Scalable ML

10605-10805

Kernel Methods

Barnabás Póczos

Roadmap I

Several algorithms **need the inner products** of features only!

We need feature maps and their inner products

Explicit (feature maps)

$$\phi(x) = [x_1, x_1 x_2^2, \sin(x_1) - x_2, \ldots]$$

Implicit (kernel functions)

$$k(x, y) = \exp(-\|x - y\|^2)$$

It is much **easier to use implicit** feature maps (kernels)

Given a function $k(x, y) = -\|x\|^{42}\|y\|^{14} + \pi$

Is it a kernel function???

Roadmap II

Given a function $k(x,y) = -\|x\|^{42}\|y\|^{14} + \pi$

Is it a kernel function???

Finite $\mathcal{X}$?

Yes

SVD,

eigenvectors, eigenvalues

Positive semi def. matrices

Finite dim feature space

Arbitrary $\mathcal{X}$

+ We have to think about the test data as well...

Mercer's theorem,

eigenfunctions, eigenvalues

Positive semi def. integral operators

Infinite dim feature space (l_2)

If the kernel is pos. semi def. function $\Leftrightarrow$ feature map construction is possible

Normed and L_p spaces

Normed space: A tuple $(\mathcal{X}, \|\cdot\|)$ is called normed space if $\mathcal{X}$ is a vector space and $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ is a norm, that is

For all $x, y \in \mathcal{X}$, $c \in \mathbb{R}$

$$\|x\| \geq 0, \text{ and } \|x\| = 0 \Leftrightarrow x = 0$$

$$\|cx\| = |c|\|x\|$$

$$\|x + y\| \leq \|x\| + \|y\|$$

$L_p(\mathcal{X})$ space: The vector space of all functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$\int_{\mathcal{X}} |f(x)|^p \, dx < \infty, \quad \text{if } p < \infty$$

$$\sup_{x \in \mathcal{X}} |f(x)| < \infty, \quad \text{if } p = \infty$$

One can prove that these are normed spaces:

$$\|f\|_p \doteq \left(\int_{\mathcal{X}} |f(x)|^p \, dx \right)^{1/p} < \infty$$

$$\|f\|_\infty \doteq \sup_{x \in \mathcal{X}} |f(x)|$$

Inner Product

Definition: inner product

$\langle \cdot, \cdot \rangle : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is an inner product in vector space $\mathcal{K}$, iff for all vectors $x, y, z \in \mathcal{K}$ and all scalars $a \in \mathbb{R}$:

* Symmetry: $\langle x, y \rangle = \langle y, x \rangle$.

* Linearity in the first argument:
$\langle ax, y \rangle = a\langle x, y \rangle$, $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.

* Positive-definite: $\langle x, x \rangle \geq 0$ with equality only for $x = 0$.

Definition: Hilbert space

A vector space with inner product

l_p spaces

$$l_p^n = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n \middle| \begin{array}{ll} \sum_{i=1}^n |x_i|^p < \infty & \text{if } 0 \leq p < \infty \\ \max_{i=1,\ldots,n} |x_i| & \text{if } p = \infty \end{array} \right\}$$

l_p **norms**:

Given $x \in l_p^n$, we define $\|x\|_p$ by:

$$\|x\|_p = \begin{cases} \sum_{i=1}^n 1_{x_i \neq 0} & \text{if } p = 0 \\ \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} & \text{if } 0 < p < \infty \\ \max_{i=1}^n |x_i| & \text{if } \infty \end{cases}$$

L_2 and l_2 special cases (Hilbert spaces)

We can define inner products in L_2 and l_2 spaces:

If $f, g \in L_2(\mathcal{X})$ then $\langle f, g \rangle \doteq \int_{\mathcal{X}} f(x) g(x) \, dx$

If $x, y \in l_2^n$ then $\langle x, y \rangle \doteq \sum_{i=1}^n x_i y_i$

Kernels

Definition: (kernel)

We are given a $\phi : \mathcal{X} \to \mathcal{K}$ feature map, where $\mathcal{K}$ is a Hilbert space

The **kernel** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the corresponding inner product function:

$$\boxed{k(x_i, x_j) \doteq \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}}}$$

Gram matrix, Feature space

Definition: (Gram matrix, kernel matrix)

Gram matrix $G \in \mathbb{R}^{m \times m}$ of kernel k at $\{x_1, \ldots, x_m\}$:

$$\left.\begin{array}{l}\text{Given a kernel } k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ \text{and a training set } \{x_1, \ldots, x_m\}\end{array}\right\} \Rightarrow G_{ij} \doteq k(x_i, x_j)$$

Definition: (Feature space, kernel space)

We are given a $\phi : \mathcal{X} \to \mathcal{K}$ feautre map.

$$\mathcal{K} \doteq span\{\phi(x) \mid x \in \mathcal{X}\}$$

PSD matrices

Definition:

Matrix $G \in \mathbb{R}^{m \times m}$ is positive semidefinite (PSD)

$$\Leftrightarrow G \text{ is symmetric, and } 0 \le \boldsymbol{\beta}^T G \boldsymbol{\beta} \;\forall \boldsymbol{\beta} \in \mathbb{R}^{m \times m}$$

Lemma [Gramm matrix is psd]:

The Gram matrix is symmetric, PSD matrix.

