Note of STAT 671

Statistical Learning I 2019

Contents

1

1.1 Kernel `10/02`

- Cat and Dog problem

1.1.1 A Simple Classifier

- $\mathcal{X} \mapsto \mathbb{R}^2$, Training set:

$$T = \{(x_i, y_i); x_i \in \mathcal{X}, y_i \in \{-1; +1\}\}$$

Notate $I_+ = \{i; y_i = +1\},\ I_- = \{i; y_i = -1\}$ Number of $I_+ = n_+;\ I_- = n_-;\ T = n = n_+ + n_-$

$$C_+ = \frac{1}{n_+} \sum_{i \in I_+}^{n} x_i; \quad C_- = \frac{1}{n_-} \sum_{i \in I_-}^{n} x_i; \quad C = \frac{1}{2}(C_+ + C_-)$$

- Deifne the generalized "simple classifier" $g : \mathbb{R}^2 \to \mathbb{R}$

$$\begin{aligned} g(x) &= \langle C_+ - C_-, X - C \rangle_{\mathbb{R}^2} = (X - C)^T (C_+ - C_-) \\ &= \langle X, C_+ \rangle - \langle X, C_- \rangle + b \end{aligned}$$

- A binary "simple classifier" is then $f(x) = \begin{cases} +1 & \text{if } g(x) \geq 0 \\ -1 & \text{if } g(x) < 0 \end{cases}$

Let us write $g(x)$ using $\langle \cdot, \cdot \rangle_{\mathbb{R}^2}$ such that we can propose other classifiers by using the kernel trick, that is reproduing $\langle \cdot, \cdot \rangle_{\mathbb{R}^2}$ by $k(\cdot, \cdot)$ a p.d. kernel.

$$g(x) = \langle C_+, X \rangle - \langle C_-, X \rangle - \langle C_+, C \rangle + \langle C_-, C \rangle$$

$\langle C_+, X \rangle = \frac{1}{n_+} \sum_{i \in I_+}^{n} \langle x_i, x \rangle;$

$\langle C_-, X \rangle = \frac{1}{n_-} \sum_{i \in I_-}^{n} \langle x_i, x \rangle;$

$\langle C_+, C\rangle = \langle C_+, \frac{1}{2}C_+\rangle + \langle C_+, \frac{1}{2}C_-\rangle = \frac{1}{2n_+^2}\sum_{(i,j)\in I_+}\langle x_i, x_j\rangle + \frac{1}{2}\langle C_+, C_-\rangle$
$\langle C_-, C\rangle = \langle C_-, \frac{1}{2}C_+\rangle + \langle C_-, \frac{1}{2}C_-\rangle = \frac{1}{2}\langle C_+, C_-\rangle + \frac{1}{2n_-^2}\sum_{(i,j)\in I_-}\langle x_i, x_j\rangle$

$$g(x) = \frac{1}{n_+}\sum_{i\in I_+}^{n}\langle x_i, x\rangle - \frac{1}{n_-}\sum_{i\in I_-}^{n}\langle x_i, x\rangle - \frac{1}{2n_+^2}\sum_{(i,j)\in I_+}\langle x_i, x_j\rangle - \frac{1}{2}\langle C_+, C_-\rangle + \frac{1}{2}\langle C_+, C_-\rangle + \frac{1}{2n_-^2}\sum_{(i,j)\in I_-}\langle x_i, x_j\rangle$$
$$= \sum_{i=1}^{n}\alpha_i\langle x_i, x\rangle + b; \text{where } \alpha_i = \begin{cases}\frac{1}{n_+} & y_i = +1\\ \frac{-1}{n_-} & y_i = -1\end{cases}; \; b = \frac{1}{2n_-^2}\sum_{(i,j)\in I_-}\langle x_i, x_j\rangle - \frac{1}{2n_+^2}\sum_{(i,j)\in I_+}\langle x_i, x_j\rangle$$

