Lecture 4: kernels and associated functions

Stéphane Canu
stephane.canu@litislab.eu

Sao Paulo 2014

March 4, 2014

Plan

1. Statistical learning and kernels
   - Kernel machines
   - Kernels
   - Kernel and hypothesis set
   - Functional differentiation in RKHS

Introducing non linearities through the feature map

SVM Val

$$f(\mathbf{x}) \quad = \quad \sum_{j=1}^{d} x_j w_j + b \quad = \quad \sum_{i=1}^{n} \alpha_i (\mathbf{x}_i^\top \mathbf{x}) + b$$

$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathbb{R}^2$

	x_1
	x_2
	x_3
	x_4
	x_5

linear in $\mathbf{x} \in \mathbb{R}^5$

Introducing non linearities through the feature map

SVM Val

$$f(\mathbf{x}) \quad = \quad \sum_{j=1}^{d} x_j w_j + b \quad = \quad \sum_{i=1}^{n} \alpha_i (\mathbf{x}_i^\top \mathbf{x}) + b$$

$$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathbb{R}^2 \qquad \phi(t) = \begin{array}{|c|c|} t_1 & x_1 \\ t_1^2 & x_2 \\ t_2 & x_3 \\ t_2^2 & x_4 \\ t_1 t_2 & x_5 \end{array}$$

linear in $\mathbf{x} \in \mathbb{R}^5$
quadratic in $\mathbf{t} \in \mathbb{R}^2$

The feature map

$$\begin{array}{rccc} \phi: & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^5 \\ & \mathbf{t} & \longmapsto & \phi(\mathbf{t}) = \mathbf{x} \end{array}$$

$$\mathbf{x}_i^\top \mathbf{x} = \phi(\mathbf{t}_i)^\top \phi(\mathbf{t})$$

Introducing non linearities through the feature map

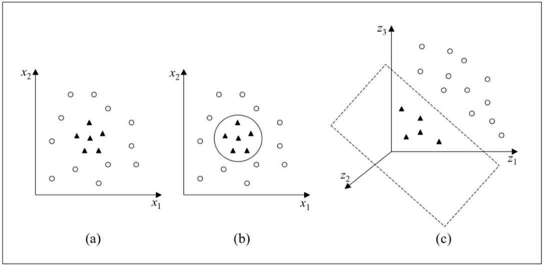


Figura 8. (a) Conjunto de dados não linear; (b) Fronteira não linear no espaço de entradas; (c) Fronteira linear no espaço de características [28]

A. Lorena & A. de Carvalho, Uma Introducão às Support Vector Machines, 2007

Non linear case: dictionnary *vs.* kernel

in the non linear case: use a dictionary of functions

$$\phi_j(\mathbf{x}), j = 1, p \qquad \text{with possibly} \qquad p = \infty$$

for instance polynomials, wavelets...

$$f(\mathbf{x}) = \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}) \qquad \text{with} \qquad w_j = \sum_{i=1}^{n} \alpha_i y_i \phi_j(\mathbf{x}_i)$$

so that

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{p} \phi_j(\mathbf{x}_i)\phi_j(\mathbf{x})}_{k(\mathbf{x_i},\mathbf{x})}$$

Non linear case: dictionnary *vs.* kernel

in the non linear case: use a dictionary of functions

$$\phi_j(\mathbf{x}), j = 1, p \qquad \text{with possibly} \qquad p = \infty$$

for instance polynomials, wavelets...

$$f(\mathbf{x}) = \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}) \qquad \text{with} \qquad w_j = \sum_{i=1}^{n} \alpha_i y_i \phi_j(\mathbf{x}_i)$$

so that

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{p} \phi_j(\mathbf{x}_i)\phi_j(\mathbf{x})}_{k(\mathbf{x}_i, \mathbf{x})}$$

$p \geq n$ so what since $k(\mathbf{x}_i, \mathbf{x}) = \sum_{j=1}^{p} \phi_j(\mathbf{x}_i)\phi_j(\mathbf{x})$

closed form kernel: the quadratic kernel

The quadratic dictionary in $\mathbb{R}^d$:

$$\begin{array}{rccl} \boldsymbol{\Phi}: & \mathbb{R}^d & \to & \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ & \mathbf{s} & \mapsto & \boldsymbol{\Phi} = \left(1, s_1, s_2, ..., s_d, s_1^2, s_2^2, ..., s_d^2, ..., s_i s_j, ...\right) \end{array}$$

in this case

$$\boldsymbol{\Phi}(\mathbf{s})^\top \boldsymbol{\Phi}(\mathbf{t}) = 1 + s_1 t_1 + s_2 t_2 + ... + s_d t_d + s_1^2 t_1^2 + ... + s_d^2 t_d^2 + ... + s_i s_j t_i t_j + ...$$

closed form kernel: the quadratic kernel

The quadratic dictionary in $\mathbb{R}^d$:

$$\begin{array}{rccl} \boldsymbol{\Phi}: & \mathbb{R}^d & \to & \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ & \mathbf{s} & \mapsto & \boldsymbol{\Phi} = \big(1, s_1, s_2, ..., s_d, s_1^2, s_2^2, ..., s_d^2, ..., s_i s_j, ...\big) \end{array}$$

in this case

$$\boldsymbol{\Phi}(\mathbf{s})^\top \boldsymbol{\Phi}(\mathbf{t}) = 1 + s_1 t_1 + s_2 t_2 + ... + s_d t_d + s_1^2 t_1^2 + ... + s_d^2 t_d^2 + ... + s_i s_j t_i t_j + ...$$

