

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



MLSS 2004 - Berder Island

Machine Learning Summer School (MLSS), Berder Island 2004

Sneak peak in the past (future?)



Outline

Warm up

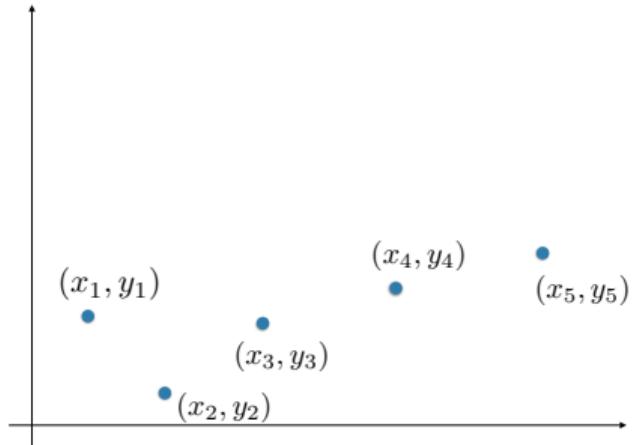
Linear models recap

Non linear features

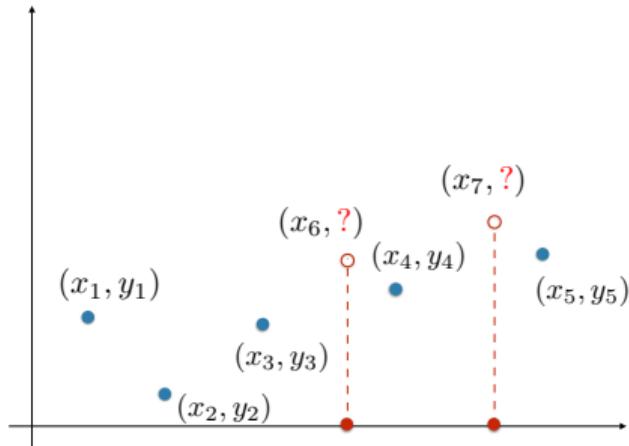
Representer theorem

Kernels

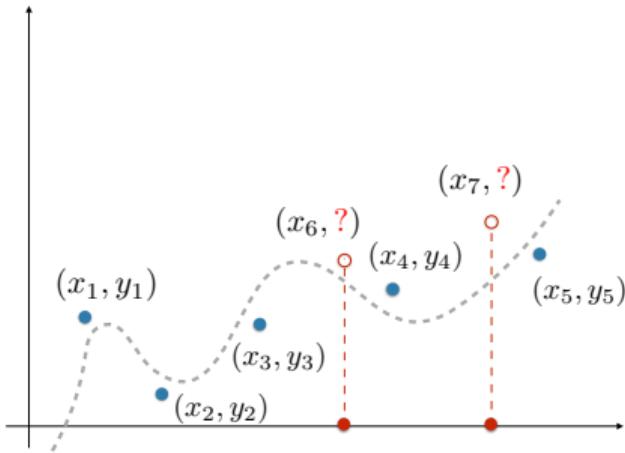
What's learning



What's learning



What's learning



Learning is about inference

Learning from data

Problem

given $\{(x_1, y_1), \dots, (x_n, y_n)\}$ **find** $\hat{f}(x_{\text{new}}) \sim y_{\text{new}}$

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

Statistical Learning

We would like to minimize
Test error

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

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

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

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

Linear least squares

My favorite

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

Linear least squares

My favorite

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

that is

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

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

Linear least squares solution

From the optimality condition

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

we can derive the equations

$$\hat{\mathbf{X}}^\top \hat{\mathbf{X}} \mathbf{w} = \hat{\mathbf{X}}^\top \hat{\mathbf{y}} \quad \Leftrightarrow \quad \hat{\mathbf{w}} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{y}}.$$

Linear least squares solution

From the optimality condition

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

we can derive the equations

$$\hat{\mathbf{X}}^\top \hat{\mathbf{X}}\mathbf{w} = \hat{\mathbf{X}}^\top \hat{\mathbf{y}} \quad \Leftrightarrow \quad \hat{\mathbf{w}} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{y}}.$$

- Non invertible matrices??