1.1.2 A simple geometric solution

1.1.3 A more general solution

1.2 RKHS 10/09

Reproducing Kernel Hilbert Space
A Hilbert Space is a complete inner product space.
A inner product space is a vector space with an inner product (dot product, scalar product).
Dot product $\vec{a}\vec{b} = a_x b_x + a_y b_y = |\vec{a}||\vec{b}|\cos(\theta)$
Start with a vector space (H,+,·) over $\mathbb{R}$ (· scalor multiplication)
An inner product is a mapping: $H \times H \to \mathbb{R}$ such that

1. $\langle f, g\rangle = \langle g, f\rangle$ symmetry for any $f, g \in H$
2. $\langle \alpha f_1 + \beta f_2, g\rangle = \alpha\langle f_1, g\rangle + \beta\langle f_2, g\rangle$ for any $f, g \in H; \alpha, \beta \in \mathbb{R}$
3. $\langle f, f\rangle \geq 0$ for all $f \in H$
4. $\langle f, f\rangle = 0 \iff f = 0_H$

We can define $\|f\|^2 = \langle f, f\rangle$ that defines a Norm on H
A metric space is complete for an inner product when it cantains the limit fo all the Cauchy sequences for this inner product.

-

$x, x' \in \mathcal{X} \neq \phi, \phi \in \mathcal{H}$
K is a positive definite kernel, $\mathcal{H}$ is a Hilbert Space of function $\mathcal{X} \mapsto \mathbb{R}$.
We known that if a function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ verifies $k(x, x') = \langle\phi(x), \phi(x')_{\mathcal{H}}$, then it is a positive kernel

- Reverse: Aronsjar Theorem

If k is a positive definite kernel then there exist $\mathcal{H}$ and ϕ such that $k(x, x') = \langle\phi(x), \phi(x')_{\mathcal{H}}$ is true.
Let us start with k and come up with $\mathcal{H}$ and $\phi : \mathcal{X}, k(\cdot, \cdot)$
Let us start $\mathcal{H}$ with the function $k(\cdot, x)$ for all $x \in \mathcal{X}$

1.2.1 Example 0: Linear kernel

$\mathcal{X} = \mathbb{R}, k(x, x') = xx', k(\cdot, x) : y \mapsto yx$

1.2.2 Example 1: Gaussian kernel with parametor σ^2

$k(\cdot, x) : y \mapsto \exp[-\frac{1}{2\sigma^2}(y - x)^2]$
Let us create a vector space by adding all the finite linear combination of $k(\cdot, x), x \in \mathcal{X}$

$$V = \{f : \mathcal{X} \to \mathbb{R},\; f(x) = \sum_{i=1}^{n}\alpha_i k(x, x_i) \text{ for some } n \geq 1;\; x_1, .., x_n \in \mathcal{X}; \alpha_1, .., \alpha_n \in \mathbb{R}\}$$

$$f \in V \leftrightarrow \begin{Bmatrix} x_1, .., x_n \\ \alpha_1, .., \alpha_n \end{Bmatrix} \quad g \in V \leftrightarrow \begin{Bmatrix} y_1, .., y_m \\ \beta_1, .., \beta_m \end{Bmatrix} \quad f + g \leftrightarrow \begin{Bmatrix} x_1, .., x_n, y_1, .., y_m \\ \alpha_1, .., \alpha_n, \beta_1, .., \beta_m \end{Bmatrix} \quad \gamma f \leftrightarrow \begin{Bmatrix} x_1, .., x_n \\ \gamma\alpha_1, .., \gamma\alpha_n \end{Bmatrix}, \gamma \in \mathbb{R}$$

$$\gamma_1 f + \gamma_2 g \leftrightarrow \begin{Bmatrix} \overbrace{x_1, .., x_n}^{z_1, .., z_n}, \overbrace{y_1, .., y_m}^{z_{n+1}, .., z_{n+m}} \\ \underbrace{\gamma_1\alpha_1, .., \gamma_1\alpha_n}_{\delta_1, .., \delta_n}, \underbrace{\gamma_2\beta_1, .., \gamma_2\beta_m}_{\delta_{n+1}, .., \delta_{n+m}} \end{Bmatrix} \leftrightarrow h(x) = \sum_{i=1}^{n+m}\delta_i k(x, z_i)$$

$$(\gamma_1 f + \gamma_2 g)(x) = \gamma_1\sum_{i=1}^{n}\alpha_i k(x, x_i) + \gamma_2\sum_{i=1}^{m}\beta_i k(x, y_i) = \gamma_1 f(x) + \gamma_2 g(x)$$

