MLCC 2022
Regularization Network II: Kernels

Simone Di Marino
Unige - DIMA, MaLGa

About this class

- Extend our model to deal with non linear problems
- Formulate the Representer Theorem
- Introduce kernel functions (+ examples)

Linear model...

- Data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$
- $\hat{X} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times d}$ and $\hat{y} = (y_1, \ldots, y_n)^\top$.
- Linear model $w \in \mathbb{R}^d$: $y \approx f_w(x) = w^\top x$
- Tikhonov regularization

Summary: our *optimal* regression function will be f_{w_λ}, where

$$w_\lambda = \operatorname{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_w(x_i)) + \lambda \|w\|^2.$$

Linear model...

- Data set $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$
- $\hat{X} = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$ and $\hat{y} = (y_1, \dots, y_n)^\top$.
- Linear model $w \in \mathbb{R}^d$: $y \approx f_w(x) = w^\top x$
- Tikhonov regularization

Summary: our *optimal* regression function will be f_{w_λ}, where

$$w_\lambda = \operatorname{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_w(x_i)) + \lambda \|w\|^2.$$

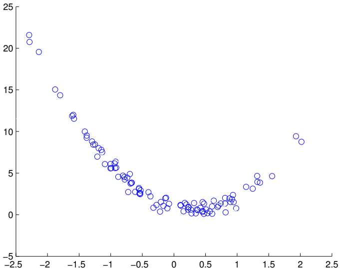
Recall: $w_\lambda = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y}$.

Linear model...

- Data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$
- Linear model $w \in \mathbb{R}^d$

$$f_w(x) = w^\top x$$

Example $d = 1$ and S as in the plot.



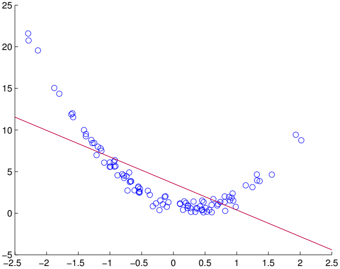
with $w_\lambda = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y}$ for a given $\lambda \geq 0$ (RLS).

Linear model...

- Data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$
- Linear model $w \in \mathbb{R}^d$

$$f_w(x) = w^\top x$$

Example $d = 1$ and S as in the plot.



with $w_\lambda = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y}$ for a given $\lambda \geq 0$ (RLS).

... and beyond

What if we want to learn a more general model?

$$f_w(x) = w_1 x^2 + w_2 x + w_3 = w^\top \phi(x)$$

… and beyond

What if we want to learn a more general model?

$$f_w(x) = w_1 x^2 + w_2 x + w_3 = w^\top \phi(x)$$

It is again a linear model (in w)! But in a different space ($\mathbb{R}^3$ instead of $\mathbb{R}$)

$$\phi(x) = (x^2, x, 1), \quad w = (w_1, w_2, w_3)$$

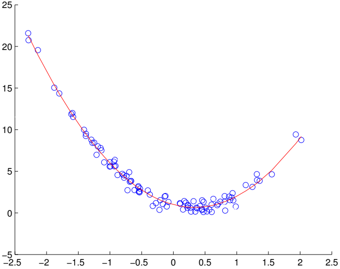
… and beyond

What if we want to learn a more general model?

$$f_w(x) = w_1 x^2 + w_2 x + w_3 = w^\top \phi(x)$$

It is again a linear model (in w)! But in a different space ($\mathbb{R}^3$ instead of $\mathbb{R}$)

$$\phi(x) = (x^2, x, 1), \quad w = (w_1, w_2, w_3)$$



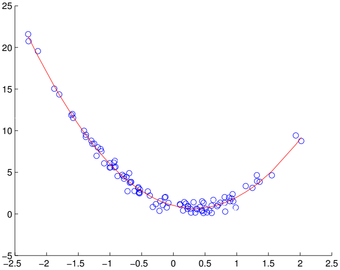
… and beyond

What if we want to learn a more general model?