The quadratic kenrel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \quad k(\mathbf{s}, \mathbf{t}) \begin{array}{l} = \big(\mathbf{s}^\top \mathbf{t} + 1\big)^2 \\ = 1 + 2\mathbf{s}^\top \mathbf{t} + \big(\mathbf{s}^\top \mathbf{t}\big)^2 \end{array}$

computes the dot product of the reweighted dictionary:

$$\begin{array}{rccl} \boldsymbol{\Phi}: & \mathbb{R}^d & \to & \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ & \mathbf{s} & \mapsto & \boldsymbol{\Phi} = \big(1, \sqrt{2}s_1, \sqrt{2}s_2, ..., \sqrt{2}s_d, s_1^2, s_2^2, ..., s_d^2, ..., \sqrt{2}s_i s_j, ...\big) \end{array}$$

closed form kernel: the quadratic kernel

The quadratic dictionary in $\mathbb{R}^d$:

$$\begin{array}{rcl} \mathbf{\Phi} : \quad \mathbb{R}^d & \rightarrow & \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} & \mapsto & \mathbf{\Phi} = \big(1, s_1, s_2, ..., s_d, s_1^2, s_2^2, ..., s_d^2, ..., s_i s_j, ...\big) \end{array}$$

in this case

$$\mathbf{\Phi}(\mathbf{s})^\top \mathbf{\Phi}(\mathbf{t}) = 1 + s_1 t_1 + s_2 t_2 + ... + s_d t_d + s_1^2 t_1^2 + ... + s_d^2 t_d^2 + ... + s_i s_j t_i t_j + ...$$

The quadratic kenrel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \quad k(\mathbf{s}, \mathbf{t}) \begin{array}{l} = \big(\mathbf{s}^\top \mathbf{t} + 1\big)^2 \\ = 1 + 2\mathbf{s}^\top \mathbf{t} + \big(\mathbf{s}^\top \mathbf{t}\big)^2 \end{array}$

computes the dot product of the reweighted dictionary:

$$\begin{array}{rcl} \mathbf{\Phi} : \quad \mathbb{R}^d & \rightarrow & \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} & \mapsto & \mathbf{\Phi} = \big(1, \sqrt{2}s_1, \sqrt{2}s_2, ..., \sqrt{2}s_d, s_1^2, s_2^2, ..., s_d^2, ..., \sqrt{2}s_i s_j, ...\big) \end{array}$$

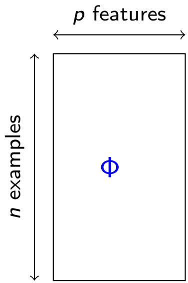
$p = 1 + d + \frac{d(d+1)}{2}$ multiplications *vs.* $d + 1$

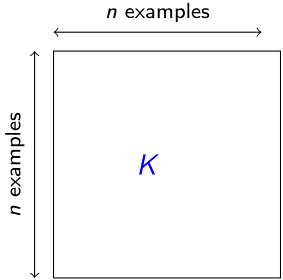
use kernel to save computration

kernel: features throught pairwise comparizons

$\mathbf{x}$ $\phi(\mathbf{x})$

e.g. a text *e.g.* BOW





$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{j=1}^{p} \phi_j(\mathbf{x}_i)\phi_j(\mathbf{x}_j)$$

K The matrix of *pairwise* comparizons ($\mathcal{O}(n^2)$)

Kenrel machine

kernel as a dictionary

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

- α_i influence of example i — depends on y_i
- $k(\mathbf{x}, \mathbf{x}_i)$ the kernel — do NOT depend on y_i

Definition (Kernel)

Let $\mathcal{X}$ be a non empty set (the input space).

A *kernel* is a function k from $\mathcal{X} \times \mathcal{X}$ onto $\mathbb{R}$.

$$\begin{array}{llll} k: & \mathcal{X} \times \mathcal{X} & \longmapsto & \mathbb{R} \\ & \mathbf{s}, \mathbf{t} & \longrightarrow & k(\mathbf{s}, \mathbf{t}) \end{array}$$

Kenrel machine

kernel as a dictionary

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

- α_i influence of example i — depends on y_i
- $k(\mathbf{x}, \mathbf{x}_i)$ the kernel — do NOT depend on y_i

Definition (Kernel)

Let $\mathcal{X}$ be a non empty set (the input space).

A *kernel* is a function k from $\mathcal{X} \times \mathcal{X}$ onto $\mathbb{R}$.

$$\begin{array}{rccc} k: & \mathcal{X} \times \mathcal{X} & \longmapsto & \mathbb{R} \\ & \mathbf{s}, \mathbf{t} & \longrightarrow & k(\mathbf{s}, \mathbf{t}) \end{array}$$

semi-parametric version: given the family $q_j(\mathbf{x}),\ j = 1, p$

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^{p} \beta_j q_j(\mathbf{x})$$

Kernel Machine

Definition (Kernel machines)

$$\mathcal{A}\big((\mathbf{x}_i, y_i)_{i=1,n}\big)(\mathbf{x}) = \psi\Big(\sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^{p} \beta_j q_j(\mathbf{x})\Big)$$

α et $\boldsymbol{\beta}$: parameters to be estimated.