Ridge regression

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

...

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

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

Beyond linear models??

Outline

Warm up

Linear models recap

Non linear features

Representer theorem

Kernels

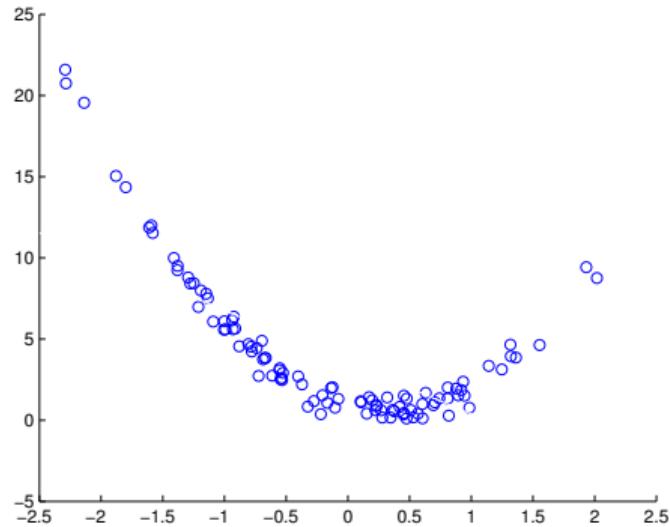
Linear combination of features

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

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

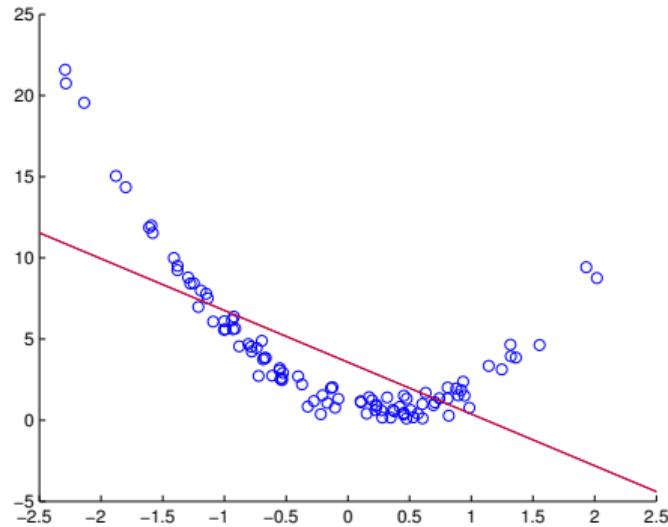
Example

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