$$f_w(x) = w_1 x^2 + w_2 x + w_3 = w^\top \phi(x)$$

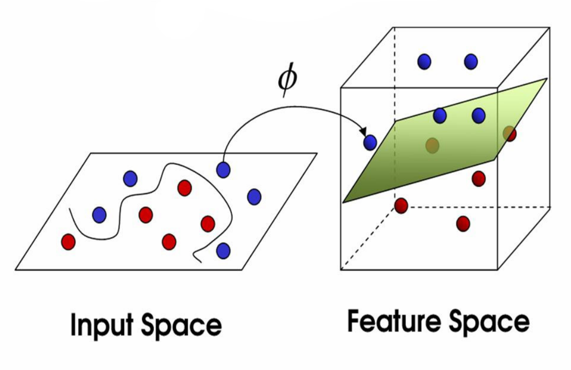
It is again a linear model (in w)! But in a different space ($\mathbb{R}^3$ instead of $\mathbb{R}$)

$$\phi(x) = (x^2, x, 1), \quad w = (w_1, w_2, w_3)$$



Geometric view

$$f(x) = w^\top \Phi(x)$$



Non linear models

- Let $\varphi_j(x) : \mathbb{R}^d \to \mathbb{R}$ with $j \in \{1, \dots, D\}$ (in general with $D >> d$)
- $\phi : \mathbb{R}^d \to \mathbb{R}^D$ is called *feature map* with $\phi(x) = (\varphi_1(x), \dots, \varphi_D(x))^\top$.
- $w \in \mathbb{R}^D$.

Nonlinear model

$$f_w(x) = w^\top \phi(x) = \sum_{j=1}^{D} w_j \varphi_j(x)$$

How to compute a non linear model (least squares)

Let $\hat{\Phi} = (\phi(x_1), \ldots, \phi(x_n))^\top \in \mathbb{R}^{n \times D}$.

$\hat{\Phi}$ is the data matrix in the feature space (simply $\hat{X}$ if ϕ is the identity).

For Regularized Least Squares the explicit minimizer w for the Empirical Loss is (same calculation because the problem is still linear in w)

$$w = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y}$$

Can we do better?

(from a computational point of view)

Note that $\hat{\Phi}^\top \hat{\Phi} \in \mathbb{R}^{D \times D}$

Can we do better?
(from a computational point of view)

Note that $\hat{\Phi}^\top \hat{\Phi} \in \mathbb{R}^{D \times D}$ when D is huge, $\hat{\Phi}^\top \hat{\Phi}$ is not computable.
Can we do better?

Can we do better?
(from a computational point of view)

Note that $\hat{\Phi}^\top \hat{\Phi} \in \mathbb{R}^{D \times D}$ when D is huge, $\hat{\Phi}^\top \hat{\Phi}$ is not computable.
Can we do better?

Representer Theorem (in the least squares context)
If w solves RLS then

$$w = \hat{\Phi}^\top c = \sum_{i=1}^{n} c_i \phi(x_i),$$

where $c = (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y} \in \mathbb{R}^n$ and $\hat{\Phi}\hat{\Phi}^\top \in \mathbb{R}^{n \times n}$.

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I)(\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)\hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I)(\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$\hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top\hat{\Phi} + \lambda n I)^{-1}\hat{\Phi}^\top\hat{y} = \hat{\Phi}^\top(\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}\hat{y}$.
Is it true that

$$(\hat{\Phi}^\top\hat{\Phi} + \lambda n I_D)^{-1}\hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top(\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$\hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top\hat{\Phi} + \lambda n I)\hat{\Phi}^\top(\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}$$