Exemples

$$\mathcal{A}(x) = \sum_{i=1}^{n} \alpha_i (x - x_i)_+^3 + \beta_0 + \beta_1 x \qquad \text{splines}$$

$$\mathcal{A}(\mathbf{x}) = \text{sign}\Big(\sum_{i \in I} \alpha_i \exp^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{b}} + \beta_0\Big) \qquad \text{SVM}$$

$$\mathbb{P}(y|\mathbf{x}) = \frac{1}{Z} \exp\Big(\sum_{i \in I} \alpha_i \mathbb{1}_{\{y=y_i\}} (\mathbf{x}^\top \mathbf{x}_i + b)^2\Big) \qquad \text{exponential family}$$

Plan

1. Statistical learning and kernels
   - Kernel machines
   - Kernels
   - Kernel and hypothesis set
   - Functional differentiation in RKHS

In the beginning was the kernel...

Definition (Kernel)

a function of two variable k from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$

Definition (Positive kernel)

A kernel $k(s,t)$ on $\mathcal{X}$ is said to be positive

- if it is symetric: $k(s,t) = k(t,s)$
- an if for any finite positive interger n:

$$\forall \{\alpha_i\}_{i=1,n} \in \mathbb{R}, \forall \{\mathbf{x}_i\}_{i=1,n} \in \mathcal{X}, \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

it is strictly positive if for $\alpha_i \neq 0$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) > 0$$

Examples of positive kernels

the linear kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \quad k(\mathbf{s}, \mathbf{t}) = \mathbf{s}^\top \mathbf{t}$

symetric: $\mathbf{s}^\top \mathbf{t} = \mathbf{t}^\top \mathbf{s}$

positive:
$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j$$
$$= \left(\sum_{i=1}^{n} \alpha_i \mathbf{x}_i \right)^\top \left(\sum_{j=1}^{n} \alpha_j \mathbf{x}_j \right) = \left\| \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \right\|^2$$

the product kernel: $k(\mathbf{s}, \mathbf{t}) = g(\mathbf{s}) g(\mathbf{t})$ for some $g : \mathbb{R}^d \to \mathbb{R}$,

symetric by construction

positive:
$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j g(\mathbf{x}_i) g(\mathbf{x}_j)$$
$$= \left(\sum_{i=1}^{n} \alpha_i g(\mathbf{x}_i) \right) \left(\sum_{j=1}^{n} \alpha_j g(\mathbf{x}_j) \right) = \left(\sum_{i=1}^{n} \alpha_i g(\mathbf{x}_i) \right)^2$$

k is positive $\Leftrightarrow$ (its square root exists) $\Leftrightarrow$ $k(\mathbf{s}, \mathbf{t}) = \langle \phi_\mathbf{s}, \phi_\mathbf{t} \rangle$

Example: finite kernel

let $\phi_j, j = 1, p$ be a finite dictionary of functions from $\mathcal{X}$ to $\mathbb{R}$ (polynomials, wavelets...)

the feature map and linear kernel

feature map:
$$\begin{array}{rccl} \mathbf{\Phi}: & \mathcal{X} & \to & \mathbb{R}^p \\ & \mathbf{s} & \mapsto & \mathbf{\Phi} = \big(\phi_1(\mathbf{s}), ..., \phi_p(\mathbf{s})\big) \end{array}$$

Linear kernel in the feature space:
$$k(\mathbf{s}, \mathbf{t}) = \big(\phi_1(\mathbf{s}), ..., \phi_p(\mathbf{s})\big)^\top \big(\phi_1(\mathbf{t}), ..., \phi_p(\mathbf{t})\big)$$

***e.g.* the quadratic kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \quad k(\mathbf{s}, \mathbf{t}) = \big(\mathbf{s}^\top \mathbf{t} + b\big)^2$**

feature map:
$$\begin{array}{rccl} \mathbf{\Phi}: & \mathbb{R}^d & \to & \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ & \mathbf{s} & \mapsto & \mathbf{\Phi} = \big(1, \sqrt{2}s_1, ..., \sqrt{2}s_j, ..., \sqrt{2}s_d, s_1^2, ..., s_j^2, ..., s_d^2, ..., \sqrt{2}s_i s_j, ...\big) \end{array}$$

Positive definite Kernel (PDK) algebra (closure)

if $k_1(\mathbf{s},\mathbf{t})$ and $k_2(\mathbf{s},\mathbf{t})$ are two positive kernels

- DPK are a convex cone: $\forall a_1 \in \mathbb{R}^+ \quad a_1 k_1(\mathbf{s},\mathbf{t}) + k_2(\mathbf{s},\mathbf{t})$
- product kernel $k_1(\mathbf{s},\mathbf{t}) k_2(\mathbf{s},\mathbf{t})$

proofs

- by linearity:

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \big(a_1 k_1(i,j) + k_2(i,j)\big) = a_1 \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_1(i,j) + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_2(i,j)$$

- assuming $\exists \psi_\ell$ s.t. $k_1(\mathbf{s},\mathbf{t}) = \sum_\ell \psi_\ell(\mathbf{s}) \psi_\ell(\mathbf{t})$