Proof:

By definition, $G \in \mathbb{R}^{m \times m}$, $G_{ij} = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}}$

Therefore,

$$0 \le \| \sum_{i=1}^m \beta_i \phi(x_i) \|_{\mathcal{K}}^2 = \langle \sum_{i=1}^m \beta_i \phi(x_i), \sum_{i=1}^m \beta_i \phi(x_i) \rangle_{\mathcal{K}} = \boldsymbol{\beta}^T G \boldsymbol{\beta}$$

Inner products

In many algorithms we need to calculate the inner product between high-dimensional features, e.g.

$$\phi(x_i) \doteq [sin(x_{i,2}), \text{exp}(x_{i,2} + x_{i,1}), x_{i,1}, x_{i,2}^{\tan(x_{i,1})}, \ldots]$$

and

$$\phi(x_j) \doteq [sin(x_{j,2}), \text{exp}(x_{j,2} + x_{j,1}), x_{j,1}, x_{j,2}^{\tan(x_{j,1})}, \ldots]$$

$$\boxed{k(x_i, x_j) \doteq \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}} = ???}$$

Looks ugly, and needs lots of computation...

Can't we just say that let

$$k(x_i, x_j) \doteq \text{exp}(-\|x_i - x_j\|^2) \;???$$

Is there a feature map $\phi(x) \in l_2$ s.t. $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$???

Kernel technique

We have seen so far how to build a kernel $k(\cdot,\cdot)$ from a given feature map $\phi : \mathcal{X} \to \mathcal{K}$

Now we want to do the opposite:

Definition:
A function $k(\cdot,\cdot)$ is kernel $\Leftrightarrow$ there exists a feature space $\mathcal{K}$ and feature map $\phi : \mathcal{X} \to \mathcal{K}$, such that $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{K}}$

Let us try to find ϕ and $\mathcal{K}$ if $k(\cdot,\cdot)$ is given!

Finite example

Goal:

$$\left.\begin{array}{l}\text{Given a kernel } k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ \text{and a FINITE set } \mathcal{X} = \{x_1, \ldots, x_r\}\end{array}\right\} \Rightarrow \text{construct } \mathcal{K} \text{ and } \phi$$

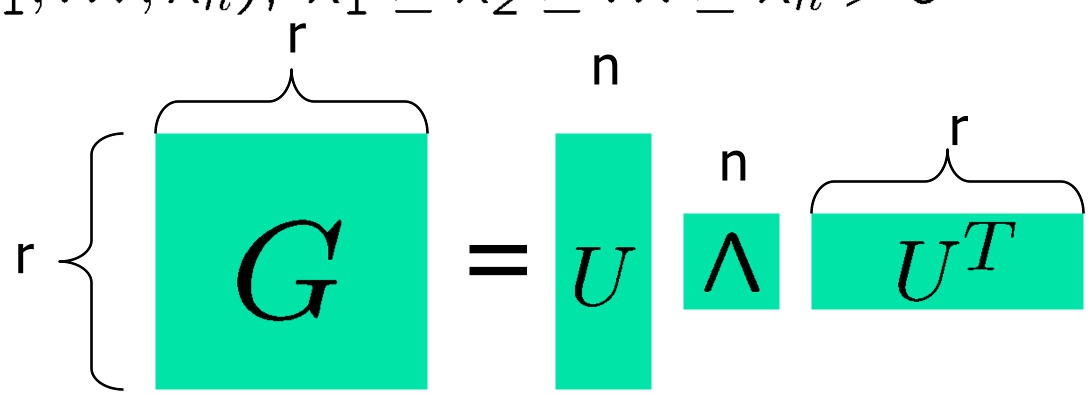
Let us calculate the $G \in \mathbb{R}^{r \times r}$, $G_{ij} = k(x_i, x_j)$ Gram matrix.

If there is such ϕ feautre map, then G is symmetric, PSD by the „Gramm matrix is psd" lemma.