$$\hat{\Phi}^\top(\hat{\Phi}\hat{\Phi}^\top + \lambda n I) \stackrel{?}{=} (\hat{\Phi}^\top\hat{\Phi} + \lambda n I)\hat{\Phi}^\top(\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}(\hat{\Phi}\hat{\Phi}^\top + \lambda n I)$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi} \hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi} \hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$\hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top (\hat{\Phi} \hat{\Phi}^\top + \lambda n I)^{-1}$$

$$\hat{\Phi}^\top (\hat{\Phi} \hat{\Phi}^\top + \lambda n I) \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top (\hat{\Phi} \hat{\Phi}^\top + \lambda n I)^{-1} (\hat{\Phi} \hat{\Phi}^\top + \lambda n I)$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$\hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}$$

$$\hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I) \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top$$

Sketch of the Proof

We want to show that $\omega = (\hat{\Phi}^\top \hat{\Phi} + \lambda n I)^{-1} \hat{\Phi}^\top \hat{y} = \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$.
Is it true that

$$(\hat{\Phi}^\top \hat{\Phi} + \lambda n I_D)^{-1} \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I_n)^{-1}$$

$$\hat{\Phi}^\top \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1}$$

$$\hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top + \lambda n I) \stackrel{?}{=} (\hat{\Phi}^\top \hat{\Phi} + \lambda n I) \hat{\Phi}^\top$$

$$\hat{\Phi}^\top \hat{\Phi}\hat{\Phi}^\top + \lambda n \hat{\Phi}^\top \stackrel{?}{=} \hat{\Phi}^\top \hat{\Phi}\hat{\Phi}^\top + \lambda n \hat{\Phi}^\top$$

Yes, it is true (in general $(BA + \lambda Id)^{-1} B = B(AB + \lambda Id)^{-1}$, whenever A and B are compatible)!

Generalization of Representer Theorem for any Loss Functions

For a given loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, let the problem be

$$w^* = \arg\min_{w \in \mathbb{R}^D} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \phi(x_i)^\top w) + \lambda \|w\|^2$$

The solution can always be written as $w^* = \hat{\Phi}^\top c$ for some coefficients vector $c = (c_1, \ldots, c_n)$

Representer Theorem for general Loss Functions

Let define the linear subspace $\hat{W}$ as $\hat{W} = \{\hat{\Phi}^\top c \mid c \in \mathbb{R}^n\}$.

Representer Theorem for general Loss Functions

Let define the linear subspace $\hat{W}$ as $\hat{W} = \{\hat{\Phi}^\top c \mid c \in \mathbb{R}^n\}$.
By definition of linear subspace we have that

$$w = \hat{w} \; + \; w_\perp \quad \text{for each } w \in \mathbb{R}^D$$

with $\hat{w} \in \hat{W}$ and $v^\top w_\perp = 0$ for each $v \in \hat{W}$.

Representer Theorem for general Loss Functions

Let define the linear subspace $\hat{W}$ as $\hat{W} = \{\hat{\Phi}^\top c \mid c \in \mathbb{R}^n\}$.
By definition of linear subspace we have that

$$w = \hat{w} \ + \ w_\perp \quad \text{for each } w \in \mathbb{R}^D$$

with $\hat{w} \in \hat{W}$ and $v^\top w_\perp = 0$ for each $v \in \hat{W}$.
Moreover note that for each $i \in \{1, \dots n, \}$ we have $\phi(x_i) \in \hat{W}$.

Representer Theorem for general Loss Functions

Let define the linear subspace $\hat{W}$ as $\hat{W} = \{\hat{\Phi}^\top c \mid c \in \mathbb{R}^n\}$.
By definition of linear subspace we have that

$$w = \hat{w} \; + \; w_\perp \quad \text{for each } w \in \mathbb{R}^D$$

with $\hat{w} \in \hat{W}$ and $v^\top w_\perp = 0$ for each $v \in \hat{W}$.
Moreover note that for each $i \in \{1, \ldots n, \}$ we have $\phi(x_i) \in \hat{W}$.
Therefore for any x_i with $i \in \{1, \ldots, n\}$