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \; k_1(\mathbf{x}_i,\mathbf{x}_j) k_2(\mathbf{x}_i,\mathbf{x}_j) &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \Big(\sum_\ell \psi_\ell(\mathbf{x}_i)\psi_\ell(\mathbf{x}_j) k_2(\mathbf{x}_i,\mathbf{x}_j)\Big) \\ &= \sum_\ell \sum_{i=1}^n \sum_{j=1}^n \big(\alpha_i \psi_\ell(\mathbf{x}_i)\big) \; \big(\alpha_j \psi_\ell(\mathbf{x}_j)\big) \; k_2(\mathbf{x}_i,\mathbf{x}_j)\end{aligned}$$

N. Cristianini and J. Shawe Taylor, kernel methods for pattern analysis, 2004

Kernel engineering: building PDK

- for any polynomial with positive coef. ϕ from $\mathbb{R}$ to $\mathbb{R}$

$$\phi\big(k(\mathbf{s},\mathbf{t})\big)$$

- if $\Psi$is a function from $\mathbb{R}^d$ to $\mathbb{R}^d$

$$k\big(\Psi(\mathbf{s}),\Psi(\mathbf{t})\big)$$

- if φ from $\mathbb{R}^d$ to $\mathbb{R}^+$, is minimum in 0

$$k(\mathbf{s},\mathbf{t}) = \varphi(\mathbf{s}+\mathbf{t}) - \varphi(\mathbf{s}-\mathbf{t})$$

- convolution of two positive kernels is a positive kernel

$$K_1 \star K_2$$

Example : the Gaussian kernel is a PDK

$$\begin{aligned}\exp(-\|\mathbf{s}-\mathbf{t}\|^2) &= \exp(-\|\mathbf{s}\|^2 - \|\mathbf{t}\|^2 + 2\mathbf{s}^\top\mathbf{t}) \\ &= \exp(-\|\mathbf{s}\|^2)\exp(-\|\mathbf{t}\|^2)\exp(2\mathbf{s}^\top\mathbf{t})\end{aligned}$$

- $\mathbf{s}^\top\mathbf{t}$ is a PDK and function exp as the limit of positive series expansion, so $\exp(2\mathbf{s}^\top\mathbf{t})$ is a PDK
- $\exp(-\|\mathbf{s}\|^2)\exp(-\|\mathbf{t}\|^2)$ is a PDK as a product kernel
- the product of two PDK is a PDK

an attempt at classifying PD kernels

- stationary kernels, (also called translation invariant):

$$k(s,t) = k_s(s-t)$$

  - radial (isotropic) gaussian: $\exp\left(-\frac{r^2}{b}\right), \quad r = \|s-t\|$
  - with compact support

    c.s. Matèrn : $\max\Big(0, 1-\left(\frac{r}{b}\right)^{\kappa}\Big) \frac{r}{b}^{k} B_k\left(\frac{r}{b}\right), \quad \kappa \geq (d+1)/2$

  - locally stationary kernels: $k(s,t) = k_1(s+t)k_s(s-t)$
    K_1 is a non negative function and K_2 a radial kernel.

- non stationary (projective kernels):

$$k(s,t) = k_p(s^\top t)$$

  - separable kernels $k(s,t) = k_1(s)k_2(t)$ with k_1 and $k_2(t)$ PDK
    in this case $K = k_1 k_2^\top$ where $k_1 = (k_1(\mathbf{x}_1), ..., k_1(\mathbf{x}_n))$

MG Genton, Classes of Kernels for Machine Learning: A Statistics Perspective - JMLR, 2002

some examples of PD kernels...

type	name	$k(s,t)$
radial	gaussian	$\exp\left(-\frac{r^2}{b}\right), \quad r = \|s-t\|$
radial	laplacian	$\exp(-r/b)$
radial	rationnal	$1 - \frac{r^2}{r^2+b}$
radial	loc. gauss.	$\max\left(0, 1-\frac{r}{3b}\right)^d \exp(-\frac{r^2}{b})$
non stat.	χ^2	$\exp(-r/b), \ r = \sum_k \frac{(s_k-t_k)^2}{s_k+t_k}$
projective	polynomial	$(s^\top t)^p$
projective	affine	$(s^\top t + b)^p$
projective	cosine	$s^\top t/\|s\|\|t\|$
projective	correlation	$\exp\left(\frac{s^\top t}{\|s\|\|t\|} - b\right)$

Most of the kernels depends on a quantity b called the bandwidth

the importance of the Kernel bandwidth

for the affine Kernel: Bandwidth = biais

$$k(\mathbf{s}, \mathbf{t}) = (s^\top t + b)^p = b^p \left(\frac{s^\top t}{b} + 1 \right)^p$$

for the gaussian Kernel: Bandwidth = influence zone

$$k(\mathbf{s}, \mathbf{t}) = \frac{1}{Z} \exp\left(-\frac{\|s - t\|^2}{2\sigma^2} \right) \qquad b = 2\sigma^2$$

the importance of the Kernel bandwidth

for the affine Kernel: Bandwidth = biais

$$k(\mathbf{s},\mathbf{t}) = (s^\top t + b)^p = b^p \left(\frac{s^\top t}{b} + 1\right)^p$$

for the gaussian Kernel: Bandwidth = influence zone

$$k(\mathbf{s},\mathbf{t}) = \frac{1}{Z}\exp\left(-\frac{\|s-t\|^2}{2\sigma^2}\right) \qquad b = 2\sigma^2$$

