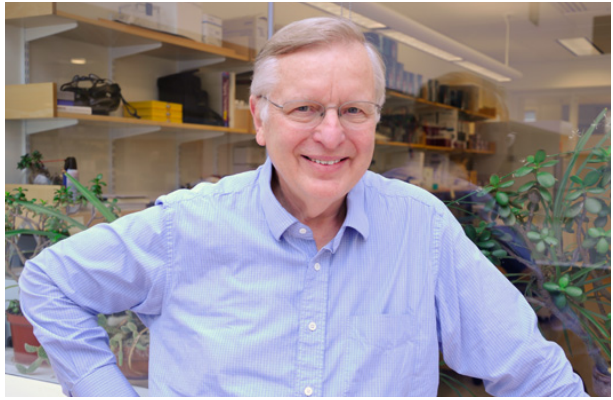
UniGe | MaLGa

MLSS 2019 London

Kernels
Part I: One weird trick

Lorenzo Rosasco
MaLGa- Machine learning Genova Center
Universit'a di Genova
MIT
IIT



Patrick Henry Winston (February 5, 1943 - July 19, 2019)

Once upon a time

Machine Learning Summer School (MLSS), Berder Island 2004

Sneak peak in the past (future?)

Outline

Warm up

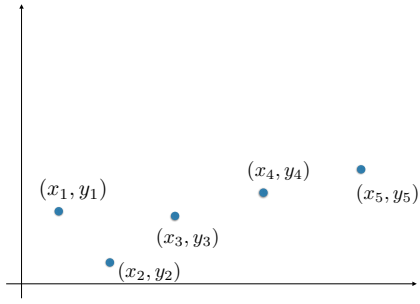
Linear models recap

Non linear features

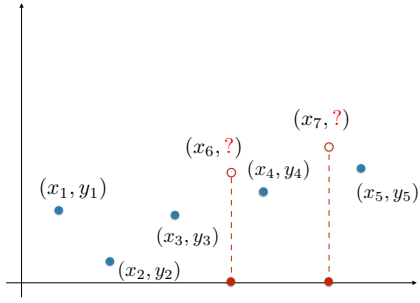
Representer theorem

Kernels

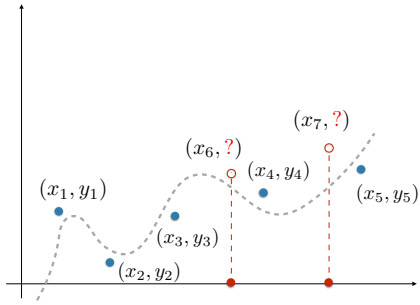
What's learning



What's learning



What's learning



Learning is about inference

Learning from data

Problem

$$\textbf{given } \{(x_1, y_1), \ldots, (x_n, y_n)\} \textbf{ find } \widehat{f}(x_{new}) \sim y_{new}$$

Note: x_i's and y_i's are **random**

Statistical Learning

We would like to minimize

Test error

$$\mathbb{E}[\ell(Y, f(X))]$$

We can minimize

Training error

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(x_i))$$

On which function class?

Outline

Warm up

Linear models recap

Non linear features

Representer theorem

Kernels

Linear functions

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

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i)$$

Linear least squares

My favorite

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2$$

Linear least squares

My favorite

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2$$

that is

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{y} - \widehat{X} w\|^2,$$

with $\widehat{X} \in \mathbb{R}^{n \times d}$ and $\widehat{y} \in \mathbb{R}^n$.