Note: the representation $\begin{Bmatrix} x_1, .., x_n \\ \alpha_1, .., \alpha_n \end{Bmatrix}$ of a function in V is not necessary unique

- Define $\langle f,g\rangle = \sum_{i=1}^{n}\alpha_i\sum_{j=1}^{m}\beta_i k(x_i,y_j)$ is a function $\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$

$f\in V \leftrightarrow \begin{Bmatrix} x_1,..,x_n \\ \alpha_1,..,\alpha_n \end{Bmatrix}; g\in V \leftrightarrow \begin{Bmatrix} y_1,..,y_m \\ \beta_1,..,\beta_m \end{Bmatrix}$

$$\langle f,g\rangle = \sum_{i=1}^{n}\alpha_i\underbrace{\sum_{j=1}^{m}\beta_i k(x_i,y_j)}_{g(x_i)} = \sum_{i=1}^{n}\alpha_i g(x_i) = \sum_{j=1}^{m}\beta_i\underbrace{\sum_{i=1}^{n}\alpha_i k(y_j,x_i)}_{f(y_j)} = \sum_{j=1}^{m}\beta_i f(y_j)$$

which shows that $\langle f,g\rangle$ does not depend on the particular representation of (f,g)
So it is a function $\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$
$\langle f,k(\cdot,x)\rangle = \sum_{i=1}^{n}\alpha_i k(x_i,x) = f(x)$
$\langle k(\cdot,y),k(\cdot,x)\rangle = k(x,y)$

1.3 RKHS construction and definitions 10/14

$,\phi\in\mathcal{H}$
K is a positive definite kernel over $\mathcal{X}\neq\phi \iff$ There is some Hilbert Space $\mathcal{H}$ and some mapping $\phi: x\mapsto\mathcal{H}$ such that $k(x,y) = \langle\phi(x),\phi(y)\rangle_{\mathcal{H}}$ is true for every $(x,y)\in\mathcal{X}\times\mathcal{X}$
For constructing $t\mapsto k(t,x), x\in\mathbb{R}$, add linear combinations

$$f:\mathcal{X}\mapsto\mathbb{R};\ f(x) = \sum_{i=1}^{n}\alpha_i k(x,x_i);\ g(x) = \sum_{j=1}^{m}\beta_j k(x,y_j)$$

- Define $\langle f,g\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_i k(x_i,y_j)$

1. not depend on the "represenation" in term of $\begin{Bmatrix} x_1,..,x_n \\ \alpha_1,..,\alpha_n \end{Bmatrix};\begin{Bmatrix} y_1,..,y_m \\ \beta_1,..,\beta_m \end{Bmatrix}$
2. $\langle f,g\rangle = \langle g,f\rangle$
3. Linearity $\langle f+g,h\rangle = \langle f,h\rangle + \langle g,h\rangle;\ \alpha\langle f,g\rangle = \alpha\langle f,g\rangle$
4. $\langle f,f\rangle \geq 0 \iff$ k has the definite positive property

$\langle f,k(\cdot,x)\rangle = \sum_{i=1}^{n}\alpha_i k(x_i,x) = f(x), f\in\begin{Bmatrix} x_1,..,x_n \\ \alpha_1,..,\alpha_n \end{Bmatrix}; k(\cdot,x) = (x,1)^T$
$k(x,y) = \langle\phi(x),\phi(y)\rangle = \langle k(\cdot,y),k(\cdot,x)\rangle = \langle k(\cdot,x),k(\cdot,y)\rangle$

- Proof $\langle f,f\rangle = 0 \implies f=0 \iff$ for any $x\in\mathcal{X}, f(x)=0$

Step 1 check that $\langle f,g\rangle$ is p.d.;
$f_1,..f_n$, scalar $\gamma_1,..,\gamma_n$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_i\gamma_j\langle f_i,f_j\rangle = \langle\sum_{i=1}^{n}\gamma_i f_i,\sum_{j=1}^{n}\gamma_j f_j\rangle \geq 0, g\in H$$