Illustration

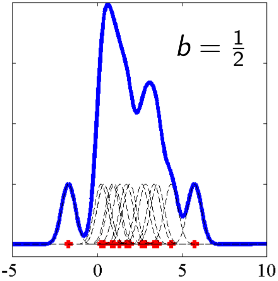
1 d density estimation

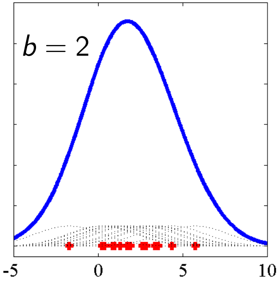
+ data

$(x_1, x_2, ..., x_n)$

– Parzen estimate

$\widehat{\mathbb{P}}(x) = \frac{1}{Z}\sum_{i=1}^{n} k(x, x_i)$





kernels for objects and structures

kernels on histograms and probability distributions
kernel on strings

- spectral string kernel $k(\mathbf{s}, \mathbf{t}) = \sum_u \phi_u(\mathbf{s})\phi_u(\mathbf{t})$
- using sub sequences
- similarities by alignements $k(\mathbf{s}, \mathbf{t}) = \sum_\pi \exp(\beta(\mathbf{s}, \mathbf{t}, \pi))$

kernels on graphs

- the pseudo inverse of the (regularized) graph Laplacian

$$L = D - A \qquad A \text{ is the adjency matrix} D \text{ the degree matrix}$$

- diffusion kernels $\frac{1}{Z(b)} \exp^{bL}$
- subgraph kernel convolution (using random walks)

and kernels on HMM, automata, dynamical system...

Shawe-Taylor & Cristianini's Book, 2004 ; JP Vert, 2006

Multiple kernel

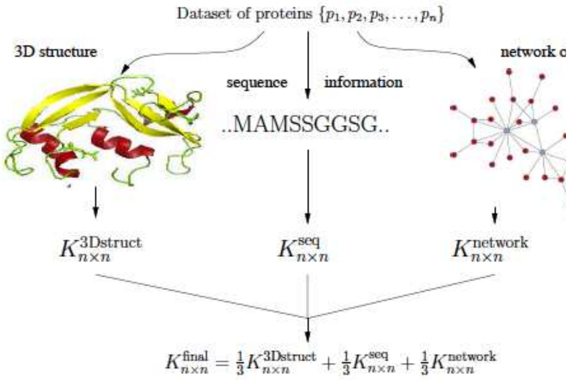


Figure 2: A dataset of proteins can be regarded in (at least) three different wa of 3D structures, a dataset of sequences and a set of nodes in a network which in other. A different kernel matrix can be extracted from each datatype, using know shapes, strings and graphs. The resulting kernels can then be combined togethe weights, as is the case above where a simple average is considered, or estimated the subject of Section 5.2

Gram matrix

Definition (Gram matrix)

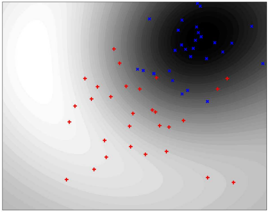
let $k(\mathbf{s}, \mathbf{t})$ be a positive kernel on $\mathcal{X}$ and $(\mathbf{x}_i)_{i=1,n}$ a sequence on $\mathcal{X}$. the Gram matrix is the square K of dimension n and of general term $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

practical trick to check kernel positivity:
K is positive $\Leftrightarrow \lambda_i > 0$ its eigenvalues are posivies: if $K\mathbf{u}_i = \lambda_i \mathbf{u}_i; i = 1, n$

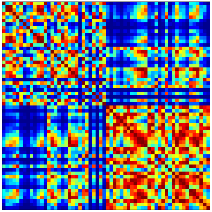
$$\mathbf{u}_i^\top K \mathbf{u}_i = \lambda_i \mathbf{u}_i^\top \mathbf{u}_i = \lambda_i$$

matrix K is the one to be used

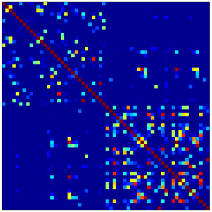
Examples of Gram matrices with different bandwidth



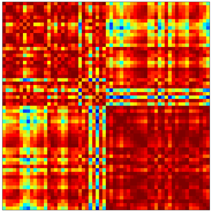
raw data



Gram matrix for $b = 2$



$b = .5$



$b = 10$

different point of view about kernels

kernel and scalar product

$$k(\mathbf{s},\mathbf{t}) = \langle \phi(\mathbf{s}), \phi(\mathbf{t}) \rangle_{\mathcal{H}}$$

kernel and distance

$$d(\mathbf{s},\mathbf{t})^2 = k(\mathbf{s},\mathbf{s}) + k(\mathbf{t},\mathbf{t}) - 2k(\mathbf{s},\mathbf{t})$$

kernel and covariance: a positive matrix is a covariance matrix

$$\mathbb{P}(\mathbf{f}) = \frac{1}{Z} \exp\Big(-\frac{1}{2}(\mathbf{f}-\mathbf{f}_0)^\top K^{-1}(\mathbf{f}-\mathbf{f}_0)\Big)$$

if $f_0 = 0$ and $f = K\alpha$, $\mathbb{P}(\alpha) = \frac{1}{Z}\exp -\frac{1}{2}\alpha^\top K\alpha$

Kernel and regularity (green's function)

$k(\mathbf{s},\mathbf{t}) = P^*P\delta_{\mathbf{s}-\mathbf{t}}$ for some operator P (*e.g.* some differential)

Let's summarize

- positive kernels
- there is a lot of them
- can be rather complex
- 2 classes: radial / projective
- the bandwith matters (more than the kernel itself)
- the Gram matrix summarize the pairwise comparizons