Linear least squares solution

From the optimality condition

$$\nabla \frac{1}{n} \|\widehat{y} - \widehat{X}w\|^2 = 0$$

we can derive the equations

$$\widehat{X}^\top \widehat{X} w = \widehat{X}^\top \widehat{y} \qquad \Leftrightarrow \qquad \widehat{w} = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top \widehat{y}.$$

Linear least squares solution

From the optimality condition

$$\nabla \frac{1}{n} \|\widehat{y} - \widehat{X}w\|^2 = 0$$

we can derive the equations

$$\widehat{X}^\top \widehat{X} w = \widehat{X}^\top \widehat{y} \qquad \Leftrightarrow \qquad \widehat{w} = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top \widehat{y}.$$

- **Non invertible matrices??**

Ridge regression

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{y} - \widehat{X}w\|^2 + \lambda \|w\|^2,$$

...

$$(\widehat{X}^\top \widehat{X} + \lambda n I) w = \widehat{X}^\top \widehat{y} \qquad \Leftrightarrow \qquad \widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda n I)^{-1} \widehat{X}^\top \widehat{y}.$$

Requires: time $O(nd^2 + d^3)$ space $O(nd \vee d^2)$.

Beyond linear models??

Side note

Linear models can be enough.

$$\widehat{X}w = \widehat{y}$$

A solution always exists as long as $n < d$.

Outline

Warm up

Linear models recap

Non linear features

Representer theorem

Kernels

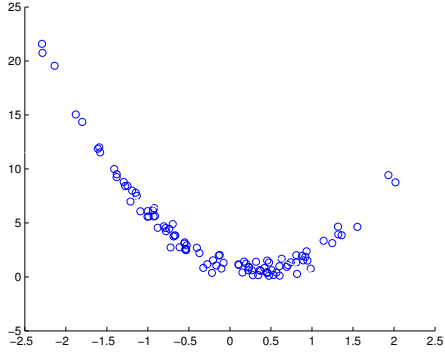
Linear combination of features

$$f(x) = \sum_{j=1}^{p} w^j \phi_j(x)$$

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \sum_{j=1}^{p} w^j \phi_j(x_i))$$

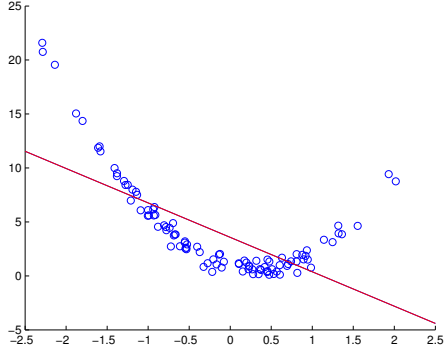
Example

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$



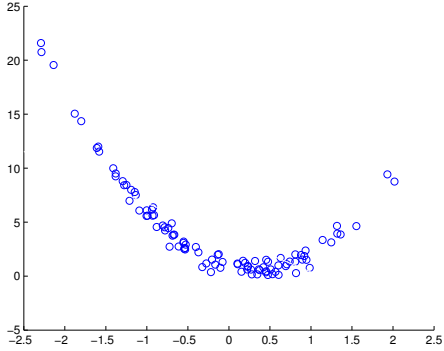
Example

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$



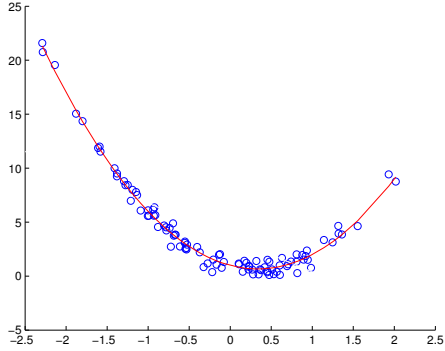
Example (cont.)

$$f(x) = w_1 x^2 + w_2 x + w_3$$



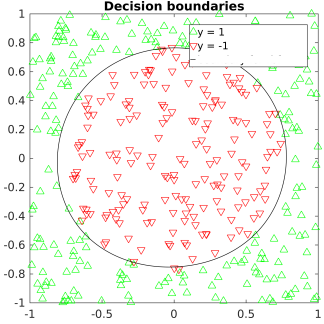
Example (cont.)

