Kernel methods

Alessandro Rudi

March 17, 2022

To learn more about the topic of this lecture, please look at the following documents:

- `http://cbio.ensmp.fr/~jvert/svn/kernelcourse/slides/master/master.pdf`
- `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf`
- `http://www.di.ens.fr/~fbach/rasma_fbach.pdf`
- `https://francisbach.com/cursed-kernels/`

In this course, we often focused on prediction methods which are *linear*, that is, the input data are vectors (i.e., $x \in \mathbb{R}^d$) and the prediction function is linear: $f(x) = w^\top x$ for $w \in \mathbb{R}^d$. In this situation, given data (x_i, y_i), $i = 1, \dots, n$, the vector w is obtained by minimizing

$$\widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w).$$

Classical examples are logistic regression or least-squares regression.

These methods look at first sight of limited practical significance, because:

- Input data may not be vectors.
- Relevant prediction functions may not be linear.

The goal of kernel methods is to go beyond theses limitations while keeping the good aspects. The underlying principle is to replace x by any function $\varphi(x) \in \mathbb{R}^d$, *explicitly* ou *implicitly*, and consider linear predictions in $\Phi(x)$, i.e., $f(x) = w^\top \varphi(x)$. We call $\varphi(x)$ the "feature" associated to x.

Examples

1. Linear regression. $\varphi(x) = x$ and $x \in \mathbb{R}^d$ in this way we have linear models, as expected, indeed

$$f(x) = w^\top x = \sum_{j=1}^{d} w_j x_j,$$

2. Polynomial regression of degree r. when $x \in \mathbb{R}$. In this case we have $\varphi(x) \in \mathbb{R}^r$ defined as $\varphi(x) = (1, x, x^2, \ldots, x^r)$ leading to general polynomials of degree at most r,

$$f(x) = w^\top \varphi(x) = \sum_{j=1}^{r} w_j (\varphi(x))_j = \sum_{j=1}^{r} w_j x^j.$$

3. Polynomial multivariate regression of degree r. By considering $x \in \mathbb{R}^d$ and

$$\varphi(x) = \left(x_1^{\alpha_1} \cdots x_d^{\alpha_d}\right)_{\sum_{i=1}^d \alpha_i = r} \quad .$$

   In this situation, $p = \binom{d+r-1}{r}$ (number of k-combinations with repetitions from a set with cardinality d), can be too big for an explicit representation to be feasible.

4. Generic set of functions. Let $\phi_1, \ldots \phi_r : \mathbb{R}^d \to \mathbb{R}$ be a set of functions of interest (e. g. a subset of the fourier basis), then we define $\varphi(x) \in \mathbb{R}^r$ as $\varphi(x) = (\phi_1(x), \ldots, \phi_r(x))$ and so

$$f(x) = w^\top \varphi(x) = \sum_{j=1}^{r} w_j \phi_j(x)$$

1 Representer theorem

The question is if there exists an easier representation for w in terms of the observed feature maps $\varphi(x_1), \ldots, \varphi(x_n)$, given a dataset $x_1, \ldots, x_n \in \mathbb{R}^d$. I.e., we want to know if it is possible to characterize the minimum $\widehat{w}$ of

$$\widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top \varphi(x_i)) + \lambda w^\top w$$

in the form of $\widehat{w} = \sum_{i=1}^n \alpha_i \varphi(x_i)$, with $\alpha_i \in \mathbb{R}$ for $i = 1, \ldots, n$. The following theorem guarantees such characterization under basic properties of $\widehat{L}$.

Theorem 1 ***(Representer theorem, 1971).***
Let $\varphi : \mathcal{X} \to \mathbb{R}^d$. Let $(x_1, \ldots, x_n) \in \mathcal{X}^n$, and assume $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ strictly increasing with respect to the last variable, then the minimum of

$$\widehat{L}(w) := \Psi(w^\top \varphi(x_1), ..., w^\top \varphi(x_n), w^\top w),$$

is attained for $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ with $\alpha \in \mathbb{R}^n$.

Proof Let $w \in \mathbb{R}^d$, and $\mathcal{F}_D = \{\sum_{i=1}^n \alpha_i \varphi(x_i) \mid \alpha \in \mathbb{R}^n\}$. Let $w_D \in \mathcal{F}_D$ and $w_\perp \in \mathcal{F}_D^\perp$ such that $w = w_D + w_\perp$, then $\forall i$, $w^\top \varphi(x_i) = w_D^\top \varphi(x_i) + w_\perp^\top \varphi(x_i)$ with $w_\perp^\top \varphi(x_i) = 0$.