Roadmap

1. Statistical learning and kernels
   - Kernel machines
   - Kernels
   - Kernel and hypothesis set
   - Functional differentiation in RKHS

From kernel to functions

$$\mathcal{H}_0 = \left\{ f \middle| \; m_f < \infty; f_j \in \mathbb{R}; t_j \in \mathcal{X}, \; f(\mathbf{x}) = \sum_{j=1}^{m_f} f_j k(\mathbf{x}, t_j) \right\}$$

let define the bilinear form $\left(g(\mathbf{x}) = \sum_{i=1}^{m_g} g_i k(\mathbf{x}, s_i) \right)$:

$$\forall f, g \in \mathcal{H}_0, \; \langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^{m_f} \sum_{i=1}^{m_g} f_j \, g_i \; k(t_j, s_i)$$

Evaluation functional: $\forall \mathbf{x} \in \mathcal{X}$

$$f(\mathbf{x}) = \langle f(\bullet), k(\mathbf{x}, \bullet) \rangle_{\mathcal{H}_0}$$

from k to $\mathcal{H}$

for any positive kernel, a hypothesis set can be constructed $\mathcal{H} = \overline{\mathcal{H}_0}$ with its metric

RKHS

Definition (reproducing kernel Hibert space (RKHS))

a Hilbert space $\mathcal{H}$ embeded with the inner product $\langle \bullet, \bullet \rangle_{\mathcal{H}}$ is said to be with reproducing kernel if it exists a positive kernel k such that

$$\begin{array}{ll} \forall s \in \mathcal{X}, & k(\bullet, s) \in \mathcal{H} \\ \forall f \in \mathcal{H}, & f(s) = \langle f(\bullet), k(s, \bullet) \rangle_{\mathcal{H}} \end{array}$$

Beware: $f = f(\bullet)$ is a function while $f(\mathbf{s})$ is the real value of f at point $\mathbf{s}$

positive kernel $\Leftrightarrow$ RKHS

- any function in $\mathcal{H}$ is pointwise defined
- defines the inner product
- it defines the regularity (smoothness) of the hypothesis set

Exercice: let $f(\bullet) = \sum_{i=1}^{n} \alpha_i k(\bullet, x_i)$. Show that $\|f\|_{\mathcal{H}}^2 = \alpha^\top K \alpha$

Other kernels (what really matters)

- finite kernels

$$k(\mathbf{s},\mathbf{t}) = \big(\phi_1(\mathbf{s}),...,\phi_p(\mathbf{s})\big)^\top\big(\phi_1(\mathbf{t}),...,\phi_p(\mathbf{t})\big)$$

- Mercer kernels

  positive on a compact set $\quad\Leftrightarrow\quad k(\mathbf{s},\mathbf{t}) = \sum_{j=1}^{p}\lambda_j\phi_j(\mathbf{s})\phi_j(\mathbf{t})$

- positive kernels
- positive semi-definite
- conditionnaly positive (for some functions p_j)

$$\forall\{\mathbf{x}_i\}_{i=1,n}, \forall\alpha_i, \sum_{i}^{n}\alpha_i p_j(\mathbf{x}_i) = 0;\ \ j=1,p, \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) \geq 0$$

- symetric non positive

$$k(\mathbf{s},\mathbf{t}) = \tanh(\mathbf{s}^\top\mathbf{t} + \alpha_0)$$

- non symetric – non positive

the key property: $\nabla_{J_t}(f) = k(t,.)$ holds

The kernel map

- observation: $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_d)^\top$
  - $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \langle \mathbf{w}, \mathbf{x} \rangle_{\mathbb{R}^d}$
- feature map: $\mathbf{x} \longrightarrow \mathbf{\Phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_j(\mathbf{x}), \ldots, \phi_p(\mathbf{x}))^\top$
  - $\mathbf{\Phi} : \mathbb{R}^d \longmapsto \mathbb{R}^p$
  - $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}) \rangle_{\mathbb{R}^p}$
- kernel dictionary: $\mathbf{x} \longrightarrow \mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_i), \ldots, k(\mathbf{x}, \mathbf{x}_n))^\top$
  - $\mathbf{k} : \mathbb{R}^d \longmapsto \mathbb{R}^n$
  - $f(\mathbf{x}) = \displaystyle\sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) = \langle \alpha, \mathbf{k}(\mathbf{x}) \rangle_{\mathbb{R}^n}$
- kernel map: $\mathbf{x} \longrightarrow k(\bullet, \mathbf{x}) \qquad p = \infty$
  - $f(\mathbf{x}) = \langle f(\bullet), K(\bullet, \mathbf{x}) \rangle_{\mathcal{H}}$

Roadmap

1. Statistical learning and kernels
   - Kernel machines
   - Kernels
   - Kernel and hypothesis set
   - Functional differentiation in RKHS

Functional differentiation in RKHS

Let J be a functional

$$J : \begin{array}{ccc} \mathcal{H} & \rightarrow & \mathbb{R} \\ f & \mapsto & J(f) \end{array} \qquad \text{examples:} \qquad J_1(f) = \|f\|^2_{\mathcal{H}}, J_2(f) = f(\mathbf{x}),$$

J directional derivative in direction g at point f

$$dJ(f,g) = \lim_{\varepsilon \rightarrow 0} \frac{J(f + \varepsilon g) - J(f)}{\varepsilon}$$