$\Rightarrow G = U \Lambda U^T$ by SVD.

$$U^T U = I_n,\ n = rank(U),\ U = \begin{bmatrix} u_1^T \\ \vdots \\ u_r^T \end{bmatrix} \in \mathbb{R}^{r \times n}$$

$$\Lambda = diag(\lambda_1, \ldots, \lambda_n),\ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$$



Finite example

Lemma:

Let $\mathcal{K} = span\{\phi(x_1), \ldots \phi(x_r)\}$, where $\phi(x_i) \doteq \Lambda^{1/2} u_i \in \mathbb{R}^n$

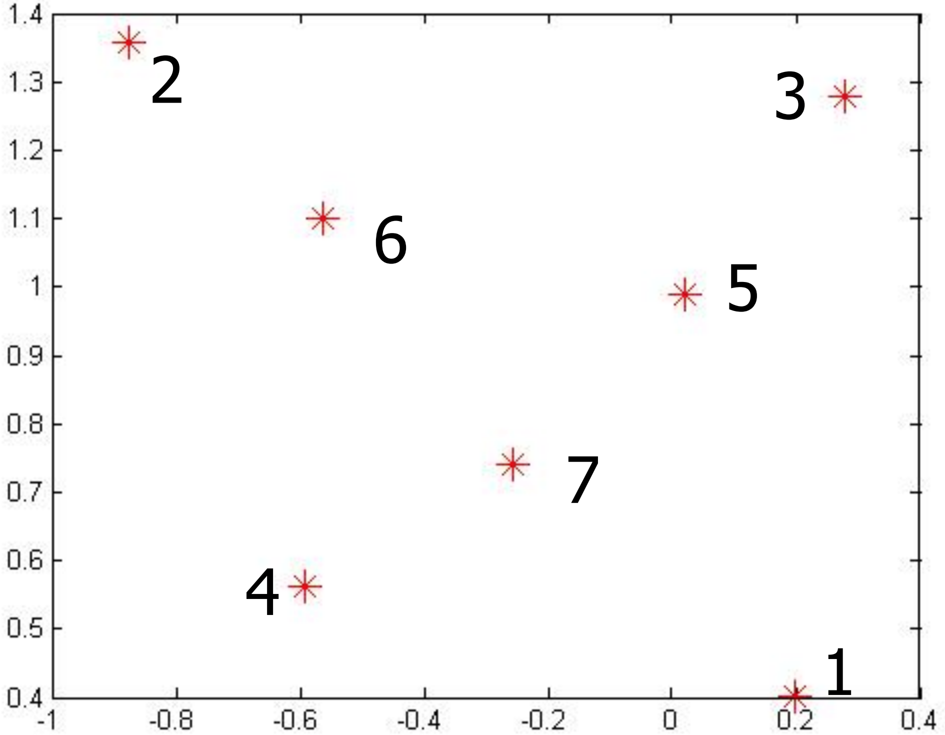
$\Rightarrow \phi(x_i)$ can be used as feature maps to produce Gram matrix G

Proof:

$$\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}} = (\Lambda^{1/2} u_i)^T \Lambda^{1/2} u_j = u_i^T \Lambda u_j = G_{ij}$$

For **general**, NOT FINITE $\mathcal{X}$ sets

the necessary and sufficient conditions of $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a kernel are given by the Mercer's theorem. (See later)



Finite example

Choose 7 2D points

Choose a kernel k

$$G_{ij} = \exp(-|x_i - x_j|^2/10) \text{ can be calculated.}$$

G =

1.0000	0.8131	0.9254	0.9369	0.9630	0.8987	0.9683
0.8131	1.0000	0.8745	0.9312	0.9102	0.9837	0.9264
0.9254	0.8745	1.0000	0.8806	0.9851	0.9286	0.9440
0.9369	0.9312	0.8806	1.0000	0.9457	0.9714	0.9857
0.9630	0.9102	0.9851	0.9457	1.0000	0.9653	0.9862
0.8987	0.9837	0.9286	0.9714	0.9653	1.0000	0.9779
0.9683	0.9264	0.9440	0.9857	0.9862	0.9779	1.0000

[U,D]=svd(G), UDU^T=G, UU^T=I

U =

-0.3709	0.5499	0.3392	0.6302	0.0992	-0.1844	-0.0633
-0.3670	-0.6596	-0.1679	0.5164	0.1935	0.2972	0.0985
-0.3727	0.3007	-0.6704	-0.2199	0.4635	-0.1529	0.1862
-0.3792	-0.1411	0.5603	-0.4709	0.4938	0.1029	-0.2148
-0.3851	0.2036	-0.2248	-0.1177	-0.4363	0.5162	-0.5377
-0.3834	-0.3259	-0.0477	-0.0971	-0.3677	-0.7421	-0.2217
-0.3870	0.0673	0.2016	-0.2071	-0.4104	0.1628	0.7531

D =

6.6315	0	0	0	0	0	0
0	0.2331	0	0	0	0	0
0	0	0.1272	0	0	0	0
0	0	0	0.0066	0	0	0
0	0	0	0	0.0016	0	0
0	0	0	0	0	0.000	0
0	0	0	0	0	0	0.000

