MIT 9.520/6.860, Fall 2018
Statistical Learning Theory and Applications

Class 04: Features and Kernels

Lorenzo Rosasco

Linear functions

Let $\mathcal{H}_{\text{lin}}$ be the space of linear functions

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

- $f \leftrightarrow w$ is one to one,
- inner product $\left\langle f, \bar{f} \right\rangle_{\mathcal{H}} := w^\top \bar{w}$,
- norm/metric $\left\| f - \bar{f} \right\|_{\mathcal{H}} := \|w - \bar{w}\|$.

An observation

Function norm controls point-wise convergence.

Since

$$|f(x) - \bar{f}(x)| \leq \|x\| \|w - \bar{w}\|, \qquad \forall x \in X$$

then

$$w_j \to w \quad \Rightarrow f_j(x) \to f(x), \qquad \forall x \in X.$$

ERM

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda\|w\|^2, \qquad \lambda \geq 0$$

- $\lambda \to 0$ ordinary least squares (bias to minimal norm),
- $\lambda > 0$ ridge regression (stable).

Computations

Let $Xn \in \mathbb{R}^{nd}$ and $\widehat{Y} \in \mathbb{R}^n$.

The ridge regression solution is

$$\widehat{w}^\lambda = (Xn^\top Xn + n\lambda I)^{-1} Xn^\top \widehat{Y} \qquad \text{time} \quad O(nd^2 \vee d^3) \quad \text{mem.} \quad O(nd \vee d^2)$$

but also

$$\widehat{w}^\lambda = Xn^\top (Xn Xn^\top + n\lambda I)^{-1} \widehat{Y} \qquad \text{time} \quad O(dn^2 \vee n^3) \quad \text{mem.} \quad O(nd \vee n^2)$$

Representer theorem in disguise

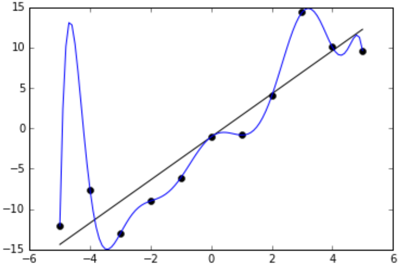
We noted that

$$\widehat{w}^{\lambda} = Xn^{\top} c = \sum_{i=1}^{n} x_i c_i \qquad \Leftrightarrow \qquad \hat{f}^{\lambda}(x) = \sum_{i=1}^{n} x^{\top} x_i c_i,$$

$$c = (XnXn^{\top} + n\lambda I)^{-1} \widehat{Y}, \qquad (XnXn^{\top})_{ij} = x_i^{\top} x_j.$$

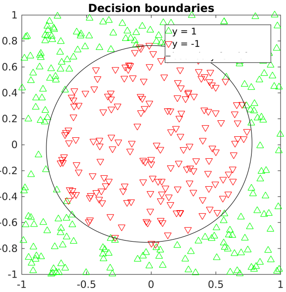
Limits of linear functions

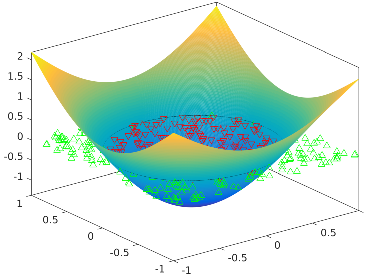
Regression



Limits of linear functions

Classification





Nonlinear functions

Two main possibilities:

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

where Φ is a non linear map.

- The former choice leads to linear spaces of functions[1].
- The latter choice can be iterated

$$f(x) = \Phi(w_L^\top \Phi(w_{L-1}^\top \dots \Phi(w_1^\top x))).$$

[1] The spaces are linear, NOT the functions!

Features and feature maps

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

where $\Phi : X \to \mathbb{R}^p$

$$\Phi(x) = (\varphi_1(x), \dots, \varphi_p(x))^\top$$

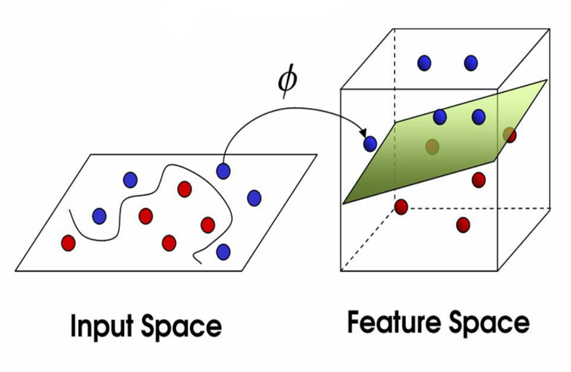
and $\varphi_j : X \to \mathbb{R}$, for $j = 1, \dots, p$.