Step 2 Use Cauchy-Schwarz inequality for $\langle f,g\rangle$
$x\in\mathcal{X}, f\in\mathcal{H}$

$$|f(x)|^2 = |\langle f,k(\cdot,x)\rangle|^2 \leq \|f\|^2\|k(\cdot,x)\|^2 = \|f\|^2 k(x,x)$$

then for any $x\in\mathcal{X}$, $\|f\|^2 = \langle f,f\rangle = 0 \implies |f(x)|^2 = 0 \implies f(x) = 0$
We have shown that $(H,\langle\cdot,\cdot\rangle)$ just constructed to a inner product space pre-Hilbert Space.
It can be completed into a Hilbert Space by including the limits of convergent Cauchy sequances

- Define RKHS 1

$X\neq\phi$, $\mathcal{H}$ is a Hilbert Space of function $\mathcal{X}\mapsto\mathbb{R}$
$\mathcal{H}$ is a Reproducing Kernel Hilbert Space when there is a function $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ such that

1. $k(\cdot,x)\in\mathcal{H}$ for all $x\in\mathcal{X}$
2. Reproducing Property $\langle\underbrace{f}_{function},\underbrace{k(\cdot,x)}_{argument}\rangle_{\mathcal{H}} = f(x)$ for any $f\in\mathcal{H}$

1.3.1 Example 0: $\mathcal{X} \in \mathbb{R}^d, \quad k(x,y) = x^T y$

The RKHS with kernel k is

$$\mathcal{H} = \{f_w : \mathbb{R}^d \mapsto \mathbb{R};\ f_w(x) = w^T x; \quad w \in \mathbb{R}^d\}$$

$$\langle f_v, f_w \rangle_{\mathcal{H}} = v^T w \implies \langle f_v, f_v \rangle = \|f_v\|^2_{\mathcal{H}} = \|v\|^2$$

Let us check that $\mathcal{H}$ is the RKHS associated with k
$t \mapsto k(t,x) = x^T t = (x^T t)^T = t^T x = f_t(x)$
Exercise:
$\langle f, k(\cdot,x)\rangle = \langle f_w, f_x\rangle = x^T w = (x^T w)^T = w^T x = f_w(x)$

1.3.2 Example 1: $\mathcal{X} \in \mathbb{R}^d, \quad k(x,y) = x^T y + c, c > 0$

What is the RKHS associated with k?

$$\mathcal{H} = \{f : \mathbb{R}^d \mapsto \mathbb{R};\ f_{w,w_0}(x) = w^T x + w_0; \quad w \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$$

$$\langle f_{v,v_0}, f_{w,w_0}\rangle_{\mathcal{H}} = v^T w + \frac{1}{c} v_0 w_0 \implies \langle f_{v,v_0}, f_{v,v_0}\rangle = \|f_{v,v_0}\|^2_{\mathcal{H}} = \|v\|^2 + \frac{v_0^2}{c}$$

- Define RKHS 2

$X \neq \phi$, $\mathcal{H}$ is a Hilbert Space of function $\mathcal{X} \mapsto \mathbb{R}$
$\mathcal{H}$ is a RKHS if and only if for any $f \in \mathcal{H}, x \in \mathcal{X}$
the evaluation function $\mathcal{H} \mapsto \mathbb{R}$: $F_x : f \mapsto f(x)$ is continuous
$f, g \in \mathcal{H}$ if $\|f - g\|$ is small then their different $|f(x) - g(x)|$ is small.

1.4 Two Definitions of RKHS (why equvalent) 10/16

$X \neq \phi$, $\mathcal{H}$: Hilbert Space of function $\mathcal{X} \mapsto \mathbb{R}$
Example: $\mathcal{X} = \{x_1, ..x_n\}; \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \subset \{\text{vector of } \mathbb{R}^n\}$

1.4.1 Definition 1:

$\mathcal{H}$ is a RKHS when there is a function $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, $K(\cdot,\cdot)$ such that

- A: $t \mapsto k(t,x) \in \mathcal{H}$ for each x
- B: $\langle f, k(\cdot,x)\rangle_{\mathcal{H}} = f(x)$ for each $f \in \mathcal{H}$, $x \in \mathcal{X}$
  – Reproducing Property