Transformed points=sqrt(D)*U^T

Feature transformed points =

-0.9551	-0.9451	-0.9597	-0.9765	-0.9917	-0.9872	-0.9966
0.2655	-0.3184	0.1452	-0.0681	0.0983	-0.1573	0.0325
0.1210	-0.0599	-0.2391	0.1998	-0.0802	-0.0170	0.0719
0.0511	0.0419	-0.0178	-0.0382	-0.0095	-0.0079	-0.0168
0.0040	0.0077	0.0185	0.0197	-0.0174	-0.0146	-0.0163
-0.0011	0.0018	-0.0009	0.0006	0.0032	-0.0045	0.0010
-0.0002	0.0004	0.0007	-0.0008	-0.0020	-0.0008	0.0028
$\phi(x_1)$	$\phi(x_2)$	$\phi(x_3)$	$\phi(x_4)$	$\phi(x_5)$	$\phi(x_6)$	$\phi(x_7)$

$$\phi(x_i) \doteq \Lambda^{1/2} u_i \in \mathbb{R}^n$$

You can check now that

$$\langle \phi(x_i), \phi(x_j) \rangle \doteq \phi(x_i)^T \phi(x_j) = \exp(-|x_i - x_j|^2/10) \ \forall i, j$$

Kernel technique, Finite example

We have seen:

If $\mathcal{X} = \{x_1, \ldots, x_r\}$ and
Gram matrix G is a symmetric, PSD matrix

$\Rightarrow$ we can construct feature space $\mathcal{K}$,
and feature map $\phi : \mathcal{X} \to \mathcal{K}$, compatible with G

Lemma:

These conditions:
(G being symmetric & PSD)
are necessary for G to be a Gram matrix of a kernel

Kernel technique, Finite example

Proof: Indirect (... wrong in the Herbrich's book...)

If $\exists \lambda_n < 0$ and $\exists v \in \mathbb{R}^r$ eigenvector s.t. $Gv = \lambda_n v$

$$\Rightarrow v^T G v = v^T \lambda_n v = \lambda_n \|v\|^2 < 0$$

G is a Gram matrix $\Rightarrow \exists \phi : \mathcal{X} \to \mathcal{K}$, s.t. $G_{ij} = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}}$

Consider the $w \doteq [\phi(x_1), \ldots \phi(x_r)]v \in \mathcal{K}$ vector.

$$\Rightarrow \|w\|^2_{\mathcal{K}} = \langle w, w \rangle_{\mathcal{K}}$$

$$= \langle [\phi(x_1), \ldots \phi(x_r)]v, [\phi(x_1), \ldots \phi(x_r)]v \rangle_{\mathcal{K}} = v^T G v < 0$$

$\lightning$

Kernel technique, Finite example

Summary:

Given a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,
and a FINITE set $\mathcal{X} = \{x_1, \ldots, x_r\}$

$k(\cdot, \cdot)$ is kernel $\Leftrightarrow$ $G = \{k(x_i, x_j)\}_{ij}$ Gram matrix is symmetric, PSD.

How can we generalize this to general domains???

Integral operators, eigenfunctions

Instead of studying the $Gv = \lambda v \;\; G \in \mathbb{R}^{r \times r}$ problem, we examine its generalization:

Let the num of objects r be infinite and let $\mathcal{X} \subseteq \mathbb{R}^d$.

Definition: Integral operator with kernel *k*(.,.)

$$(T_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, x) f(x) dx$$

Remark: This integral operator is a generalization of the matrix vector product

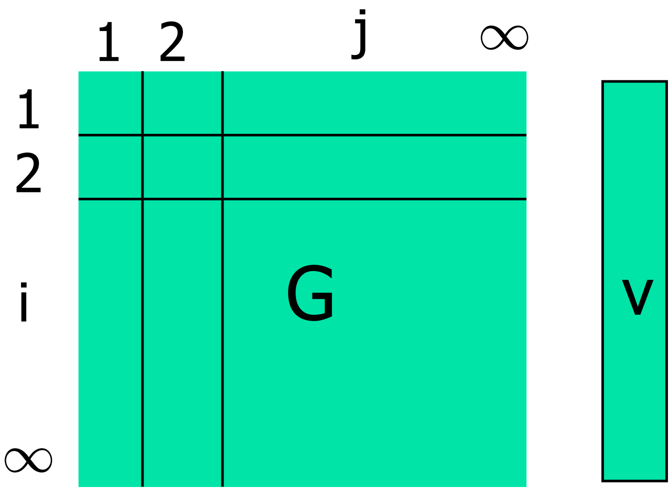
$(T_G v)(i) \doteq (Gv)(i) \;\; i = 1, \ldots, r$ is a special case of this, when the integral is replaced by a finite sum.