Gradient $\nabla_J(f)$

$$\nabla_J : \begin{array}{ccc} \mathcal{H} & \rightarrow & \mathcal{H} \\ f & \mapsto & \nabla_J(f) \end{array} \qquad \text{if} \qquad dJ(f,g) = \langle \nabla_J(f), g \rangle_{\mathcal{H}}$$

exercise: find out $\nabla_{J_1}(f)$ et $\nabla_{J_2}(f)$

Hint

$$dJ(f, g) = \left. \frac{dJ(f + \varepsilon g)}{d\varepsilon} \right|_{\varepsilon=0}$$

Solution

$$
\begin{aligned}
dJ_1(f,g) &= \lim_{\varepsilon \to 0} \frac{\|f+\varepsilon g\|^2 - \|f\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\|f\|^2 + \varepsilon^2\|g\|^2 + 2\varepsilon\langle f,g\rangle_{\mathcal{H}} - \|f\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \varepsilon\|g\|^2 + 2\langle f,g\rangle_{\mathcal{H}} \\
&= \langle 2f, g\rangle_{\mathcal{H}}
\end{aligned}
\qquad \Leftrightarrow \qquad \nabla_{J_1}(f) = 2f
$$

$$
\begin{aligned}
dJ_2(f,g) &= \lim_{\varepsilon \to 0} \frac{f(\mathbf{x}) + \varepsilon g(\mathbf{x}) - f(\mathbf{x})}{\varepsilon} \\
&= g(\mathbf{x}) \\
&= \langle k(x,.), g\rangle_{\mathcal{H}}
\end{aligned}
\qquad \Leftrightarrow \qquad \nabla_{J_2}(f) = k(\mathbf{x},.)
$$

Solution

$$
\begin{aligned}
dJ_1(f,g) &= \lim_{\varepsilon \to 0} \frac{\|f+\varepsilon g\|^2 - \|f\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\|f\|^2 + \varepsilon^2 \|g\|^2 + 2\varepsilon \langle f,g\rangle_{\mathcal{H}} - \|f\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \varepsilon \|g\|^2 + 2\langle f,g\rangle_{\mathcal{H}} \\
&= \langle 2f, g\rangle_{\mathcal{H}}
\end{aligned}
\qquad \Leftrightarrow \qquad \nabla_{J_1}(f) = 2f
$$

$$
\begin{aligned}
dJ_2(f,g) &= \lim_{\varepsilon \to 0} \frac{f(\mathbf{x}) + \varepsilon g(\mathbf{x}) - f(\mathbf{x})}{\varepsilon} \\
&= g(\mathbf{x}) \\
&= \langle k(x,.), g\rangle_{\mathcal{H}}
\end{aligned}
\qquad \Leftrightarrow \qquad \nabla_{J_2}(f) = k(\mathbf{x},.)
$$

$$
\text{Minimize}_{f \in \mathcal{H}} J(f) \quad \Leftrightarrow \quad \forall g \in \mathcal{H},\ dJ(f,g) = 0 \quad \Leftrightarrow \quad \nabla_J(f) = 0
$$

Subdifferential in a RKHS $\mathcal{H}$

Definition (Sub gradient)

a subgradient of $J : \mathcal{H} \longmapsto \mathbb{R}$ at f_0 is any function $g \in \mathcal{H}$ such that

$$\forall f \in \mathcal{V}(f_0), \qquad J(f) \;\geq\; J(f_0) + \langle g, (f - f_0)\rangle_{\mathcal{H}}$$

Definition (Subdifferential)

$\partial J(f)$, the subdifferential of J at f is the set of all subgradients of J at f .

$$\mathcal{H} = \mathbb{R} \qquad J_3(x) = |x| \qquad \partial J_3(0) = \{g \in \mathbb{R} \mid -1 < g < 1\}$$
$$\mathcal{H} = \mathbb{R} \qquad J_4(x) = \max(0, 1 - x) \qquad \partial J_4(1) = \{g \in \mathbb{R} \mid -1 < g < 0\}$$

Theorem (Chain rule for linear Subdifferential)

Let T be a linear operator $\mathcal{H} \longmapsto \mathbb{R}$ and φ a function from $\mathbb{R}$ to $\mathbb{R}$
If $J(f) = \varphi(Tf)$
Then $\partial J(f) = \{T^* g \mid g \in \partial\varphi(Tf)\}$, where T^* denotes T's adjoint operator

example of subdifferential in $\mathcal{H}$

evaluation operator and its adjoint

$$
\begin{aligned}
T : \quad \mathcal{H} &\longrightarrow \quad \mathbb{R}^n \\
f &\longmapsto \quad Tf = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top
\end{aligned}
\qquad\qquad
\begin{aligned}
T^* : \quad \mathbb{R}^n &\longrightarrow \mathcal{H} \\
\alpha &\longmapsto T^*\alpha
\end{aligned}
$$

build the adjoint $\langle Tf, \alpha \rangle_{\mathbb{R}^n} = \langle f, T^*\alpha \rangle_{\mathcal{H}}$

example of subdifferential in $\mathcal{H}$

evaluation operator and its adjoint

$$\begin{array}{rrcl} T: & \mathcal{H} & \longrightarrow & \mathbb{R}^n \\ & f & \longmapsto & Tf = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top \end{array}$$