1.4.2 Definition 2:

$\mathcal{H}$ is a RKHS when the evaluation functions

$$\begin{aligned} F_x : \quad & \mathcal{H} && \mapsto \mathbb{R} \\ & f && \mapsto f(x) \quad \text{are continuous.} \end{aligned}$$

1.4.3 Definition 1 $\implies$ Definition 2

F_x is continuous. if

$$\begin{aligned} & \|f - g\|_{\mathcal{H}} && < \delta \quad \text{(might depend on x)} \\ \implies & |f(x) - g(x)| && < \varepsilon \end{aligned}$$

F_x is *C-Lipschitz* continuous when

$$|f(x) - g(x)| \leq c\|f - g\|_{\mathcal{H}}, \quad c > 0, \quad \text{for any } f,\ g \in \mathcal{H}$$

C-Lipschitz $\implies$ continuity.

$$|f(x) - g(x)| = |(f-g)(x)| = |\langle f - g, k(\cdot,x)\rangle_{\mathcal{H}}| \leq \|f - g\|_{\mathcal{H}} \underbrace{\langle k(\cdot,x), k(\cdot,x)\rangle^{\frac{1}{2}}}_{k^{\frac{1}{2}}(x,x)}$$

1.4.4 Definition 2 $\implies$ Definition 1

Riesz Representation Theorem: In any Hilber Space of function $\mathcal{X} \mapsto \mathbb{R}$ for which F_x is continuous for each $x \in \mathcal{X}$, then there is an unique element of $\mathcal{H}$, notated g_x, for which $f(x) = \langle f, g_x \rangle_{\mathcal{H}}$ for each $f \in \mathcal{H}$, $\quad g_x(\cdot) = k(\cdot, x)$.

1.5 Examples

1.5.1 Example 0: $\mathcal{X} \in \mathbb{R}^d, \quad k(x,y) = x^T y$

1.5.2 Example 1: $\mathcal{X} = \{x_1, ..x_n\}$,

notate $\underset{(n,n)}{k}$; $[k]_{ij} = k(x_i, x_j)$. k is symmetric and positive semi-definite.
Assume that k is positive definite,

$$f: \mathcal{X} \mapsto \mathbb{R}, \quad \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \subset \mathbb{R}^n$$

$$k(\cdot, x_i) = \begin{bmatrix} k_{1i} \\ \vdots \\ k_{ni} \end{bmatrix} = k_i \,; \quad k = (k_1, ..k_n)$$

$$\begin{aligned} \mathcal{H} &= \{\alpha_1 k_1 + \cdots + \alpha_n k_n;\ \alpha_1, \cdots, \alpha_n \in \mathbb{R}\} \\ &= \text{Span}\{k_1, ..k_n\} = \mathbb{R}^n \quad \text{is a vector space.} \end{aligned}$$

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}} &= f^T k^{-1} g \\ \langle f, k(\cdot, x_i) \rangle &= \langle f, k e_i \rangle, \quad e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i \\ &= f^T \underbrace{k^{-1} k}_{I} e_i \\ &= f^T e_i \\ &= \begin{bmatrix} f(x_1) & \dots & f(x_n) \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i \\ &= f(x_i) \end{aligned}$$

1.5.3 Example 2: $\mathcal{X} \in \mathbb{R}^n, \quad k(x,y) = (x^T y)^2$

$$\mathcal{H} = \{f:\ f(x) = x^T S_x; \quad \underset{(n,n)}{S} \text{ is a symmetric Matrix}\}$$

verify this is a Hilbert Space.

$$\langle f_{S_1}, f_{S_2} \rangle_{\mathcal{H}} = \langle S_1, S_2 \rangle_{\mathcal{F}} = \sum_{i,j=1}^{n} [S_1]_{ij} [S_2]_{ij}$$

$$\langle f_{S_1}, k(\cdot, x_i) \rangle = f_{S_1}(x) \quad \text{check it}$$

$$k(y,x) = (y^T x)(y^T x) = y^T \cdot \underbrace{x x^T}_{\substack{(n,n) \\ \text{symmetric} \\ \text{matrix}}} \cdot y$$