$(T_G v)(\cdot) = \sum_j G(\cdot, j) v(j) \qquad (T_G v)(i) = \sum_j G_{ij} v_j$

From Vectors to Functions

Observe that each vector $v = (v_1, \ldots v_n)$ is a function that maps from integers $\{1, \ldots, n\}$ to $\mathbb{R}$

We can generalize vectors easily to countibly infinite domain $\{1, 2, \ldots\}$: $v = (v_1, v_2, \ldots, v_n, \ldots)$



$$(T_G v)(i) \doteq (Gv)(i) = \underbrace{\sum_{j=1}^{\infty}}_{\int_{\mathcal{X}}} \underbrace{G_{ij}}_{k(i,j)} \underbrace{v_j}_{f(j)}$$

And we can even generalize vectors further to the uncountibly infinite domain $\mathbb{R}$: $v = v_x \; x \in \mathbb{R}$, or in other words $v : \mathbb{R} \to \mathbb{R}$

Integral operators, eigenfunctions

Definition: Eigenvalue, Eigenfunction

- λ is the eigenvalue,
- $\Psi \in L_2(\mathcal{X})$ is the eigenfunction

  of integral opreator $(T_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, x) f(x) dx$

$$\Leftrightarrow \begin{cases} \int_{\mathcal{X}} k(x, \bar{x}) \psi(\bar{x}) d\bar{x} = \lambda \psi(x) \quad \forall x \in \mathcal{X} \\ \\ \|\psi\|_{L_2}^2 \doteq \int_{\mathcal{X}} \psi^2(x) dx = 1 \end{cases}$$

The previous $Gv = \lambda v$ is a special case of this, when $\mathcal{X} = \{x_1, \ldots, x_r\}$ is a finite set.

Positive (semi) definite operators

Definition: Positive Definite Operator

$k(\cdot,\cdot)$ is symmetric kernel,

$$\Rightarrow (T_k f)(\cdot) \doteq \int_{\mathcal{X}} k(\cdot,x) f(x) dx$$

$T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ operator is positive semi definit

$$\Leftrightarrow \int_{\mathcal{X}} \int_{\mathcal{X}} k(\tilde{x},x) f(x) f(\tilde{x}) dx d\tilde{x} \geq 0 \quad \forall f \in L_2(\mathcal{X})$$

The previous $v^T G v \geq 0$ is a special case of this, when $\mathcal{X} = \{x_1, \ldots, x_r\}$ is a finite set.

Mercer's theorem

$$(*) \begin{cases} k(\cdot,\cdot) \in L_2(\mathcal{X} \times \mathcal{X}), \\ k \text{ is symmetric: } k(x,\tilde{x}) = k(\tilde{x},x) \\ (T_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot,x) f(x) dx \text{ operator is pos. semi definit} \\ \psi_i,\ i = 1,2,\dots \text{ are the eigenfunctions of } T_k \\ \text{with eigenvalues } \lambda_i \end{cases}$$

$$\Rightarrow \begin{cases} (\lambda_1, \lambda_2, \dots) \in l_1, \quad \lambda_i \geq 0\ \forall i \\ \psi_i \in L_\infty(\mathcal{X}), \quad \forall i = 1,2,\dots \\ k(x,\tilde{x}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(\tilde{x}) \quad \forall x, \tilde{x} \end{cases}$$

2 variables

1 variable

Mercer's theorem

We like the Mercer's theorem becuase of the **expansion**:

$$k(x,\tilde{x}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(\tilde{x}) \quad \forall x, \tilde{x}$$

It shows the **existence of the feature map** $\phi : \mathcal{X} \to \mathcal{K} \subset l_2$

Let $\mathcal{K} \doteq l_2(\mathcal{X})$,
and let $\phi(x) \doteq (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots)^T$

$$\begin{aligned}
&\Rightarrow \langle \phi(x), \phi(\tilde{x}) \rangle_{l_2} \\
&= (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots)^T (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots) \\
&= \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(\tilde{x}) = k(x,\tilde{x}) \quad ☺
\end{aligned}$$

$\phi(x) = (\psi_1(x), \psi_2(x), \ldots) \in l_2$ is known as **Mercer map**

A nicer characterization

The (*) condition in the Mercer's theorem is a bit ugly, but we have a nicer form that characterizes when a function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel
(i.e. scalar product in some inner product space)

Theorem: nicer kernel characterization

$k(\cdot,\cdot)$ is a (Mercer) kernel

$\Leftrightarrow$ $(T_k f)(\cdot)$ is a pos. semi definite operator

$\Leftrightarrow$ $G = (k(x_i, x_j))_{ij}^r \in \mathbb{R}^{r\times r}$ Gram matrix is pos. semi definite $\forall r$, $\forall (x_1, \ldots, x_r) \in \mathcal{X}^r$