- X need not be $\mathbb{R}^d$.
- We can also write

$$f(x) = \sum_{i=1}^{p} w^j \varphi_j(x).$$

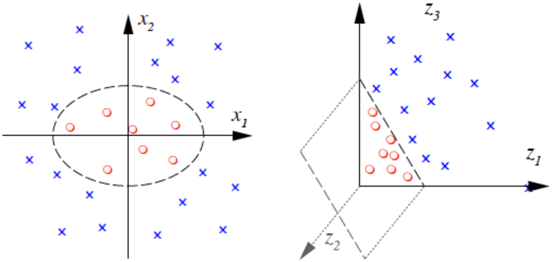
Geometric view

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



An example

$$\Phi : R^2 \to R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$



More examples

The equation

$$f(x) = w^\top \Phi(x) = \sum_{i=1}^{p} w^j \varphi_j(x)$$

suggests to think of features as some form of basis.

Indeed we can consider

- Fourier basis,
- wave-lets + their variations,
- ...

And even more examples

Any set of functions

$$\varphi_j : X \to \mathbb{R}, \qquad j = 1, \dots, p$$

can be considered.

Feature design/engineering

- vision: SIFT, HOG
- audio: MFCC
- ...

Nonlinear functions using features

Let $\mathcal{H}_\Phi$ be the space of linear functions

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

- $f \leftrightarrow w$ is one to one, if $(\varphi_j)_j$ are lin. indip.
- inner product $\left\langle f, \bar{f} \right\rangle_{\mathcal{H}_\Phi} := w^\top \bar{w}$,
- norm/metric $\left\| f - \bar{f} \right\|_{\mathcal{H}_\Phi} := \|w - \bar{w}\|$.

In this case

$$|f(x) - \bar{f}(x)| \leq \|\Phi(x)\| \|w - \bar{w}\|, \qquad \forall x \in X.$$

Back to ERM

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top \Phi(x_i))^2 + \lambda \|w\|^2, \qquad \lambda \geq 0,$$

Equivalent to,

$$\min_{f \in \mathcal{H}_\Phi} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_\Phi}, \qquad \lambda \geq 0.$$

Computations using features

Let $\widehat{\Phi} \in \mathbb{R}^{np}$ with

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

The ridge regression solution is

$$\widehat{w}^{\lambda} = (\widehat{\Phi}^{\top}\widehat{\Phi} + n\lambda I)^{-1}\widehat{\Phi}^{\top}\widehat{Y} \qquad \text{time} \quad O(np^2 \vee p^3) \quad \text{mem.} \quad O(np \vee p^2),$$

but also

$$\widehat{w}^{\lambda} = \widehat{\Phi}^{\top}(\widehat{\Phi}\widehat{\Phi}^{\top} + n\lambda I)^{-1}\widehat{Y} \qquad \text{time} \quad O(pn^2 \vee n^3) \quad \text{mem.} \quad O(np \vee n^2).$$

Representer theorem a little less in disguise

Analogously to before

$$\widehat{w}^\lambda = \widehat{\Phi}^\top c = \sum_{i=1}^n \Phi(x_i) c_i \qquad \Leftrightarrow \qquad \hat{f}^\lambda(x) = \sum_{i=1}^n \Phi(x)^\top \Phi(x_i) c_i$$

$$c = (\widehat{\Phi}\widehat{\Phi}^\top + \lambda I)^{-1} \widehat{Y}, \qquad (\widehat{\Phi}\widehat{\Phi}^\top)_{ij} = \Phi(x_i)^\top \Phi(x_j)$$

$$\Phi(x)^\top \Phi(\bar{x}) = \sum_{s=1}^p \varphi_s(x) \varphi_s(\bar{x}).$$

Unleash the features

- Can we consider linearly dependent features?

- Can we consider $p = \infty$?

An observation

For $X = \mathbb{R}$ consider

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

with $\varphi_1(x) = 1$.

Then

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

From features to kernels

$$\Phi(x)^\top \Phi(\bar{x}) = \sum_{j=1}^{\infty} \varphi_j(x)\varphi_j(\bar{x}) = k(x, \bar{x})$$

We might be able to compute the series in closed form.

The function k is called kernel.

Can we run ridge regression ?

Kernel ridge regression

We have

$$\hat{f}^{\lambda}(x) = \sum_{i=1}^{n} \Phi(x)^\top \Phi(x_i) c_i = \sum_{i=1}^{n} k(x, x_i) c_i$$

$$c = (\widehat{K} + \lambda I)^{-1} \widehat{Y}, \qquad (\widehat{K})_{ij} = \Phi(x_i)^\top \Phi(x_j) = k(x_i, x_j)$$

$\widehat{K}$ is the kernel matrix, the Gram (inner products) matrix of the data.

"The kernel trick"

Kernels

- Can we start from kernels instead of features?

- Which functions $k : X \times X \to \mathbb{R}$ define kernels we can use?