$$\begin{array}{rrcl} T^*: & \mathbb{R}^n & \longrightarrow & \mathcal{H} \\ & \alpha & \longmapsto & T^*\alpha = \sum_{i=1}^n \alpha_i k(\bullet, x_i) \end{array}$$

build the adjoint $\langle Tf, \alpha \rangle_{\mathbb{R}^n} = \langle f, T^*\alpha \rangle_{\mathcal{H}}$

$$\begin{aligned} \langle Tf, \alpha \rangle_{\mathbb{R}^n} &= \sum_{i=1}^n f(x_i)\alpha_i \\ &= \sum_{i=1}^n \langle f(\bullet), k(\bullet, x_i) \rangle_{\mathcal{H}} \alpha_i \\ &= \langle f(\bullet), \underbrace{\sum_{i=1}^n \alpha_i k(\bullet, x_i)}_{T^*\alpha} \rangle_{\mathcal{H}} \end{aligned}$$

example of subdifferential in $\mathcal{H}$

evaluation operator and its adjoint

$$\begin{array}{rrcl} T: & \mathcal{H} & \longrightarrow & \mathbb{R}^n \\ & f & \longmapsto & Tf = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top \end{array}$$

$$\begin{array}{rrcl} T^*: & \mathbb{R}^n & \longrightarrow & \mathcal{H} \\ & \alpha & \longmapsto & T^*\alpha = \sum_{i=1}^n \alpha_i k(\bullet, x_i) \end{array}$$

build the adjoint $\langle Tf, \alpha\rangle_{\mathbb{R}^n} = \langle f, T^*\alpha\rangle_{\mathcal{H}}$

$$\begin{aligned} \langle Tf, \alpha\rangle_{\mathbb{R}^n} &= \sum_{i=1}^n f(x_i)\alpha_i \\ &= \sum_{i=1}^n \langle f(\bullet), k(\bullet, x_i)\rangle_{\mathcal{H}} \alpha_i \\ &= \langle f(\bullet), \underbrace{\sum_{i=1}^n \alpha_i k(\bullet, x_i)}_{T^*\alpha} \rangle_{\mathcal{H}} \end{aligned}$$

$$\begin{array}{rrcl} TT^*: & \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ & \alpha & \longmapsto & TT^*\alpha = \sum_{j=1}^n \alpha_j k(x_j, x_i) \\ & & & = K\alpha \end{array}$$

example of subdifferential in $\mathcal{H}$

evaluation operator and its adjoint

$$\begin{array}{rccl} T: & \mathcal{H} & \longrightarrow & \mathbb{R}^n \\ & f & \longmapsto & Tf = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top \end{array}$$

$$\begin{array}{rccl} T^*: & \mathbb{R}^n & \longrightarrow & \mathcal{H} \\ & \alpha & \longmapsto & T^*\alpha = \sum_{i=1}^n \alpha_i k(\bullet, x_i) \end{array}$$

build the adjoint $\langle Tf, \alpha\rangle_{\mathbb{R}^n} = \langle f, T^*\alpha\rangle_{\mathcal{H}}$

$$\begin{aligned} \langle Tf, \alpha\rangle_{\mathbb{R}^n} &= \sum_{i=1}^n f(x_i)\alpha_i \\ &= \sum_{i=1}^n \langle f(\bullet), k(\bullet, x_i)\rangle_{\mathcal{H}} \alpha_i \\ &= \langle f(\bullet), \underbrace{\sum_{i=1}^n \alpha_i k(\bullet, x_i)}_{T^*\alpha} \rangle_{\mathcal{H}} \end{aligned}$$

$$\begin{array}{rccl} TT^*: & \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ & \alpha & \longmapsto & TT^*\alpha = \sum_{j=1}^n \alpha_j k(x_j, x_i) \\ & & & = K\alpha \end{array}$$

Example of subdifferentials

x given $\quad J_5(f) = |f(x)| \qquad \partial J_5(f_0) = \{g(\bullet) = \alpha k(\bullet, \mathbf{x})\ ; -1 < \alpha < 1\}$

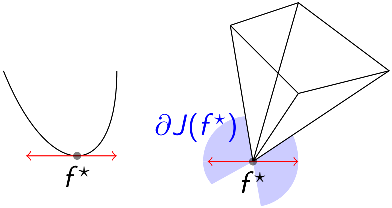
x given $\quad J_6(f) = \max(0, 1 - f(x)) \qquad \partial J_6(f_1) = \{g(\bullet) = \alpha k(\bullet, \mathbf{x})\ ; -1 < \alpha < 0\}$

Optimal conditions

Theorem (Fermat optimality criterion)

When $J(f)$ is convex, $f^\star$ is a stationary point of problem $\min_{f \in \mathcal{H}} J(f)$

If and only if $0 \in \partial J(f^\star)$



exercice: find for a given $y \in \mathbb{R}$ (from Obozinski)

$$\min_{x \in \mathbb{R}} \; \frac{1}{2}(x-y)^2 + \lambda |x|$$

Let's summarize

- positive kernels $\Leftrightarrow$ RKHS $= \mathcal{H}$ $\Leftrightarrow$ regularity $\|f\|^2_{\mathcal{H}}$
- the key property: $\nabla_{J_t}(f) = k(t,.)$ holds not only for positive kernels
  $f(\mathbf{x}_i)$ exists (pointwise defined functions)
- universal consistency in RKHS
- the Gram matrix summarize the pairwise comparizons