Kernel Families

So far we have seen two ways for making a linear classifier nonlinear in the input space:

1. (explicit) Choosing a mapping ϕ
   $\Rightarrow$ Mercer kernel k

2. (implicit) Choosing a Mercer kernel k
   $\Rightarrow$ Mercer map ϕ

Common Kernels

- Polynomials of degree d

$$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\mathbf{v})^d$$

- Polynomials of degree up to d

$$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\mathbf{v}+1)^d$$

- Sigmoid

$$K(\mathbf{u},\mathbf{v}) = \tanh(\eta\mathbf{u}\cdot\mathbf{v}+\nu)$$

- Gaussian kernels

$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||^2}{2\sigma^2}\right)$$

Equivalent to $\phi(x)$ of infinite dimensionality!

Reproducing Kernel Hilbert Spaces

RKHS, Motivation

1., For a given kernel $k(\cdot,\cdot)$ we already know how to define feature space $\mathcal{K}$, and $\phi : \mathcal{X} \to \mathcal{K}$ feature map (Mercer map):

$$\mathcal{K} = l_2, \text{ and } \phi(x) \doteq (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots)^T$$

We will show another way using RKHS

2., Is there a way to efficiently optimize objectives over functions?

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{m} |y_i - f(x_i)| + \lambda \|f\|_{\mathcal{F}}$$

$$\text{or } f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{m} |y_i - f(x_i)|^k + \lambda \|f\|_{\mathcal{F}}^j$$

$$\text{or } f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{m} |y_i - f(x_i)|^k + \lambda \exp\exp\exp(\|f\|_{\mathcal{F}}^j)$$

or ???

Reproducing Kernel Hilbert Spaces

For a given kernel $k(\cdot,\cdot)$ we already know how to define feature space $\mathcal{K}$, and $\phi : \mathcal{X} \to \mathcal{K}$ feature map (Mercer map):

$$\mathcal{K} = l_2, \text{ and } \phi(x) \doteq (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots)^T$$

Now, we show another way using RKHS

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given kernel $\Rightarrow \mathcal{F}_0 \doteq \{k(x,\cdot) | x \in \mathcal{X}\}$ function space

We will add inner product to $\mathcal{F}_0$ function space
$\Rightarrow$ Pre-Hilbert space

Completing (closing) a pre-Hilbert space $\Rightarrow$ Hilbert space

Reproducing Kernel Hilbert Spaces

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given kernel $\Rightarrow \mathcal{F}_0 \doteq \{k(x,\cdot) | x \in \mathcal{X}\}$ function space

$$(x_1, \ldots, x_r) \text{ given } \Rightarrow f(\cdot) \doteq \sum_{i=1}^{r} \alpha_i k(x_i, \cdot) \in \mathcal{F}_0$$

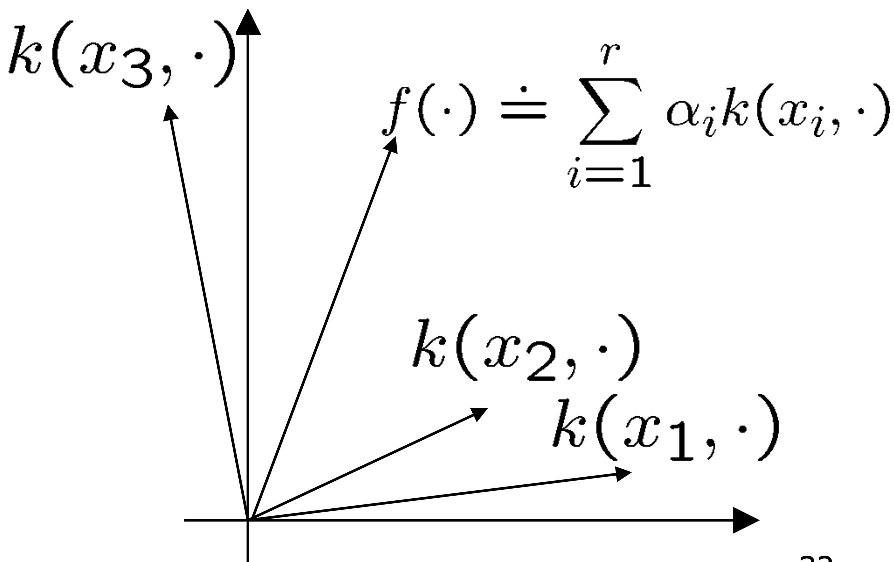
$$(\tilde{x}_1, \ldots, \tilde{x}_s) \text{ given } \Rightarrow g(\cdot) \doteq \sum_{j=1}^{s} \beta_j k(\tilde{x}_j, \cdot) \in \mathcal{F}_0$$