Positive definite kernels

A function $k : X \times X \to \mathbb{R}$ is called positive definite:

- if the matrix $\hat{K}$ is positive semidefinite for all choice of points $x_1, \dots, x_n$, i.e.
$$a^\top \widehat{K} a \geq 0, \qquad \forall a \in \mathbb{R}^n.$$

- Equivalently
$$\sum_{i,j=1}^{n} k(x_i, x_j) a_i a_j \geq 0,$$
for any $a_1, \dots, a_n \in \mathbb{R}$, $x_1, \dots, x_n \in X$.

Inner product kernels are pos. def.

Assume $\Phi : X \to \mathbb{R}^p$, $p \leq \infty$ and

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

Note that

$$\sum_{i,j=1}^{n} k(x_i, x_j) a_i a_j = \sum_{i,j=1}^{n} \Phi(x_i)^\top \Phi(x_j) a_i a_j = \left\| \sum_{i=1}^{n} \Phi(x_i) a_i \right\|^2 .$$

Clearly k is symmetric.

But there are many pos. def. kernels

Classic examples

- linear $k(x, \bar{x}) = x^\top \bar{x}$
- polynomial $k(x, \bar{x}) = (x^\top \bar{x} + 1)^s$
- Gaussian $k(x, \bar{x}) = e^{-\|x-\bar{x}\|^2 \gamma}$

But one can consider

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

It is natural to think of a kernel as a measure of similarity.

From pos. def. kernels to functions

Let X be any set/ Given a pos. def. kernel k.

- consider the space $\mathcal{H}_k$ of functions

$$f(x) = \sum_{i=1}^{N} k(x, x_i) a_i$$

  for any $a_1, \dots, a_n \in \mathbb{R}$, $x_1, \dots, x_n \in X$ and any $N \in \mathbb{N}$.

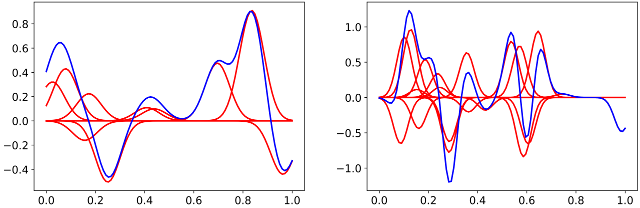
- Define an inner product on $\mathcal{H}_k$

$$\left\langle f, \bar{f} \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} \sum_{j=1}^{\bar{N}} k(x_i, \bar{x}_j) a_i \bar{a}_j.$$

- $\mathcal{H}_k$ can be *completed* to a Hilbert space.

A key result

Functions defind by Gaussian kernels with large and small widths.



An illustration

Theorem

Given a pos. def. k there exists Φ *s.t.* $k(x,\bar{x}) = \langle \Phi(x), \Phi(\bar{x}) \rangle_{\mathcal{H}_k}$ *and*

$$\mathcal{H}_\Phi \simeq \mathcal{H}_k$$

Roughly speaking

$$f(x) = w^\top \Phi(x) \qquad \simeq \qquad f(x) = \sum_{i=1}^{N} k(x, x_i) a_i$$

From features and kernels to RKHS and beyond

$\mathcal{H}_k$ and $\mathcal{H}_\Phi$ have many properties, characterizations, connections:

- reproducing property
- reproducing kernel Hilbert spaces (RKHS)
- Mercer theorem (Kar hunen Loéve expansion)
- Gaussian processes
- Cameron-Martin spaces

Reproducing property

Note that by definition of $\mathcal{H}_k$

- $k_x = k(x, \cdot)$ is a function in $\mathcal{H}_k$
- For all $f \in \mathcal{H}_k$, $x \in X$
$$f(x) = \langle f, k_x \rangle_{\mathcal{H}_k}$$
called the reproducing property

- Note that
$$|f(x) - \bar{f}(x)| \leq \|k_x\|_{\mathcal{H}_k} \left\|f - \bar{f}\right\|_{\mathcal{H}_k}, \qquad \forall x \in X.$$

The above observations have a converse.

RKHS

Definition

A RKHS $\mathcal{H}$ is a Hilbert with a function $k : X \times X \to \mathbb{R}$ s.t.

- $k_x = k(x, \cdot) \in \mathcal{H}_k$,
- *and*

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}_k}.$$

Theorem

If $\mathcal{H}$ is a RKHS then k is pos. def.

Evaluation functionals in a RKHS

If $\mathcal{H}$ is a RKHS then the evaluation functionals

$$e_x(f) = f(x)$$

are continuous. i.e.

$$|e_x(f) - e_x(\bar{f})| \lesssim \left\|f - \bar{f}\right\|_{\mathcal{H}_k}, \qquad \forall x \in X$$

since

$$e_x(f) = \langle f, k_x \rangle_{\mathcal{H}_k}.$$

Note that $L^2(\mathbb{R}^d)$ or $C(\mathbb{R}^d)$ don't have this property!

Alternative RKHS definition

Turns out the previous property also characterizes a RKHS.

Theorem

A Hilbert space with continuous evaluation functionals is a RKHS.

Summing up

- From linear to non linear functions
- using features
- using kernels

plus

- pos. def. functions
- reproducing property
- RKHS