From Pythagoreas theorem, we get: $w^\top w = w_D^\top w_D^2 + w_\perp^\top w_\perp$. Therefore we have:

$$\begin{aligned}\Psi(w^\top \varphi(x_1), ..., w^\top \varphi(x_n), w^\top w) =& \Psi(w_D^\top \varphi(x_1), ..., w_D^\top \varphi(x_n), w_D^\top w_D + w_\perp^\top w_\perp) \\ \geq& \Psi(w_D^\top \varphi(x_1), ..., w_D^\top \varphi(x_n), w_D^\top w_D).\end{aligned}$$

Thus

$$\inf_{w \in \mathbb{R}^d} \Psi(w^\top \varphi(x_1), ..., w^\top \varphi(x_n), w^\top w) = \inf_{w \in \mathcal{F}_D} \Psi(w^\top \varphi(x_1), ..., w^\top \varphi(x_n), w^\top w).$$

■

Corollary 1 *For $\lambda > 0$, $\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum \ell(y_i, w^\top \varphi(x_i)) + \frac{\lambda}{2} w^\top w$ is attained at $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$.*

- It is important to note that there is no assumption on ℓ (no convexity).
- This result is extendable to Hilbert spaces (RKHS) (see next section).

1.1 Finite dimensional representation of the learning problem

Since by representer theorem we know that the minimum of $\widehat{L}$ is of the form $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ we can directly optimize this characterization, i.e. we can then write:

$$\min_{w \in \mathbb{R}^r} \frac{1}{n} \sum \ell(y_i, w^\top \varphi(x_i)) + \frac{\lambda}{2} w^\top w = \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha.$$

where K is a $n \times n$ matrix with values

$$K_{ij} = \varphi(x_i)^\top \varphi(x_j).$$

Indeed

$$\varphi(x_i)^\top w = \sum_{j=1}^n \alpha_j \varphi(x_i)^\top \varphi(x_j) = (K\alpha)_i,$$

moreover

$$\|w\|^2 = w^\top w = \sum_{i,j=1}^n \alpha_i \alpha_j \varphi(x_i)^\top \varphi(x_j) = \alpha^\top K \alpha.$$

Finally we have a closed form representation also for the function evaluation. Define the **kernel function** $k(x, x') = \varphi(x)^\top \varphi(x')$, we have

$$f(x) = w^\top \varphi(x) = \sum_{i=1}^n \alpha_i \varphi(x_i)^\top \varphi(x) = \sum_{i=1}^n \alpha_i k(x_i, x).$$

Remark: Kernel Trick. The whole learning problem can be written in terms of the kernel k, indeed f depends only on k, $\widehat{L}$ depends on K and $K_{ij} = k(x_i, x_j)$. Then we have the so called *kernel trick*, i.e. we

don't need to compute explicitly the features φ to be able to represent and solve the learning problem, we need just to be able to compute their inner product.

Example: Power of the kernel trick, infinite dimensional feature maps. Consider $X = (-1, 1)$ and the feature map

$$\varphi(x) = (1, x, x^2, \dots).$$

The resulting model space would have the form

$$f(x) = \sum_{j \in \mathbb{N}} w_j x^j,$$

with $\sum_{j \in \mathbb{N}} w_j^2 < \infty$, corresponding to the set of the analytic functions on $(-1, 1)$, that is a very rich space. In particular it is dense in the space of continuous functions. However it is not possible to compute $\varphi(x)$ explicitly since it is infinite dimensional. The kernel trick provides an elegant way to compute the solution of the learning problem in closed form, indeed the inner product can be computed in closed form in $O(1)$,

$$k(x, x') = \varphi(x)^\top \varphi(x') = \sum_{j \in \mathbb{N}} x^j x'^j = \frac{1}{1 - xx'}.$$

The kernel trick allows to:

- replace $\mathbb{R}^d$ by $\mathbb{R}^n$; this is interesting when d is very large.
- separate the representation problem (design a kernel on a set $\mathcal{X}$) and algorithms and analysis (which only use the kernel matrix K).

2 Kernels

Since the learning problem is completely defined in terms of the kernel function, the explicit knowledge of the feature map is not required anymore. In particular given a function $k : X \times X \to \mathbb{R}$, to use it in a learning problem, we need to be sure that it is a *positive definite kernel*, i.e., that there exists a feature map φ such that

$$k(x, x') = \varphi(x)^\top \varphi(x'), \quad \forall\, x, x' \in X.$$

Kernel functions admits many characterizations

- **Characterization in terms of positive-definitness**: k is a positive definite kernel if and only if the kernel matrix K ispositive semi-definite (i.e. all their eigenvalues are non-negative) for all $n \in \mathbb{N}$ and $x_1, \dots, x_n \in X$.