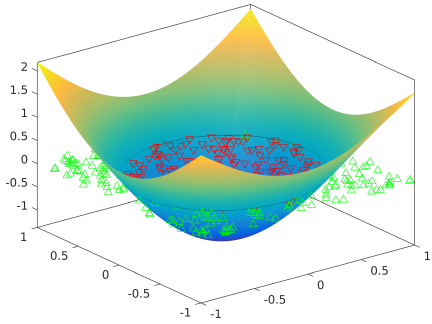
$$f(x) = w_1 x^2 + w_2 x + w_3$$



Another example

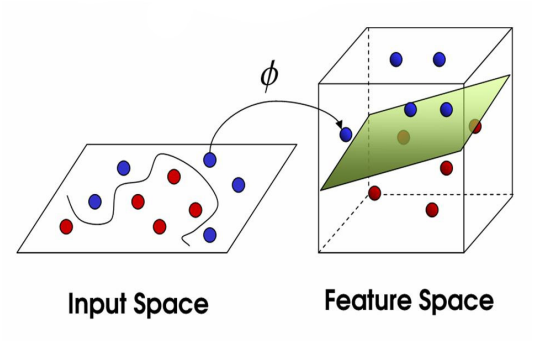
$$\mathbf{f}((\mathbf{x}^1, \mathbf{x}^2)) = \mathbf{w}_1(\mathbf{x}^1)^2 + \sqrt{2}\mathbf{w}_2\mathbf{x}^1\mathbf{x}^2 + \mathbf{w}_3(\mathbf{x}^2)^2$$





Geometric view

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}), \qquad \Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_\mathbf{p}(\mathbf{x}))$$



Feature map

$$\Phi : \mathbf{X} \to \mathcal{F}$$

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$$

Non linear least squares

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top \Phi(x_i))^2$$

Non linear least squares

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top \Phi(x_i))^2$$

that is

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{y} - \widehat{\Phi} w\|^2,$$

with $\widehat{\Phi} \in \mathbb{R}^{n \times p}$ and $(\widehat{\Phi})_{ij} = \phi_j(x_i)$

Non linear least squares solution

From the optimality condition

$$\nabla_w \frac{1}{n} \|\widehat{y} - \widehat{\Phi} w\|^2 = 0$$

$$\widehat{\Phi}^\top \widehat{\Phi} w = \widehat{\Phi}^\top \widehat{y} \qquad \Leftrightarrow \qquad \widehat{w} = (\widehat{\Phi}^\top \widehat{\Phi})^{-1} \widehat{\Phi}^\top \widehat{y}.$$

For ridge regression

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

Requires: time $O(np^2 + p^3)$ space $O(np \vee p^2)$.

Complaints

- **Better computations??**
- Infinite features??

Outline

Warm up

Linear models recap

Non linear features

Representer theorem

Kernels

Key for magic

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

Key for magic

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

- **Why is this true?**
- Why do we care?

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

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

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = (V\Sigma U^\top U\Sigma V^\top + \lambda n I)^{-1} V\Sigma U^\top \widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = (V\Sigma^2 V^\top + \lambda n I)^{-1} V\Sigma U^\top \widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = V(\Sigma^2 + \lambda n I)^{-1}\Sigma U^\top \widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = V\Sigma(\Sigma^2 + \lambda n I)^{-1} U^\top \widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = V\Sigma U^\top U(\Sigma^2 + \lambda n I)^{-1} U^\top \widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = V\Sigma U^\top (U\Sigma^2 U^\top + \lambda n U U^\top)^{-1} \widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

$$w = V\Sigma U^\top (U\Sigma V^\top V\Sigma U^\top + \lambda n I^\top)^{-1}\widehat{y}$$

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Sketch of the Proof

Consider the SVD[1]

$$\widehat{\Phi} = U\Sigma V^\top$$

Then

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

[1] $U^\top U = I_{n\times n}$, $V^\top V = I_{n\times n}$ $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geqslant 0$. (Note that $\Sigma = \Sigma^\top$)

