Machine learning - Master ICFP 2019-2020

Kernel methods

Francis Bach

March 6, 2020

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`

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

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

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.

Example. Polynomial regression of degree r, by considering $x \in \mathbb{R}^d$ and

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

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.

WARNING. The type of kernel is different from the ones in lecture 2. The ones here are "positive definite"; the ones from lecture 2 are "non-negative". See more details in `https://francisbach.com/cursed-kernels/`

1 Representer theorem

Theorem 1 ***(Representer theorem, 1971).***
Let $\varphi : \mathcal{X} \to \mathbb{R}^d$. Let $(x_1, \dots, x_n) \in \mathcal{X}^n$, and assume $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ strictly increasing with respect to the last variable, then the minimum of $\Psi(w^\top \varphi(x_1), ..., w^\top \varphi(x_n), w^\top w)$ is attained for $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$ with $\alpha \in \mathbb{R}^n$.

Proof Let $w \in \mathbb{R}^d$, and $\mathcal{F}_D = \{\sum \alpha_i \Phi(x_i) / \alpha \in \mathbb{R}^n\}$. 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).
- We have: $\forall j \in \{1, \dots, n\}$, $w^\top \varphi(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j) = (K\alpha)_j$ where K is the kernel matrix and $w^\top w = \alpha^\top K \alpha$. We can then write:

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

For a test point, we have $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$.

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

- **Definition**: k is a positive definite kernel if and only if all kernel matrices are positive semi-definite.

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

- $\mathcal{F}$ 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:

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

  We have: $\Phi(x) = \{ \binom{r}{\alpha_1, \ldots, \alpha_p}^{\frac{1}{2}} x_1^{\alpha_1} \ldots x_p^{\alpha_p} \}$. 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.,

$$k(x,y) = \nu_0 + \sum_{m \geqslant 1} 2\nu_m \cos 2\pi m x \cos 2\pi m y + 2\nu_m \sin 2\pi m x \sin 2\pi m y$$

  with $\nu \geqslant 0$.
  The (infinite-dimensional) feature vector is composed of $\nu_0^{1/2}$, and of $\sqrt{2\nu_m} \cos 2\pi m x$ and $\sqrt{2\nu_m} \sin 2\pi m x$, for $m \geqslant 1$.
  If $f(x)$ can be written $f(x) = \Phi(x)^\top w$, then

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

  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) Let $x_1, ...x_n \in \mathbb{R}^d$, let $\alpha_1, .., \alpha_n \in \mathbb{R}$,

$$\begin{aligned}\sum \alpha_s \alpha_j k(x_s, x_j) &= \sum \alpha_s \alpha_j q(x_s - x_j) \\ &= \sum \alpha_s \alpha_j \int \exp^{-i\omega^\top (x_s - x_j)} d\mu(\omega) \\ &= \int (\sum \alpha_s \alpha_j \exp^{-i\omega^\top x_s} \overline{\exp^{-i\omega^\top x_j}}) d\mu(\omega) \\ &= \int |\sum \alpha_s \exp^{-i\omega^\top x_s}|^2 d\mu(\omega) \geq 0.\end{aligned}$$

■

Consruction 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:

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

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 $w = (\frac{1}{n}\Phi^\top\Phi + \lambda I)^{-1}\frac{1}{n}\Phi^\top y \Leftrightarrow \Phi f = \Phi(\frac{1}{n}\Phi^\top\Phi+\lambda I)^{-1}\frac{1}{n}\Phi^\top y$.

With $K = \Phi\Phi^\top$, 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: $O(nr^2)$

5 Using distances in feature space (exercise)

Given sets of negative examples $x_i, i \in I_-$ and positive examples $x_i, i \in I_+$, we consider the average μ_- of negative points and the average μ_+ of positive points *in the feature space*.

(1) For a testing point x, compute $\|\Phi(x) - \mu_+\|^2$ and $\|\Phi(x) - \mu_-\|^2$.

(2) We classify x as positive if $\|\Phi(x) - \mu_+\|^2 > \|\Phi(x) - \mu_-\|^2$. Relate the classification rule to existing classifiers. We can consider that $|I_+| = |I_-|$.