The inner product:

$$\langle f, g \rangle_{\mathcal{F}_0} \doteq \sum_{i=1}^{r} \sum_{j=1}^{s} \alpha_i \beta_j k(x_i, \tilde{x}_j)$$

$$= \sum_{i=1}^{r} \alpha_i g(x_i)$$

$$= \sum_{j=1}^{s} \beta_j f(\tilde{x}_j) \quad (*)$$



Reproducing Kernel Hilbert Spaces

Note:

While for calculating $\langle f,g\rangle_{\mathcal{F}_0}$ we use their representations: $\alpha \in \mathbb{R}^r, \beta \in \mathbb{R}^s, \{x_i\}_{i=1}^r, \{\tilde{x}_j\}_{j=1}^s$ the $\langle f,g\rangle_{\mathcal{F}_0}$ is independent of the representation of f, g

Proof:

If we change $\alpha \in \mathbb{R}^r$ or $x_i \Rightarrow \langle f,g\rangle_{\mathcal{F}_0}$ doesn't change (because of (*)) The same for $\beta \in \mathbb{R}^s$

$$\boxed{\langle f,g\rangle_{\mathcal{F}_0} = \sum_{i_1}^{r} \alpha_i f(x_i) = \sum_{j=1}^{s} \beta_j f(\tilde{x}_j) \quad (*)}$$

Reproducing Kernel Hilbert Spaces

Lemma:

$\langle f, g \rangle$ is an inner product of $\mathcal{F}_0$

$\Rightarrow \mathcal{F}_0$ is pre-Hilbert space

$\mathcal{F} \doteq close(\mathcal{F}_0)$ is a Hilbert space

- **Pre-Hilbert** space:
  Like the Euclidean space with *rational* scalars only
- **Hilbert space**:
  Like the Euclidean space with *real* scalars

Proof:

1., $\langle f, g \rangle_{\mathcal{F}_0} = \langle g, f \rangle_{\mathcal{F}_0}$

2., $\langle cf + dg, h \rangle_{\mathcal{F}_0} = c\langle f, h \rangle_{\mathcal{F}_0} + d\langle g, h \rangle_{\mathcal{F}_0}, \ \forall c, d \in \mathbb{R}, \ \forall f, g, h \in \mathcal{F}_0$

3., $\langle f, f \rangle_{\mathcal{F}_0} \geq 0$

4., $\langle f, f \rangle_{\mathcal{F}_0} = 0 \Leftrightarrow f = 0$

Reproducing Kernel Hilbert Spaces

Lemma: (Reproducing property)

$$\langle f, k(x,\cdot)\rangle_{\mathcal{F}} = f(x)$$

Proof: definition of $\langle f, g\rangle_{\mathcal{F}}$

Lemma:

$$\langle \underbrace{k(x_i,\cdot)}_{\phi(x_i)}, \underbrace{k(x_j,\cdot)}_{\phi(x_j)}\rangle_{\mathcal{F}} = k(x_i, x_j)$$

Proof: reproducing property

Reproducing Kernel Hilbert Spaces

Proof of property 4.,:

$$0 \leq (f(x))^2 = \langle f, k(x,\cdot)\rangle_{\mathcal{F}}^2, \ \forall x$$

rep. property

$$\langle f, k(x,\cdot)\rangle_{\mathcal{F}}^2 \leq \langle f, f\rangle_{\mathcal{F}} \langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{F}} \quad \forall x$$

Cauchy-Schwarz

For Cauchy-Schwarz we don't need 4.,
we need only that <0,0>=0

$$\text{Hence, if } \langle f, f\rangle_{\mathcal{F}} = 0 \Rightarrow (f(x))^2 = 0, \ \forall x \in \mathcal{X}$$

$$\Rightarrow f(x) = 0, \ \forall x \in \mathcal{X}$$

$$\Rightarrow f = 0$$

Methods to Construct Feature Spaces

We now have two methods to construct feature maps from kernels

1., **Mercer map**:
$\mathcal{K} = l_2$, and $\phi(x) \doteq (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots)^T \in l_2$

2., **RKHS map**:
$\mathcal{K} = \mathcal{F}$, and $\phi(x) \doteq k(x, \cdot) \in \mathcal{F}$

For finite discrete $\mathcal{X}$, $|\mathcal{X}| = r$ we already know a 3^{rd} **method:**

3., $\mathcal{K} \subset \mathbb{R}^n$, $\phi(x_i) = \Lambda^{1/2} u_i \in \mathbb{R}^n$, $i = 1, \ldots r$

Well, these feature spaces are all isomorph with each other... ☺

The Representer Theorem

In the SVM problem we could use the dual algorithm, because we had this representation:

$$f(x) \doteq sign(\langle \phi(x), \mathbf{w} \rangle_{\mathcal{K}}) = sign(\sum_{i=1}^{m} \alpha_i k(x, x_i))$$

and thus we had to update $\alpha_1, \ldots, \alpha_m$ only, and not $\mathbf{w} \in \mathcal{K}$!