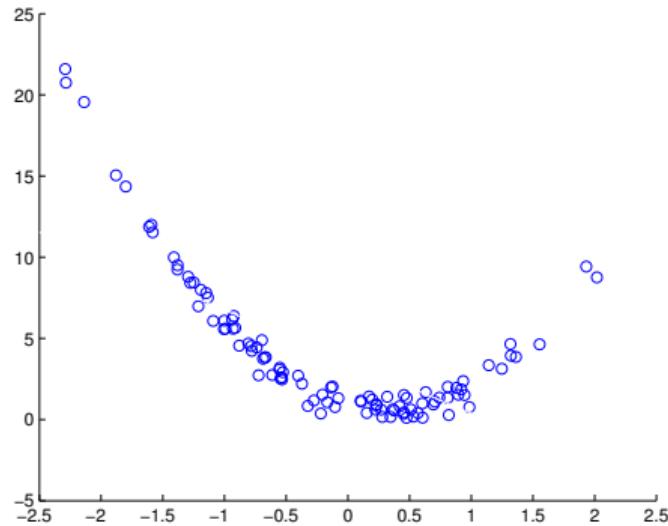
Example

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



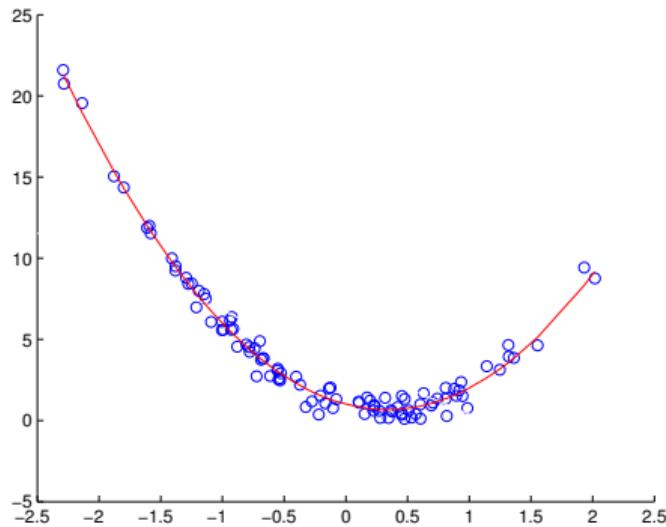
Example (cont.)

$$f(x) = w_1x^2 + w_2x + w_3$$



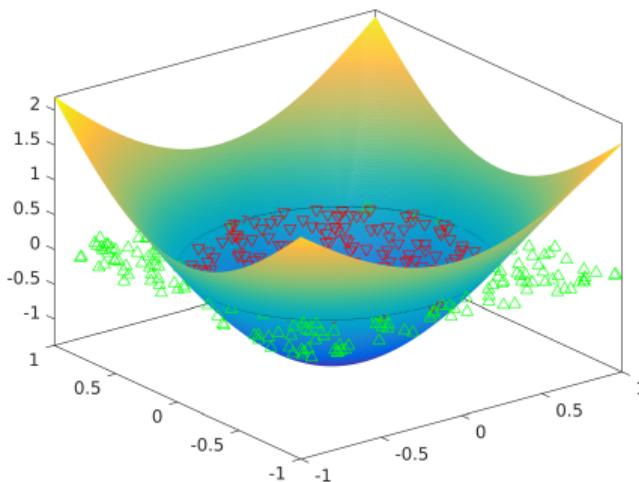
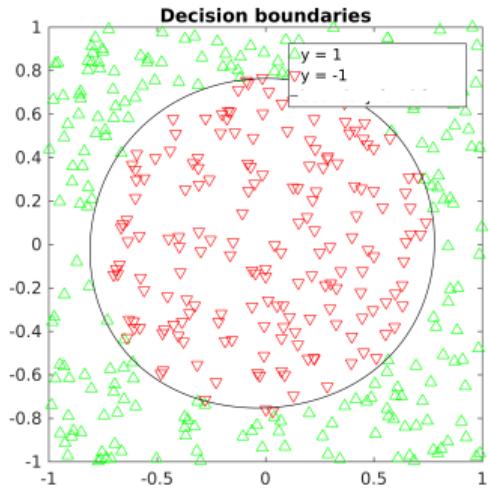
Example (cont.)

$$f(x) = w_1x^2 + w_2x + w_3$$



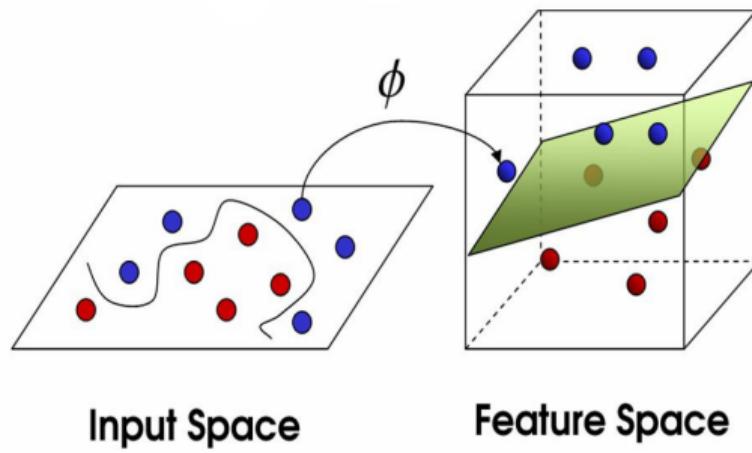
Another example

$$f((x^1, x^2)) = w_1(x^1)^2 + \sqrt{2}w_2x^1x^2 + w_3(x^2)^2$$



Geometric view

$$f(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}), \quad \Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_p(\mathbf{x}))$$