$$\phi(x_i)^\top w = \phi(x_i)^\top \hat{w} \; + \; \phi(x_i)^\top w_\perp$$

Representer Theorem for general Loss Functions

Let define the linear subspace $\hat{W}$ as $\hat{W} = \{\hat{\Phi}^\top c \mid c \in \mathbb{R}^n\}$.
By definition of linear subspace we have that

$$w = \hat{w} \, + \, w_\perp \quad \text{for each } w \in \mathbb{R}^D$$

with $\hat{w} \in \hat{W}$ and $v^\top w_\perp = 0$ for each $v \in \hat{W}$.
Moreover note that for each $i \in \{1, \dots n, \}$ we have $\phi(x_i) \in \hat{W}$.
Therefore for any x_i with $i \in \{1, \dots, n\}$

$$\phi(x_i)^\top w = \phi(x_i)^\top \hat{w} \, + \, \underbrace{\phi(x_i)^\top w_\perp}_{=0}$$

Representer Theorem for general Loss Functions

Let define the linear subspace $\hat{W}$ as $\hat{W} = \{\hat{\Phi}^\top c \mid c \in \mathbb{R}^n\}$.
By definition of linear subspace we have that

$$w = \hat{w} \; + \; w_\perp \quad \text{for each } w \in \mathbb{R}^D$$

with $\hat{w} \in \hat{W}$ and $v^\top w_\perp = 0$ for each $v \in \hat{W}$.
Moreover note that for each $i \in \{1, \dots n, \}$ we have $\phi(x_i) \in \hat{W}$.
Therefore for any x_i with $i \in \{1, \dots, n\}$

$$\phi(x_i)^\top w = \phi(x_i)^\top \hat{w}$$

Representer Theorem for general Loss Functions

Therefore the problem become

$$w^* = \arg\min_{w \in \mathbb{R}^D} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \phi(x_i)^\top \hat{w}) + \lambda \|w\|^2$$

Moreover, considering that $\hat{w}^\top w_\perp = 0$ we have

$$\|\hat{w}\| \leq \|\hat{w}\| + \|w_\perp\| = \|w\|$$

Representer Theorem for general Loss Functions

Therefore the problem become

$$w^* = \arg\min_{w\in\mathbb{R}^D} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \phi(x_i)^\top \hat{w}) + \lambda\|w\|^2$$

Moreover, considering that $\hat{w}^\top w_\perp = 0$ we have

$$\|\hat{w}\| \leq \|\hat{w}\| + \|w_\perp\| = \|w\|$$

Now let $w^* = \hat{w}^* + w_\perp^*$. The problem is minimized when $w_\perp^* = 0$.

Representer Theorem for general Loss Functions

Therefore the problem become

$$w^* = \arg\min_{w \in \mathbb{R}^D} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \phi(x_i)^\top \hat{w}) + \lambda \|w\|^2$$

Moreover, considering that $\hat{w}^\top w_\perp = 0$ we have

$$\|\hat{w}\| \leq \|\hat{w}\| + \|w_\perp\| = \|w\|$$

Now let $w^* = \hat{w}^* + w_\perp^*$. The problem is minimized when $w_\perp^* = 0$. That is

$$w^* = \hat{\Phi}^\top c$$

for some $c \in \mathbb{R}^n$.

Why we need Kernels...

Let us analyze the regression function f_w, which minimizes RLS for the Generalized Linear model:

$$f_w(x) = \phi(x)^\top \omega = \phi(x)^\top \hat{\Phi}^\top c = \sum_{i=1}^{n} \phi(x)^\top \phi(x_i) c_i$$

where

$$c = (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$$

and $\hat{\Phi}\hat{\Phi}^\top \in \mathbb{R}^{n \times n}$ is

$$(\hat{\Phi}\hat{\Phi}^\top)_{ij} = \phi(x_i)^\top \phi(x_j).$$

Why we need Kernels...

