lambda}{2}\alpha^T K\alpha \tag{7}$$

Combined the dual form of Eq. (6), we summarize the different form for machine learning problem

Primal form	$\underset{f \in \mathcal{F}}{min} \sum_{i=1}^{n} l(y_i, f^T\Phi_i(\mathbf{x}_i)) + \frac{\lambda}{2}\|\|f\|\|^2$
Primal form + Representer Theorem	$\underset{\alpha \in \mathbb{R}^n}{min} \sum_{i=1} l((K\alpha)_i) + \frac{\lambda}{2}\alpha^T K\alpha$
Dual form	$\underset{\alpha \in \mathbb{R}^n}{max} - \sum_{i=1} l^*(\lambda\alpha_i) - \frac{\lambda}{2}\alpha^T K\alpha$

Since the concepts in Reproducting Kernel Hilbert Space is very hard to understand, we use one concrete example to illustrate the important concepts for that.

2 The Proof of Representer Theorem

From linear algobria, the space $\mathcal{H}$ can be divided into two orthogonal space $\mathcal{H}_s$ and $\mathcal{H}_\perp$, assume $\mathcal{H}_s = span\{k(\cdot, \mathbf{x}_1), \cdots, k(\cdot, \mathbf{x}_i), \cdots, k(\cdot, \mathbf{x}_n)\}$. For any f in the space $\mathcal{H}$, f can be considerred being composing from two part f_s and $f_\perp$, put that formally, $f = f_s + f_\perp$

- 1). $\mathcal{H}_s = span\{k(\cdot, \mathbf{x}_1), \cdots, k(\cdot, \mathbf{x}_i), \cdots, k(\cdot, \mathbf{x}_n)\}$

- 2). orthogonal decomposition: $\mathcal{H} = \mathcal{H}_s \oplus \mathcal{H}_\perp, \forall f \in \mathcal{H}, f = f_s + f_\perp$

- 3). pointwise evaluation decompostion

$$\begin{aligned} f(\mathbf{x}_i) &= f_s(\mathbf{x}_i) + f_\perp(\mathbf{x}_i) \\ &=< f_s, k(\cdot, \mathbf{x}_i) > + < f_\perp(\cdot), k(\cdot, x_i) > \\ &=< f_s, k(\cdot, \mathbf{x}_i) >= f_s(\mathbf{x}_i) \end{aligned}$$

  Since $f_\perp$ and $k(\cdot, \mathbf{x}_i)$ are orthogonal.

- 4). norm decomposition $||f||^2 = ||f_s||^2 + ||f_\perp||^2 \geq ||f_s||^2$

- 5). decompose the global cost

$$\begin{aligned} \sum_{i=1}^{n} l(y_i, f(x_i)) + ||f||^2 &= \sum_{i=1}^{n} l(y_i, f_s(x_i)) + ||f_s||^2 + ||f_\perp||^2 \\ &= \sum_{i=1}^{n} l(y_i, f_s(x_i)) + ||f_s||^2 \end{aligned}$$

So, we have $\underset{f \in \mathcal{H}}{argmin}$ Obj func $= \underset{f \in \mathcal{H}_s}{argmin}$ Obj func. In a word, if in small space, we can find a function satisfying the requirement and if we go beyond that small space, definitely we will increase the cost. If you can finish your job in a small space, don't go to the bigger space, which makes worse.

3 Reference

http://asi.insa-rouen.fr/enseignants/~scanu/tutorial_03_Noyaux.pdf
http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/Slides4A.pdf
http://www.di.ens.fr/~fbach/eccv08_fbach.pdf
http://www.cs.berkeley.edu/~bartlett/courses/281b-sp08/7.pdf