Feature map

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

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

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

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{w}^\top \Phi(\mathbf{x}_i))^2$$

that is

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

with $\hat{\Phi} \in \mathbb{R}^{n \times p}$ and $(\hat{\Phi})_{ij} = \phi_j(\mathbf{x}_i)$

Non linear least squares solution

From the optimality condition

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

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

For ridge regression

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

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

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

Key for magic

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

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

Sketch of the Proof

Consider the SVD¹

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

Then

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Sketch of the Proof

Consider the SVD¹

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

Then

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

Key for magic

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

► Why is this true?

► Why do we care?

Representer theorem

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

\Leftrightarrow

$$\hat{\mathbf{w}}_\lambda = \hat{\Phi}^\top \mathbf{c} = \sum_{i=1}^n \Phi(\mathbf{x}_i) \mathbf{c}_i \quad \mathbf{c} = \underbrace{(\hat{\Phi}\hat{\Phi}^\top + \lambda n\mathbf{I})^{-1}\hat{\mathbf{y}}}_{n \times n}$$

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

Representer theorem (cont.)

$$\hat{f}_\lambda(\mathbf{x}) = \Phi(\mathbf{x})^\top \hat{\mathbf{w}}_\lambda = \sum_{i=1}^n \Phi(\mathbf{x})^\top \Phi(\mathbf{x}_i) c_i, \quad \mathbf{c} = (\hat{\Phi} \hat{\Phi}^\top + \lambda n I)^{-1} \hat{\mathbf{y}}$$

$$(\hat{\Phi} \hat{\Phi}^\top)_{ij} = \Phi(\mathbf{x}_j)^\top \Phi(\mathbf{x}_i)$$

All depends just on

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

...

Kernel time!

Outline

Warm up

Linear models recap

Non linear features

Representer theorem

Kernels

Meet the kernel

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

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

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

Examples: Affine Kernel

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

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

Proof

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

with $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+1}$ 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

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

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

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

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

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

with $q_{p,k} = \frac{p!}{k!(p-k)!}$ and $\Phi : \mathbb{R} \rightarrow \mathbb{R}^{p+1}$ 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

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

Meet the kernel

$$k(x, x') = \Phi(x)^\top \Phi(x') = \sum_{i=1}^p \phi_i(x)\phi_i(x')$$

Can we take $p = \infty$?!

Examples: Polynomial Kernel of any degree

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

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

Proof

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

with

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

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

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

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

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

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

$$\varphi_1(x) = 1 \quad \varphi_j(x) = x^{j-1} e^{-x^2\gamma} \sqrt{\frac{(2\gamma)^{(j-1)}}{(j-1)!}}, \quad j = 2, \dots, \infty$$

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

$$\varphi_1(x) = 1 \quad \varphi_j(x) = x^{j-1} e^{-x^2\gamma} \sqrt{\frac{(2\gamma)^{(j-1)}}{(j-1)!}}, \quad j = 2, \dots, \infty$$

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

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

Meet the kernel

$$k(x, x') = \Phi(x)^\top \Phi(x') = \sum_{i=1}^p \phi_i(x)\phi_i(x')$$

Can we take $p = \infty$?!

YES!

From learning with features...

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

All depends just on

$$(\hat{\Phi})_{ij} = \Phi(\mathbf{x}_i)^j$$

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

... to learning with kernels

$$\widehat{f}_\lambda(\mathbf{x}) = \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i) c_i, \quad \mathbf{c} = (\widehat{\mathbf{K}} + \lambda n \mathbf{I})^{-1} \widehat{\mathbf{y}}$$
$$(\widehat{\mathbf{K}})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

All depends just on

$$k(\mathbf{x}, \mathbf{x}')$$

Requires²: time $O(c_k n^2 + n^3)$ space $O(n^2)$.