Let us analyze the regression function f_w, which minimizes RLS for the Generalized Linear model:

$$f_w(x) = \phi(x)^\top \omega = \phi(x)^\top \hat{\Phi}^\top c = \sum_{i=1}^{n} \phi(x)^\top \phi(x_i) c_i$$

where

$$c = (\hat{\Phi}\hat{\Phi}^\top + \lambda n I)^{-1} \hat{y}$$

and $\hat{\Phi}\hat{\Phi}^\top \in \mathbb{R}^{n \times n}$ is

$$(\hat{\Phi}\hat{\Phi}^\top)_{ij} = \phi(x_i)^\top \phi(x_j).$$

$f_w(x)$ **is expressed only by using inner products between feature vectors**

Why we need Kernels...

Idea: if we known $\phi(x)^\top\phi(x')$ for each couple $x, x' \in \mathbb{R}^d$ then we can express $f_w(x)$.

Why we need Kernels...

Idea: if we known $\phi(x)^\top \phi(x')$ for each couple $x, x' \in \mathbb{R}^d$ then we can express $f_w(x)$. Therefore we define the *Kernel* as

$$K(x, x') = \phi(x)^\top \phi(x')$$

Why we need Kernels...

Idea: if we known $\phi(x)^\top \phi(x')$ for each couple $x, x' \in \mathbb{R}^d$ then we can express $f_w(x)$. Therefore we define the *Kernel* as

$$K(x, x') = \phi(x)^\top \phi(x')$$

In this way we have

$$f_w(x) = \hat{K}_x^\top (\hat{K} + \lambda n I)^{-1} \hat{y}$$

with $\hat{K}_x = (K(x, x_1), \ldots, K(x, x_n)), \quad (\hat{K})_{ij} = K(x_i, x_j)$.

Why we need Kernels...

Idea: if we known $\phi(x)^\top \phi(x')$ for each couple $x, x' \in \mathbb{R}^d$ then we can express $f_w(x)$. Therefore we define the *Kernel* as

$$K(x, x') = \phi(x)^\top \phi(x')$$

In this way we have

$$f_w(x) = \hat{K}_x^\top (\hat{K} + \lambda n I)^{-1} \hat{y}$$

with $\hat{K}_x = (K(x, x_1), \ldots, K(x, x_n)), \quad (\hat{K})_{ij} = K(x_i, x_j)$.
We don't have to define an explicit ϕ, we need only to define a Kernel K

Why we need Kernels...

Idea: if we known $\phi(x)^\top \phi(x')$ for each couple $x, x' \in \mathbb{R}^d$ then we can express $f_w(x)$. Therefore we define the *Kernel* as

$$K(x, x') = \phi(x)^\top \phi(x')$$

In this way we have

$$f_w(x) = \hat{K}_x^\top (\hat{K} + \lambda n I)^{-1} \hat{y}$$

with $\hat{K}_x = (K(x, x_1), \dots, K(x, x_n)), \quad (\hat{K})_{ij} = K(x_i, x_j)$.
We don't have to define an explicit ϕ, we need only to define a Kernel K
The same holds for general loss functions indeed

$$f_{w^*}(x) = \phi(x)^\top w^* = \phi(x)^\top \hat{\Phi}^\top c = \hat{K}_x^\top c = \sum_{i=1}^n c_i K(x, x_i).$$

Examples of Kernel: Linear Kernel

For $x, z \in \mathbb{R}^d$

$$K(x, z) = x^\top z$$

Proof

$$K(x, z) = \phi(x)^\top \phi(z)$$

with $\phi : \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$\phi(x) = x.$$