The **Representer theorem** provides us a big class of problems, where the solution can be represented by

$$\boxed{f(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot), \quad \alpha \in \mathbb{R}^m}$$

The Representer Theorem

Theorem: $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, Mercer kernel on $\mathcal{X}$

$z = (x_1, y_1), \ldots, (x_m, y_m) \in (\mathcal{X} \times \mathcal{Y})^m$ training sample

Empirical risk: $g_{emp} : (\mathcal{X} \times \mathcal{Y} \times \mathbb{R})^m \to \mathbb{R} \cup \{\infty\}$

Regularizer: $g_{reg} : \mathbb{R} \to [0, \infty)$ strictly increasing function

$\mathcal{F}$: RKHS induced by $k(\cdot,\cdot)$

$\Rightarrow$

$$\Rightarrow f^* = \arg\min_{f \in \mathcal{F}} R_{reg}[f, z]$$

$$\doteq \arg\min_{f \in \mathcal{F}} \underbrace{g_{emp}[(x_i, y_i, f(x_i))_{i \in \{1 \ldots m\}}]}_{\text{1st term, empirical risk}} + \underbrace{g_{reg}(\|f\|)}_{\text{2nd term, regularization}}$$

admits the following representation:

$$f^*(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot), \quad \alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$$

The Representer Theorem

Message:

Optimizing in general function classes is difficult, but in RKHS it is only finite! (m) dimensional problem

Proof of Representer Theorem:

$$\phi(x) \doteq k(x, \cdot) = \phi(x)(\cdot)$$

$x_1, \ldots, x_m$ training samples are given

$$f \in \mathcal{F} \Rightarrow f(\cdot) = \sum_{i=1}^{m} \alpha_i \phi(x_i)(\cdot) + v(\cdot)$$

where $\mathcal{F} \ni v \perp span\{\phi(x_1), \ldots, \phi(x_m)\}$,

thus $\langle v, \phi(x_i) \rangle_{\mathcal{F}} = 0 \quad \forall i = 1, \ldots, m$

Proof of Representer Theorem

$$f^* = \arg\min_{f \in \mathcal{F}} R_{reg}[f, z] \doteq \arg\min_{f \in \mathcal{F}} \underbrace{g_{emp}[(x_i, y_i, f(x_i))_{i \in \{1\ldots m\}}]}_{\text{1st term, empirical loss}} + \underbrace{g_{reg}(\|f\|)}_{\text{2nd term, regularization}}$$

$$\Rightarrow f(x_j) = \langle f, \underbrace{k(x_j, \cdot)}_{\phi(x_j)} \rangle_{\mathcal{F}} = \langle \sum_{i=1}^{m} \alpha_i \phi(x_i) + v, \phi(x_j) \rangle_{\mathcal{F}}$$

$$= \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{F}} = \sum_{i=1}^{m} \alpha_i k(x_i, x_j)$$

$\Rightarrow f(x_j)$ depends only on $\alpha_1, \ldots, \alpha_m$, but independent from v!

$\Rightarrow 1^{st}$ term depends only on $\alpha_1, \ldots, \alpha_m$, but not on v

Proof of the Representer Theorem

$$f^* = \arg\min_{f \in \mathcal{F}} R_{reg}[f, z] \doteq \arg\min_{f \in \mathcal{F}} \underbrace{g_{emp}[(x_i, y_i, f(x_i))_{i \in \{1 \ldots m\}}]}_{\text{1st term, empirical loss}} + \underbrace{g_{reg}(\|f\|)}_{\text{2nd term, regularization}}$$

Let us examine the 2^{nd} term.

$$g_{reg}(\|f\|) = g_{reg}(\|\sum_{i=1}^{m} \alpha_i \phi(x_i) + v\|)$$

$$= g_{reg}\left(\sqrt{\|\sum_{i=1}^{m} \alpha_i \phi(x_i)\|_{\mathcal{F}}^2 + \|v\|_{\mathcal{F}}^2}\right)$$

since $\mathcal{F} \ni v \perp span\{\phi(x_1), \ldots, \phi(x_m)\}$

$$\geq g_{reg}(\|\sum_{i=1}^{m} \alpha_i \phi(x_i)\|_{\mathcal{F}})$$

with equality only if $v = 0$!

$\Rightarrow$ any minimizer f^* must have $v = 0$

$$\Rightarrow f^*(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$$

Thanks for Your Attention!