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

$$\hat{\mathbf{w}}_\lambda = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{y}_i, \Phi(\mathbf{x}_i)^\top \mathbf{w}) + \lambda \|\mathbf{w}\|^2$$

then there is $\mathbf{a} = (a_1, \dots, a_n)$ such that

$$\hat{\mathbf{w}}_\lambda = \hat{\Phi}^\top \mathbf{a}$$

Hint³

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

$$\mathbf{w} = \hat{\mathbf{w}} + \mathbf{w}_\perp$$

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

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: $C_j = \{x_i, \dots, x_n \mid \|x_i - m_j\| \leq \|x_i - m_\ell\|\}$
- ▶ update: $m_j = \frac{1}{\#C_j} \sum_{x \in C_j} x$

...to kernel K-means

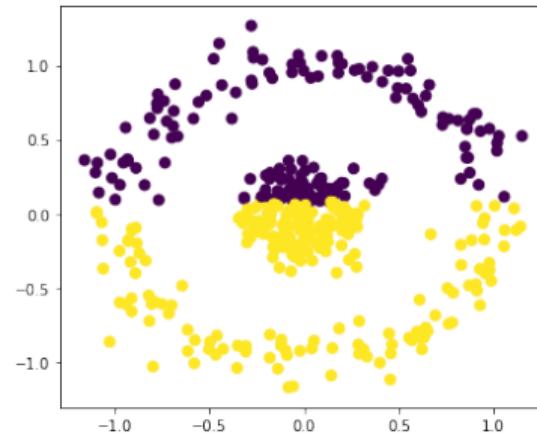
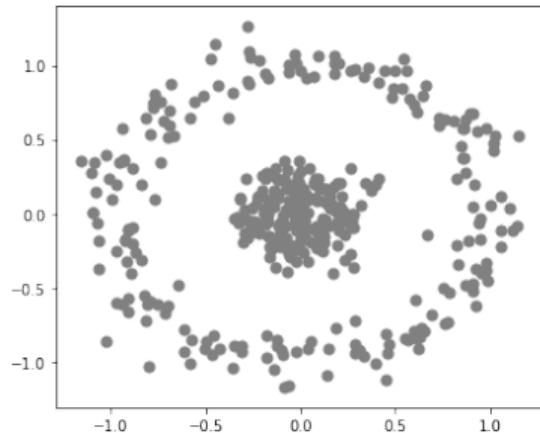
$$\min_{m_1 \in \mathbb{R}^p, \dots, m_k \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \min_{j=1, \dots, K} \|\Phi(x_i) - m_j\|^2, \quad p \leq \infty$$

Show that all depends just on

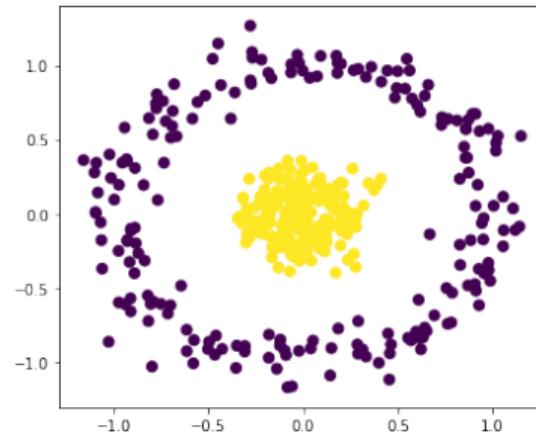
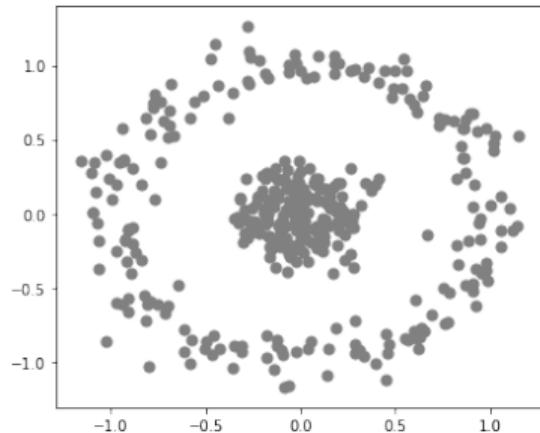
$$k(x, x') = \Phi(x)^\top \Phi(x')$$

Hint⁴

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?