Examples of Kernel: Affine Kernel

For $x, z \in \mathbb{R}^d$

$$K(x, z) = x^\top z + \alpha^2$$

Proof

$$K(x, z) = \phi(x)^\top \phi(z)$$

with $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ defined as

$$\phi(x) = (x, \alpha).$$

Examples of Kernel: Polynomial Kernel of degree p

For $p \in \mathbb{N}$

$$K(x,z) = (xz+1)^p \quad \text{with } x, z \in \mathbb{R}$$

Proof

$$(xz+1)^p = \sum_{k=0}^{p} q_{p,k}(xz)^k = \phi(x)^\top \phi(z)$$

with $q_{p,k} = \frac{p!}{k!(p-k)!}$ and $\phi : \mathbb{R} \to \mathbb{R}^{p+1}$ defined as

$$\phi(x) = (\sqrt{q_{p,0}}, \sqrt{q_{p,1}}x, \sqrt{q_{p,2}}x^2, \ldots, \sqrt{q_{p,k}}x^k, \ldots, \sqrt{q_{p,p}}x^p)$$

Examples of Kernel: Polynomial Kernel of degree p

For $p \in \mathbb{N}$

$$K(x, z) = (xz + 1)^p \quad \text{with } x, z \in \mathbb{R}$$

Proof

$$(xz + 1)^p = \sum_{k=0}^{p} q_{p,k} (xz)^k = \phi(x)^\top \phi(z)$$

with $q_{p,k} = \frac{p!}{k!(p-k)!}$ and $\phi : \mathbb{R} \to \mathbb{R}^{p+1}$ defined as

$$\phi(x) = (\sqrt{q_{p,0}}, \sqrt{q_{p,1}} x, \sqrt{q_{p,2}} x^2, \dots, \sqrt{q_{p,k}} x^k, \dots, \sqrt{q_{p,p}} x^p)$$

For $x, z \in \mathbb{R}^d$ it is defined as

$$K(x, z) = (x^\top z + 1)^p$$

Examples of Kernel: Polynomial Kernel of any degree

For $x, z \in [0, 1]$ and $0 < \alpha < 1$

$$K(x, z) = \frac{1}{1 - \alpha^2 xz}$$

Proof

$$\frac{1}{1 - \alpha^2 xz} = \sum_{k=0}^{\infty} (\alpha^2 xz)^k = \phi(x)^\top \phi(z)$$

with $\phi : \mathbb{R} \to \mathbb{R}^{\mathbb{N}}$ defined as

$$\phi(x) = (1, \alpha x, \alpha^2 x^2, \alpha^3 x^3, \dots)$$

Examples of Kernel: Polynomial Kernel of any degree

For $x, z \in [0,1]$ and $0 < \alpha < 1$

$$K(x,z) = \frac{1}{1 - \alpha^2 xz}$$

Proof

$$\frac{1}{1 - \alpha^2 xz} = \sum_{k=0}^{\infty} (\alpha^2 xz)^k = \phi(x)^\top \phi(z)$$

with $\phi : \mathbb{R} \to \mathbb{R}^{\mathbb{N}}$ defined as

$$\phi(x) = (1, \alpha x, \alpha^2 x^2, \alpha^3 x^3, \dots)$$

ϕ is infinite dimensional, but $\phi(x)^\top \phi(x')$ is computed in constant time!!

Examples of Kernel: Polynomial Kernel of any degree

For $x, z \in [0, 1]$ and $0 < \alpha < 1$

$$K(x, z) = \frac{1}{1 - \alpha^2 xz}$$

Proof

$$\frac{1}{1 - \alpha^2 xz} = \sum_{k=0}^{\infty} (\alpha^2 xz)^k = \phi(x)^\top \phi(z)$$

with $\phi : \mathbb{R} \to \mathbb{R}^{\mathbb{N}}$ defined as

$$\phi(x) = (1, \alpha x, \alpha^2 x^2, \alpha^3 x^3, \dots)$$

ϕ is infinite dimensional, but $\phi(x)^\top \phi(x')$ is computed in constant time!!