  Theorem 2 ***(Aronszajn, 1950)***
  k is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{F}$, and $\varphi : \mathcal{X} \to \mathcal{F}$ such that $\forall x, y$, $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$.

- $\mathcal{F}$ is called the "feature space", and φ the "feature map".

- Simple properties (to be done as exercises): the sum and product of kernels are kernels. What are their associated feature space and feature map?
- Linear kernel: $k(x,y) = x^\top y$
- Polynomial kernel: the kernel $k(x,y) = (x^\top y)^r$ can be expanded as:

$$k(x,y) = (\sum_{i=1}^{d} x_i y_i)^r = \sum_{\alpha_1+...+\alpha_p=r} \binom{r}{\alpha_1,...,\alpha_p} \underbrace{(x_1 y_1)^{\alpha_1}...(x_p y_p)^{\alpha_p}}_{(x_1^{\alpha_1}...x_p^{\alpha_p})(y_1^{\alpha_1}...y_p^{\alpha_p})}$$

  We have: $\Phi(x) = \{\binom{r}{\alpha_1,...,\alpha_p}^{\frac{1}{2}} x_1^{\alpha_1}...x_p^{\alpha_p}\}$. Exercise: how can we go beyond homogeneous polynomials?

- **Translation-invariant kernels on** $[0,1]$. $k(x,y) = q(x-y)$ where q is 1-periodic. k is a positive definite kernel if and only if the Fourier series of q is non-negative (using the complex representation), i.e.,

$$k(x,y) = \nu_0 + \sum_{m\in\mathbb{N}} 2\nu_m \cos 2\pi mx \cos 2\pi my + 2\nu_m \sin 2\pi mx \sin 2\pi my$$

  with $\nu \geqslant 0$.

  The (infinite-dimensional) feature vector is composed of $\nu_0^{1/2}$, and of $\sqrt{2\nu_m}\cos 2\pi mx$ and $\sqrt{2\nu_m}\sin 2\pi mx$, for $m \geqslant 1$.

  If $f(x)$ can be written $f(x) = \Phi(x)^\top w$, then

$$\|w\|^2 = \big(\int_0^1 f(x)\big)^2 + \sum_{m\geqslant 1} \frac{2}{\nu_m}\big(\int_0^1 f(x)\cos 2\pi mx\big)^2 + \frac{2}{\nu_m}\big(\int_0^1 f(x)\sin 2\pi mx\big)^2.$$

  For $\nu_m = \frac{1}{m^{2s}}$, $m \geqslant 1$, this norm is equal to

$$\|w\|^2 = \big(\int_0^1 f(x)\big)^2 + \frac{1}{(2\pi)^{2s}}\int_0^2 |f^{(s)}(x)|^2 dx$$

  and the kernel has an analytical expression $k(x,y) = \nu_0 + (-1)^{s-1}\frac{(2\pi)^{2s}}{(2s)!}B_{2s}(\{x-y\})$, where B_{2s} is Bernoulli's polynomial.

- **Translation-invariant kernels on** $\mathbb{R}^d$: $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = q(x-y)$ with $q : \mathbb{R}^d \to \mathbb{R}$.

  Theorem 3 ***(Böchner)****: k is positive definite $\Leftrightarrow$ q is the Fourier transform of a non-negative Borel measure $\Leftarrow$ $q \in L^1$ and its Fourier transform is non-negative.*

  Proof (partial) First we recall that the Fourier transform $\widehat{f}$ of a function f is defined as

$$\widehat{f}(\omega) = \int e^{i\omega x} f(x) dx$$

Let $x_1, ...x_n \in \mathbb{R}^d$, let $\alpha_1, .., \alpha_n \in \mathbb{R}$, you know that $q(x) = \int e^{-i\omega x} \widehat{q}(\omega) d\omega$

$$\begin{aligned} K \text{ is p.d.} \equiv \alpha^\top K \alpha \geq 0 \forall \alpha \in \mathbb{R}^n \equiv \sum \alpha_s \alpha_j k(x_s, x_j) &= \sum \alpha_s \alpha_j q(x_s - x_j) \\ &= \sum \alpha_s \alpha_j \int e^{-i\omega^\top (x_s - x_j)} \widehat{q}(\omega) d\omega \\ &= \int (\sum \alpha_s \alpha_j e^{-i\omega^\top x_s} \overline{e^{-i\omega^\top x_j}}) \widehat{q}(\omega) d\omega \\ &= \int |\sum \alpha_s e^{-i\omega^\top x_s}|^2 \widehat{q}(\omega) d\omega \geq 0. \end{aligned}$$