Key for magic

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

- Why is this true?
- **Why do we care?**

Representer theorem

$$\widehat{w}_\lambda = \widehat{\Phi}^\top \underbrace{(\widehat{\Phi}\widehat{\Phi}^\top + \lambda n I)^{-1}\widehat{y}}_{c \in \mathbb{R}^n}$$

$$\Leftrightarrow$$

$$\widehat{w}_\lambda = \widehat{\Phi}^\top c = \sum_{i=1}^{n} \Phi(x_i) c_i \qquad\qquad c = (\underbrace{\widehat{\Phi}\widehat{\Phi}^\top}_{n \times n} + \lambda n I)^{-1}\widehat{y}$$

Requires: time $O(pn^2 + n^3)$ space $O(n^2)$.

Representer theorem (cont.)

$$\widehat{f}_\lambda(x) = \Phi(x)^\top \widehat{w}_\lambda = \sum_{i=1}^{n} \Phi(x)^\top \Phi(x_i) c_i, \qquad c = (\widehat{\Phi}\widehat{\Phi}^\top + \lambda n I)^{-1} \widehat{y}$$

$$(\widehat{\Phi}\widehat{\Phi}^\top)_{ij} = \Phi(x_j)^\top \Phi(x_i)$$

All depends just on

$$\Phi(x)^\top \Phi(x') = \sum_{j=1}^{p} \phi_j(x) \phi_j(x')$$

...

Kernel time!

Outline

Warm up

Linear models recap

Non linear features

Representer theorem

Kernels

Meet the kernel

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}') = \sum_{j=1}^{p} \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$$

k is often known in closed form

Examples: Linear Kernel

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

$$k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z}$$

Proof

$$k(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^\top \Phi(\mathbf{z})$$

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

$$\Phi(\mathbf{x}) = \mathbf{x}.$$

Examples: Affine Kernel

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

$$\mathbf{k}(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z} + \alpha^2$$

Proof

$$\mathbf{k}(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^\top \Phi(\mathbf{z})$$

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

$$\Phi(\mathbf{x}) = (\mathbf{x}, \alpha).$$

Examples: 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

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

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

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

Examples: 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

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

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

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

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

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

Meet the kernel

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}') = \sum_{j=1}^{p} \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$$

Can we take $p = \infty$**?!?**

Examples: 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

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

with

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

Examples: 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

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

with

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

Φ is infinite dimensional, but $k(x, z)$ is computed in finite time!!

Examples: 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

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

with

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

Φ is infinite dimensional, but $k(x, z)$ is computed in finite time!!

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

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

Examples: Gaussian Kernel

For $x, z \in [0,1]$ and $\gamma > 0$

$$k(x,z) = e^{-|x-z|^2\gamma}$$

Proof

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

with

$$\phi_1(x) = 1 \qquad \phi_j(x) = x^{j-1}e^{-x^2\gamma}\sqrt{\frac{(2\gamma)^{(j-1)}}{(j-1)!}}, \qquad j = 2, \ldots, \infty$$

Examples: Gaussian Kernel

For $x, z \in [0,1]$ and $\gamma > 0$

$$k(x,z) = e^{-|x-z|^2\gamma}$$

Proof

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

with

$$\phi_1(x) = 1 \qquad \phi_j(x) = x^{j-1}e^{-x^2\gamma}\sqrt{\frac{(2\gamma)^{(j-1)}}{(j-1)!}}, \qquad j = 2, \ldots, \infty$$

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

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

Meet the kernel

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}') = \sum_{j=1}^{p} \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$$

Can we take $p = \infty$**?!?**

YES!

From learning with features…

$$\widehat{f}_\lambda(x) = \widehat{w}_\lambda^\top \Phi(x) \qquad \widehat{w}_\lambda = (\widehat{\Phi}^\top \widehat{\Phi} + \lambda n I)^{-1} \widehat{\Phi}^\top \widehat{y}$$

All depends just on

$$(\widehat{\Phi})_{ij} = \Phi(x_i)^j$$

Requires: time $O(np^2 + p^3)$ space $O(np \vee p^2)$.