For $x, z \in \mathbb{R}^d$ it is defined as

$$K(x, z) = \frac{1}{1 - \alpha^2 x^\top z}$$

Examples of Kernel: Gaussian Kernel

For $X = \mathbb{R}$ and $\gamma > 0$ consider

$$K(x, x') = e^{-|x-\bar{x}|^2\gamma}$$

Proof Let

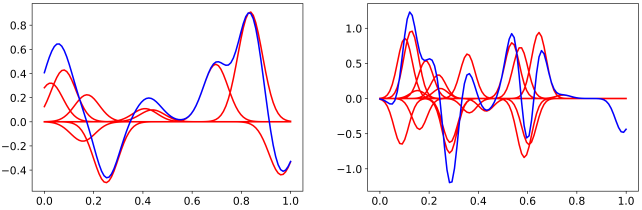
$$\varphi_j(x) = x^{j-1}e^{-x^2\gamma}\sqrt{\frac{(2\gamma)^{(j-1)}}{(j-1)!}}, \qquad j = 2, \dots, \infty$$

with $\varphi_1(x) = e^{-x^2\gamma}$.
Then

$$\begin{aligned}
\sum_{j=1}^{\infty} \varphi_j(x)\varphi_j(\bar{x}) &= \sum_{j=1}^{\infty} x^{j-1}e^{-x^2\gamma}\sqrt{\frac{(2\gamma)^{j-1}}{(j-1)!}}\bar{x}^{j-1}e^{-\bar{x}^2\gamma}\sqrt{\frac{(2\gamma)^{j-1}}{(j-1)!}} \\
&= e^{-x^2\gamma}e^{-\bar{x}^2\gamma}\sum_{j=1}^{\infty}\frac{(2\gamma)^{j-1}}{(j-1)!}(x\bar{x})^{j-1} = e^{-x^2\gamma}e^{-\bar{x}^2\gamma}e^{2x\bar{x}^2\gamma} \\
&= e^{-|x-\bar{x}|^2\gamma}
\end{aligned}$$

A key result

Functions defind by Gaussian kernels with large and small widths.



Kernel - Characterization

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *Kernel* if it behaves like an inner product that is

1. it is symmetric

$$K(x, z) = K(z, x) \quad \text{for all } x, z \in \mathbb{R}^d$$

2. it is positive definite (p.d.).

Kernel - Characterization

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *Kernel* if it behaves like an inner product that is

1. it is symmetric

$$K(x, z) = K(z, x) \quad \text{for all } x, z \in \mathbb{R}^d$$

2. it is positive definite (p.d.). For any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ define $\hat{K}$ as $(\hat{K})_{ij} = K(x_i, x_j)$.

Kernel - Characterization

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *Kernel* if it behaves like an inner product that is

1. it is symmetric

$$K(x,z) = K(z,x) \quad \text{for all } x, z \in \mathbb{R}^d$$

2. it is positive definite (p.d.). For any $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathbb{R}^d$ define $\hat{K}$ as $(\hat{K})_{ij} = K(x_i, x_j)$.

$$K \text{ is p.d.} \quad \text{iff} \quad \hat{K} \text{ is p.d. for any } n \in \mathbb{N}, x_1, \dots, x_n \in \mathbb{R}^d$$

The first is easy to check, the second is quite difficult!

Kernel properties

Let $K_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_3 : \mathbb{R}^t \times \mathbb{R}^t$ be Kernels and $x, x' \in \mathbb{R}^d$, $z, z' \in \mathbb{R}^t$ and $\alpha, \beta > 0$ then the following are Kernels too

1. $\alpha K_1(x, x') + \beta K_2(x, x')$
2. $K_1(x, x')K_2(x, x')$
3. $p(K_1(x, x'))$ for any p a function whose polynomial expansion has only non-negative coefficients
4. $f(x)K_1(x, x')f(x')$ for any $f : \mathbb{R}^d \to \mathbb{R}$
5. $\frac{K_1(x,x')}{\sqrt{K_1(x,x)K_1(x',x')}}$
6. $K_3(\psi(x), \psi(x))$ for any $\psi : \mathbb{R}^d \to \mathbb{R}^t$
7. $\alpha K_1(x, x') + \beta K_3(z, z')$
8. $K_1(x, x')K_3(z, z')$