■

Construction of the norm. Intuitive (non-rigorous) reasoning: if q is in L^1, then $\widehat{q}(\omega)$ exists and, with $d\mu(\omega) = \widehat{q}(\omega) d\omega$, we have an explicit representation of

$$k(x, y) = \int \langle \sqrt{\widehat{q}(\omega)} \exp^{-i\omega^\top x}, \sqrt{\widehat{q}(\omega)} \exp^{-i\omega^\top x} \rangle d\omega = \int \langle \varphi_\omega(x), \varphi_\omega(y) \rangle d\omega = \langle \varphi(x), \varphi(y) \rangle.$$

If we consider $f(x) = \int \varphi_\omega(x) w_\omega d\omega$, then $w_\omega = \widehat{f}(\omega)/\sqrt{\widehat{q}(\omega)}$, and the squared norm of w is equal to $\int \frac{|\widehat{f}(w)|^2}{\widehat{q}(w)} dw$, where $\widehat{f}$ denotes the Fourier transform of f.

Examples: Exponential kernel $\exp(-\alpha|x - y|)$ and Gaussian kernel $\exp(-\alpha|x - y|^2)$.

- Many applications of the kernel trick!
  - Exercise: show that on $\mathcal{X} = \mathbb{R}^+$, $k(x, y) = \min(x, y)$ and $k(x, y) = \frac{xy}{x+y}$ are positive definite kernels.
- Non vectorial data (sequences, graphes, images).
  - Exercise: for $\mathcal{X}$ the set of all subsets of a given set V, show that $k(A, B) = \frac{|A \cap B|}{|A \cup B|}$ is a positive definite kernel.
  - Examples of kernels on sequences

3 Ridge regression (mostly as an exercise)

We consider the optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \|y - \Phi w\|_2^2 + \frac{\lambda}{2} \|w\|_2^2.$$

with $\Phi \in \mathbb{R}^{n \times r} = (\varphi(x_1)^\top, \ldots, \varphi(x_n)^\top)$ and $y \in \mathbb{R}^n = (y_1, \ldots, y_n)$. We can solve it in two ways (done as an execise):

1. **Direct** : $\min_{w\in\mathbb{R}^d} \frac{1}{2n}\|y-\Phi w\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
2. **With representer theorem** : $\min_{\alpha\in\mathbb{R}^n} \frac{1}{2n}||y-K\alpha||_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha$

1. Using the representer theorem:
   gradient with respect to α : $\frac{1}{n}K(K\alpha-y)+\lambda K\alpha = 0 \Leftrightarrow (K^2+n\lambda K)\alpha = Ky \Leftrightarrow K((K+n\lambda I)\alpha-y) = 0$.
   If $K$is non invertible, the solution is not unique : $\alpha = (K+n\lambda I)^{-1}y+Ker(K)$. However the prediction is unique : $K\alpha = K(K+n\lambda I)^{-1}y$.
2. Direct method: minimizing with respect to w

   gradient w.r.t. w : $\frac{1}{n}\Phi^\top(\Phi w-y)$
   This leads to $w = (\frac{1}{n}\Phi^\top\Phi+\lambda I)^{-1}\frac{1}{n}\Phi^\top y \Leftrightarrow \Phi f = \Phi(\frac{1}{n}\Phi^\top\Phi+\lambda I)^{-1}\frac{1}{n}\Phi^\top y$.

By noting that $K = \Phi\Phi^\top$, we get :

$$\overbrace{\Phi\Phi^\top(\underbrace{\Phi\Phi^\top}_{n\times n}+n\lambda I)^{-1}y}^{\text{kernel}} = \overbrace{\Phi(\underbrace{\Phi^\top\Phi}_{d\times d}+n\lambda I)^{-1}\Phi^\top y}^{\text{direc}}$$

This is simply:

Lemma 1 *(matrix inversion lemma)* $\forall A$ *matrix,* $(AA^\top+I)^{-1}A = A(A^\top A+I)^{-1}$

There is thus an "equivalence" betweeen this lemma and the representer theorem.

4 Complexity of linear algebra computations

If $K\in\mathbb{R}^{n\times n}$ and $L\in\mathbb{R}^{n\times n}$ are two matrices

- computing KL has complexity $O(n^3)$
- computing K^{-1}has complexity $O(n^3)$
- computing Ky has complexity $O(n^2)$
- Solving $K^{-1}y$ has complexity $O(n^3)$
- Decomposing K in eigenvalues / eigenvectors $O(n^3)$
- Largest eigenvector: $O(n^2)$

Low-rank approximation

- Eigenvector basis (complexity $O(n^2r)$)
- Orthogonal projection on first r columns: $O(nr^2)$