...to learning with kernels

$$\widehat{f}_\lambda(x) = \sum_{i=1}^{n} k(x, x_i) c_i, \qquad c = (\widehat{K} + \lambda n I)^{-1} \widehat{y}$$

$$(\widehat{K})_{ij} = k(x_i, x_j)$$

All depends just on

$$k(x, x')$$

Requires[2]: time $O(c_k n^2 + n^3)$ space $O(n^2)$.

[2] c_k cost of evaluating k

Kernels for all

- kernels on probability distributions
- kernels on strings
- kernels on functions
- kernels on groups
- kernels graphs
- ...

See later.

Not just least squares

- General loss (an exercise for you!)
- Kernel PCA
- Kernel ICA
- Kernel CCA
- Kernel K-means (also an exercise for you!)
- ...

Kernels and general loss

Representer Theorem for general loss

Let

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

then there is $c = (c_1, \dots, c_n)$ such that

$$\widehat{w}_\lambda = \widehat{\Phi}^\top c$$

Hint[3]

---

[3] Consider $\widehat{W} = \{\widehat{\Phi}^\top c \mid c \in \mathbb{R}^n\} \subset \mathbb{R}^p$. so that $\forall w \in \mathbb{R}^p$

$$w = \widehat{w} + w_\perp$$

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

From K-means...

$$\min_{m_1 \in \mathbb{R}^d, \dots, m_k \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,K} \|x_i - m_j\|^2$$

Lloyd's algorithm

- assignment: $C_j = \{x_i, \dots, x_n \mid \|x_i - m_j\| \leqslant \|x_i - m_\ell\|\}$
- uptdate: $m_j = \frac{1}{\#C_j} \sum_{x \in C_j} x$

...to kernel K-means

$$\min_{m_1 \in \mathbb{R}^p, \dots, m_k \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,K} \|\Phi(x_i) - m_j\|^2, \qquad p \leqslant \infty$$

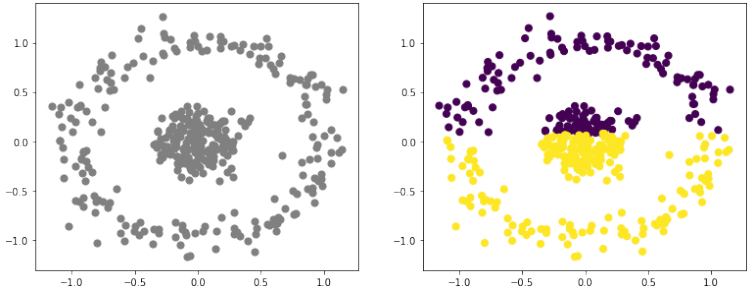
Show that all depends just on

$$k(x, x') = \Phi(x)^\top \Phi(x')$$

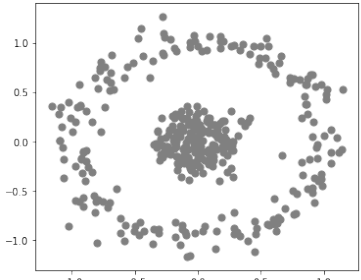
Hint[4]

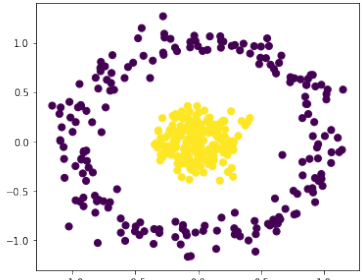
[4] Note that $\|\Phi(x) - \Phi(x')\|^2$ hence $\|\Phi(x) - m_j\|^2$ can be written using the kernel.

K-means vs kernel K-means



K-means vs kernel K-means





Summing up

- Kernels and non linear models
- Kernels and feature maps
- Kernels and nonparametrics

What more?