Gaussian Kernel

Let $x, x' \in \mathbb{R}^d$ and $\sigma > 0$, the gaussian kernel is

$$K(x, x') = e^{-\frac{1}{2\sigma^2}\|x-x'\|^2}$$

Gaussian Kernel

Let $x, x' \in \mathbb{R}^d$ and $\sigma > 0$, the gaussian kernel is

$$K(x, x') = e^{-\frac{1}{2\sigma^2}\|x-x'\|^2}$$

Proof $K_1(x, x') = \frac{x^\top x'}{2\sigma^2}$ is a Kernel by Point 1

Gaussian Kernel

Let $x, x' \in \mathbb{R}^d$ and $\sigma > 0$, the gaussian kernel is

$$K(x, x') = e^{-\frac{1}{2\sigma^2}\|x-x'\|^2}$$

Proof $K_1(x, x') = \frac{x^\top x'}{2\sigma^2}$ is a Kernel by Point 1
Let $e^t = \sum_{k=1}^{\infty} \frac{t^k}{k!}$ has polynomial expansion with positive coefficients therefore the following is a Kernel (Point 3)

$$K_2(x, x') = e^{K_1(x,x')} = e^{\frac{x^\top x'}{2\sigma^2}}$$

is a Kernel.

Gaussian Kernel

Let $x, x' \in \mathbb{R}^d$ and $\sigma > 0$, the gaussian kernel is

$$K(x, x') = e^{-\frac{1}{2\sigma^2}\|x-x'\|^2}$$

Proof $K_1(x, x') = \frac{x^\top x'}{2\sigma^2}$ is a Kernel by Point 1
Let $e^t = \sum_{k=1}^{\infty} \frac{t^k}{k!}$ has polynomial expansion with positive coefficients therefore the following is a Kernel (Point 3)

$$K_2(x, x') = e^{K_1(x,x')} = e^{\frac{x^\top x'}{2\sigma^2}}$$

is a Kernel.
Let define $f(x) = e^{-\frac{x^\top x}{2\sigma^2}}$ then the following is a Kernel (Point 4)

$$K_3(x, x') = f(x) K_2(x, x') f(x')$$

Gaussian Kernel

Let $x, x' \in \mathbb{R}^d$ and $\sigma > 0$, the gaussian kernel is

$$K(x, x') = e^{-\frac{1}{2\sigma^2}\|x-x'\|^2}$$

Proof $K_1(x, x') = \frac{x^\top x'}{2\sigma^2}$ is a Kernel by Point 1
Let $e^t = \sum_{k=1}^{\infty} \frac{t^k}{k!}$ has polynomial expansion with positive coefficients therefore the following is a Kernel (Point 3)

$$K_2(x, x') = e^{K_1(x,x')} = e^{\frac{x^\top x'}{2\sigma^2}}$$

is a Kernel.
Let define $f(x) = e^{-\frac{x^\top x}{2\sigma^2}}$ then the following is a Kernel (Point 4)

$$K_3(x, x') = f(x) K_2(x, x') f(x')$$

But $K_3 = K$ indeed

$$K_3(x, x') = f(x) e^{\frac{x^\top x'}{\sigma^2}} f(x') = e^{-\frac{x^\top x + x'^\top x' - 2x^\top x'}{2\sigma^2}} = e^{\frac{-\|x-x'\|^2}{2\sigma^2}} = K(x, x')$$

Wrapping up

In this class we discussed how to deal with high dimensional non linear problems (feature maps and kernels). We also introduced the Representer Theorem.

Next class

Definitely what this course is not about... Neural Networks!